

CHARACTERIZATION AND MITIGATION OF FALSE
INFORMATION ON THE WEB

by
Anu Shrestha



A dissertation
submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computer Science
Boise State University

May 2022

© 2022

Anu Shrestha

ALL RIGHTS RESERVED

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Anu Shrestha

Thesis Title: Characterization and Mitigation of False Information on the Web

Date of Final Oral Examination: 11th March, 2022

The following individuals read and discussed the dissertation submitted by student Anu Shrestha, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Francesca Spezzano Ph.D.	Chair, Supervisory Committee
Edoardo Serra Ph.D.	Member, Supervisory Committee
Maria Soledad Pera Ph.D	Member, Supervisory Committee

The final reading approval of the thesis was granted by Francesca Spezzano Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

DEDICATION

To my parents and my sisters for their unconditional love and support.

ACKNOWLEDGMENT

I owe my gratitude to all the people who have made this dissertation possible.

When I started planning for my higher studies in the United States, I was uncertain about the various process involved in converting my plan into a reality. As I am moving closer to seeing that plan unfold into a reality, I am thankful to many people who helped me in one way or the other during my Ph.D. journey.

First and foremost, I can't thank enough to my advisor Dr. Francesca Spezzano for providing me the necessary guidance, support, suggestions, and much-needed motivation during this entire time of Ph.D. studies. I am very certain that the learning and experience that I gained while working with my advisor in multiple research works, paper publications, conference presentations, and workshops are going to be lifelong useful to me. I am equally grateful to my committee members Dr. Edoardo Serra and Dr. Maria (Sole) Pera for their valuable advice and support during my entire graduate school and for serving as my committee. The discussions with Dr. Serra on various topics related to the research papers, collaborations and his resourcefulness have always helped in making my understanding of the subject much clear. Dr. Pera's thoughtful feedback and input on collaborations and research publications have helped me gain a perspective of the bigger picture of the work.

My passion for learning and contributing to the science and technology field would be very limited had I not been provided the full scholarship by Boise State Univer-

sity for my graduate studies. I feel very fortunate and grateful for this opportunity provided by the Department of Computer Science at Boise State University.

This acknowledgment would be incomplete without mentioning how thankful I am to my friends who have cheered me through the rough times and celebrated my every success. And, last but not the least, my parents and siblings who have endured every hardship to see me succeed, I owe this and every bit of achievement to them.

ABSTRACT

Social media and Web sources have made information available, accessible, and shareable any time and anywhere nearly without friction. This information can be truthful, falsified, or can only be the opinion of the writer as users in such platforms are both information creators and consumers. In any case, it has the power to affect the decision of an individual, the beliefs of the society, activities, and the economy of the whole country. Thus, it is imperative to identify false information and mitigate the effects of false information that are ubiquitous across the Web and social media. Therefore, the main goal of this dissertation is to proactively combat false information by defining three objectives. First, analyze the reason behind the success of its motive, second, recognize and quantify the impacts made on information systems, and third, develop novel ways of identifying false information and the actors responsible for creating and spreading them. The achievement of these three objectives enhanced our understanding of false information and helped in building strategies to mitigate this phenomenon.

Overall, this dissertation presents our research on in-depth analysis of malicious entities, their impact in the information ecosystem, and the models we build to accurately detect different malicious entities like fraudulent reviewers, fake news, fake news spreaders in real-world scenarios. We show that each of our methods outperforms the existing state-of-the-art methods in the detection of false information and

malicious actors in real-world opinion-based systems and social media.

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGMENT	v
ABSTRACT	vii
LIST OF TABLES	xviii
LIST OF FIGURES	xxiv
LIST OF ABBREVIATIONS	xxviii
I Introduction and Background	1
1 INTRODUCTION	2
1.1 Motivation	2
1.2 Overview and contributions	5
1.3 Overarching thesis statement	11
2 BACKGROUND AND RELATED WORKS	13
2.1 Motivation and success of false information	13
2.2 Impact of false information	16

2.3	Detection of false information, fake news spreaders and fraudulent reviewers	17
II	Investigating why deceptive information is successful	22
3	THAT’S FAKE NEWS! INVESTIGATING HOW READERS IDENTIFY THE RELIABILITY OF NEWS WHEN PROVIDED TITLE, IMAGE, SOURCE BIAS, AND FULL ARTICLES	23
3.1	Introduction	23
3.2	Related Work	26
3.3	Methods	31
3.3.1	Dataset: FakeNewsNet	31
3.3.2	Evaluation by People	32
3.3.3	Reasoning Given for Evaluations	34
3.3.4	Evaluation by Computer – Automated Detector	35
3.4	Results	37
3.4.1	Comparing Conditions Evaluated by People	37
3.4.2	Reasons for Participant’s Judgement of Real or Fake	40
3.4.3	Comparing Conditions Evaluated by Computer	42
3.5	Discussion	43
3.6	Limitations	48
3.7	Conclusion & Future Work	49
III	Measuring the impact of deceptive information in in-	

formation systems **51**

4 AN EMPIRICAL ANALYSIS OF COLLABORATIVE RECOMMENDER SYSTEMS ROBUSTNESS TO SHILLING ATTACKS 52

4.1 Introduction 52

4.2 Related work 54

4.3 Experimental Settings 56

 4.3.1 Datasets 56

 4.3.2 Algorithms 58

 4.3.3 Evaluation Framework 59

4.4 Results and Discussion 61

 4.4.1 Do spammers ratings impact recommendations? 62

 4.4.2 Who is really affected by spammers? 64

4.5 Conclusions 68

IV Detecting trustworthy entities in the information ecosystem **70**

5 DEEPTRUST: AN AUTOMATIC FRAMEWORK TO DETECT TRUSTWORTHY USERS IN OPINION-BASED SYSTEMS 71

5.1 Introduction 71

5.2 Related Work 75

5.3 DeepTrust User Embedding 78

5.4 DeepTrust Architecture 81

5.5 Dataset 85

5.6	Experiments	86
5.6.1	Experimental Settings	87
5.6.2	Detecting Trustworthy, Unreliable, and Fraudulent Users	88
5.6.3	Detecting Fraudulent Users	91
5.6.4	Classifying Trustworthy vs. Untrustworthy Users	92
5.6.5	Classifying Unknown Users	93
5.7	Discussion	94
5.8	Conclusions	97
6	TEXTUAL CHARACTERISTICS OF NEWS TITLE AND BODY TO DE- TECT FAKE NEWS: A REPRODUCIBILITY STUDY	99
6.1	Introduction	99
6.2	Overview of the paper by Horne and Adali	101
6.2.1	Approach	101
6.2.2	Features	102
6.2.3	Observation and Evaluation	103
6.3	Reproducibility	104
6.3.1	Datasets	104
6.3.2	Features	106
6.3.3	Analysis	108
6.3.4	Results	110
6.4	How to Reproduce our Experiments	115
6.5	Conclusions	116
7	CHARACTERIZING AND PREDICTING FAKE NEWS SPREADERS IN	

SOCIAL NETWORKS	117
7.1 Introduction	117
7.2 Related Work	120
7.3 Datasets	122
7.4 Features	125
7.4.1 Demographics	125
7.4.2 Behavioral-based features	127
7.4.3 Network-based features	128
7.4.4 Emotions	128
7.4.5 Personality	129
7.4.6 Readability	131
7.4.7 Writing Style	132
7.5 User Characterization	132
7.5.1 Demographics	133
7.5.2 User Behavior	135
7.5.3 User Network	135
7.5.4 User Emotions	136
7.5.5 User Personality Traits	138
7.5.6 User Readability Level and Writing Style	139
7.6 Experiments	141
7.6.1 Experimental Setting	141
7.7 Classification Results	144
7.8 Feature Importance and Shapley Additive Explanations	145
7.9 Conclusions	147

8	JOINT CREDIBILITY ESTIMATION OF PUBLISHER, NEWS, AND USER VIA HETEROGENEOUS GRAPH REPRESENTATION LEARNING.	150
8.1	Introduction	150
8.2	Related Work	154
8.3	Methodology	156
8.3.1	Relational Graph Convolutional Networks	157
8.3.2	Role-Relational Graph Convolutional Network (Role-RGCN)	158
8.4	Datasets	160
8.5	Experiments	162
8.5.1	Baselines	162
8.5.2	Experimental Setting	164
8.5.3	Experimental Results	165
8.6	Conclusion	166
8.7	Appendices	167
8.7.1	Markov Random Field-based model	167
9	MODELING THE DIFFUSION OF FAKE AND REAL NEWS THROUGH THE LENS OF THE DIFFUSION OF INNOVATIONS THEORY	171
9.1	Introduction	171
9.2	Related Work	174
9.2.1	Diffusion of Innovations Theory	176
9.3	Dataset	177
9.4	Features	180
9.4.1	User-Based Features	180
9.4.2	Network-Based Features	185

9.4.3	News-Based Features	187
9.5	Experiments	188
9.5.1	Experimental Settings	188
9.5.2	Baselines for Comparison	189
9.5.3	Results and Analysis	190
9.6	Conclusion	196
V	Concluding Remarks	198
10	CONCLUSION AND FUTURE DIRECTIONS	199
10.1	Conclusion	199
10.2	Future Directions	202
	REFERENCES	203
	APPENDICES	244
A	QUALITATIVE CODES	245
B	MULTI-MODAL ANALYSIS OF MISLEADING POLITICAL NEWS	252
B.1	Introduction	252
B.2	Related Work	255
B.3	Datasets	257
B.4	Multi-modal Features	260
B.4.1	Textual Features	261
B.4.2	Image Features	263

B.4.3	Source Bias	265
B.5	Multi-modal Analysis	265
B.5.1	Do We Need to “Read”?	272
B.5.2	Can we Detect Misleading News from its Snippet?	273
B.6	Conclusion	274
C	AN ANALYSIS OF PEOPLE’S REASONING FOR SHARING REAL AND FAKE NEWS	275
C.1	Introduction	275
C.2	Related Work	278
C.3	Data Collection	280
C.4	Data Analysis	281
C.5	Predicting News Sharing	285
C.5.1	Textual Features Extraction	285
C.5.2	Experimental Setting and Results	286
C.6	Conclusion and Future Work	288
D	MULTI-MODAL SOCIAL AND PSYCHO-LINGUISTIC EMBEDDING VIA RECURRENT NEURAL NETWORKS TO IDENTIFY DEPRESSED USERS IN ONLINE FORUMS	290
D.1	Introduction	291
D.2	Related Work	293
D.3	The ReachOut Forum	295
D.3.1	Dataset	297
D.4	Methodology	299

D.4.1	Unsupervised Learning of User Representation from Their Posts	300
D.4.2	Network Features	303
D.4.3	Anomaly Detection	307
D.5	Experiments	309
D.5.1	Experimental Setting	309
D.5.2	Baselines for comparison	310
D.5.3	Results	311
D.6	Conclusion	314

LIST OF TABLES

3.1	Comparison of accuracy between conditions: news text excerpt (E); title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B). Note: with Bonferroni correction, only a p-value < 0.0083 would be significant at a family-wise alpha level of 0.05. . . .	38
3.2	The accuracy of a computer model in identifying real and fake news when using: excerpt (E); title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B).	43
4.1	Details on the datasets considered for our analysis.	56
4.2	Performance analysis using different metrics on datasets with and without spam. Statistically significant differences are shaded in gray, <i>pvalue</i> ≤ 0.001	62
4.3	Statistically significant RMSE differences between recommendations without spammers (W/o) and with spammers (W) across different user fairness ranges (* means <i>pvalue</i> < 0.03 and ** means <i>pvalue</i> ≤ 0.01). Cases where removing spammers reduces the RMSE are shaded. . . .	65
4.4	Statistically significant NDCG@5 differences between recommendations without spammers (W/o) and with spammers (W) across different user fairness ranges (* means <i>pvalue</i> < 0.03 and ** means <i>pvalue</i> ≤ 0.01). Cases where removing spammers increases the NDCG@5 are shaded.	67

4.5	Statistically significant HR@5 differences between recommendations without spammers (W/o) and with spammers (W) across different user fairness ranges (* means $pvalue < 0.03$ and ** means $pvalue \leq 0.01$). Cases where removing spammers reduces the HR@5 are shaded. . . .	67
5.1	Number of users for each class in the Amazon dataset.	87
5.2	Precision, recall, and F1-measure results of detecting trustworthy, fraudulent, and unreliable users with DeepTrust and comparison with related work.	88
5.3	Precision, recall, and F1-measure for DeepTrust combined with related work for detecting trustworthy, fraudulent, and unreliable users. . . .	88
5.4	Precision, recall, and F1-measure results of detecting trustworthy, fraudulent, and unreliable <i>cold start</i> users with DeepTrust and comparison with related work.	90
5.5	Precision, recall, and F1-measure for DeepTrust combined with related work for detecting trustworthy, fraudulent, and unreliable <i>cold start</i> users.	90
5.6	Classification result for fraudulent user detection: precision, recall, F1-measure, AUROC, and average precision (AvgP).	92
5.7	Classification result for trustworthy vs.untrustworthy user detection: precision, recall, F1-measure, AUROC, and average precision (AvgP).	95
5.8	Precision, recall, F1-measure, AUROC, and average precision (AvgP) for DeepTrust combined with related work for detecting fraudulent users.	97

5.9	Precision, recall, F1-measure, AUROC and average precision (AvgP) of combining DeepTrust with related work for detecting trustworthy users.	97
6.1	Size of datasets used in our study.	105
6.2	Features that differ in body of news content. ($p < 0.05$).	109
6.3	Features that differ in the title of news content. All differences are statistically significant ($p < 0.05$).	110
6.4	News title vs. news body features for detecting fake news on the PolitiFact, BuzzFeedNews, and GossipCop datasets: stylistic, psychology, and complexity features. Best results for both news title and body are in bold. Best overall results between news title and body are shaded.	111
6.5	News title vs. news body features for detecting fake news on the PolitiFact, BuzzFeedNews, and GossipCop datasets: same four features as in Horne and Adali [1] – NN, TTR, WC, and Quote for news body and FK, NN, per_stop, and avg_wlen for title. Best results for both news title and body are in bold. Best overall results between news title and body are shaded.	112
6.6	Feature group ablation for news title and body when the best classifier (Random Forest) is used on the PolitiFact, BuzzFeedNews, and GossipCop datasets. Best results for both news title and body are in bold.	113
7.1	Datasets and statistics.	125

7.2	Writing style features that differ in user feed. All differences are statistically significant ($p \leq 0.002$ for PolitiFact and $p \leq 0.04$ for PAN 2020). Shaded cells indicate the same pattern in both datasets. . . .	140
7.3	Average precision of our proposed features (in input to a Random Forest classifier) on PolitiFact and PAN 2020 datasets and comparison with baselines. Best values are in bold.	144
7.4	Average precision per feature group on PolitiFact and PAN 2020 datasets.	145
7.5	Average precision of important features from Figure 7.9 vs. all features on PolitiFact and PAN 2020 datasets.	146
8.1	Datasets and statistics.	161
8.2	Macro F1 score of our proposed model on PolitiFact and comparison with baselines. Best values are in bold.	166
8.3	Macro F1 score of our proposed model on PolitiFact and comparison with classical RGCN. Best values are in bold.	166
8.4	Edge potentials for each entity in news ecosystem. Here FNS represents fake news spreader and RNS represents real news spreader.	168
9.1	Size of the PolitiFact dataset.	178
9.2	Size of the datasets used in our experiments.	180
9.3	Performance of our proposed features according to different classifiers on both real and fake news sharing and comparison with baselines. Best values are in bold.	191

9.4	Features that differ in fake and real news sharing. <i>S</i> means shared and <i>NS</i> means not shared. All differences are statistically significant ($p < 0.05$). Same feature with same trend for both title and text is denoted as ‘t&t’.	193
A.1	Thematic codes (and examples) that were inductively developed by analyzing the question “why did you identify the news item as real or fake”. The table identifies the code, provides a description, and example quotes [with the associated accuracy]. The accuracy is indicated as: FN = false negative (identified fake news as real); FP = false positive (identified real news as fake); TN = true negative (identified real news as real); TP = true positive (identified fake news as fake).	245
B.1	Available datasets for misleading news detection.	258
B.2	Feature ablation for <u>FakeNewsNet</u> (left) and <u>BuzzFeedNews</u> (right) datasets.	264
B.3	Top-30 most important news <u>body content</u> features and their corresponding logistic regression coefficients for the FakeNewsNet (left) and BuzzFeedNews (right).	266
B.4	Top-30 most important <u>headline</u> features and their corresponding logistic regression coefficients for FakeNewsNet (left) and BuzzFeedNews (right) datasets.	268
B.5	Top-10 most important <u>image</u> features and corresponding logistic regression coefficients for <u>FakeNewsNet</u> .	271

B.6	Results comparing news snippet feature combination (headline, image, and source bias) with news body content for <u>FakeNewsNet</u> (left) and <u>BuzzFeedNews</u> (right).	272
C.1	News Sharing Behavior.	281
C.2	Comparison of emotion, psycho-linguistic, and demographic features to predict whether a news item will be shared or not. We used a random forest classifier. Best results among feature groups considered separately are in bold. Best overall results are shaded.	287
D.1	Summary of the CLPsych 2017 dataset.	298
D.2	Precision (Pr), Recall (Re), and F1-measure (F1) of anomaly detection with social network features, psycho-linguistic features and combination.	311

LIST OF FIGURES

3.1	Recent examples of false social media posts (as identified by snopes.com) that illustrate use of title and image (T+I) and text excerpts (E). . .	24
3.2	Example stimuli from each condition: news text excerpt (E); title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B). Participants saw 3 examples for the Excerpt condition, and 5 examples for each of the other conditions.	33
3.3	Mean accuracy for each condition. Error bars are standard error of the means (SEM).	38
3.4	(a) Distribution of participants' self-identified political ideological leaning; (b) Accuracy in Title+Image condition for each political group. Accuracy in other conditions did not differ significantly between political groups. Error bars are SEM.	39
3.5	The five most frequent codes for the Title and Image (3.5a) and Excerpt (3.5b) conditions differentiated by when participants responded correctly and incorrectly.	41
4.1	Rating distribution across the four datasets considered in our analysis.	57
4.2	Rating distribution for attacked products by spammers and benign users.	58
4.3	RMSE differences across fairness range.	64

4.4	NDCG differences across fairness range.	66
5.1	Distribution of the Amazon rank among fraudulent reviewers in the dataset from Section 5.5.	72
5.2	Sample user-item bipartite rating network ($t_i \leq t_j$ if $i \leq j$).	75
5.3	DeepTrust architecture.	79
5.4	Description of an LSTM cell C . Figure adapted from [2].	83
5.5	Classification of unknown users as trustworthy (green), unreliable (yellow), and fraudulent (red) and correlation with the Amazon rank.	95
7.1	Distribution of user demographics on the PolitiFact dataset.	133
7.2	Boxplots of user behavioral features on the PolitiFact dataset.	135
7.3	Average Twitter follower to following (TFF) ratio on the PolitiFact dataset. The difference is statistically significant ($p \leq 0.001$)	136
7.4	Radar charts of the emotion features: PolitiFact and PAN 2020 datasets.	137
7.5	Box plots of user stress level on the PolitiFact and PAN 2020 datasets.	137
7.6	Radar charts of the Big-Five personality scores: PolitiFact and PAN 2020 datasets.	138
7.7	Readability index of tweets written by fake news spreaders vs real news spreaders in PolitiFact.	140
7.8	Readability index of tweets written by fake news spreaders vs real news spreaders in PAN 2020.	141

7.9	SHAP summaries of the important features: PolitiFact and PAN datasets. Y-axis represents the features in order of importance. X-axis represents the shap values, positive values (greater than zero) represents a higher chance of classifying a user as a fake news spreader and negative values represent a higher chance of classifying a user as a real news spreader.	143
8.1	Sample news ecosystem in social media.	152
8.2	Role-RGCN Architecture	159
9.1	News sharing network between users in Twitter.	179
9.2	Feature group analysis: AUROC and average precision per feature group for shared real and fake news.	192
B.1	Number of publishers per category in the MediaBias/FactCheck dataset.	260
B.2	Publisher credibility per bias and bias distribution within questionable sources in the MediaBias/FactCheck dataset.	260
B.3	Most important features for news <u>body content</u> with average values for factual and misleading news: FakeNewsNet (top) and BuzzFeedNews (bottom).	267
B.4	Most important features for news <u>headline</u> with average values for factual and misleading news: FakeNewsNet (top) and BuzzFeedNews (bottom).	269
B.5	Most important features for news <u>image</u> and average values for factual and misleading news.	271
C.1	News items used in our survey instrument.	280
C.2	Distribution of participant’s gender.	282

C.3	Distribution of participant’s self-identified political orientation. . . .	283
D.1	The triage annotation decision tree [3].	296
D.2	Overview of the proposed unsupervised technique to identify depressed users in online forums.	299
D.3	Autoencoder architecture	300
D.4	Description of an LSTM cell C . Figure adapted from [2].	302
D.5	PageRank (a), Reciprocity (b), and Local Clustering Coefficient (c) distribution of depressed (blue) and non-depressed (orange) users. . .	304
D.6	Plot the user embeddings computed with the autoencoder and reduced in a 2-dimensional space via PCA.	312

LIST OF ABBREVIATIONS

AMT Amazon Mechanical Turk

ANOVA Analysis of Variance

AUROC Area Under the Receiver Operating Characteristics

AvgP Average Precision

BERT Bidirectional Encoder Representations from Transformers

GCN Graph convolution Network

GRU Gated Recurrent Units

HR Hit Ratio

ICM Independent Cascade Model

LIWC Linguistic Inquiry and Word Count

LTM Linear Threshold Model

MAE Mean Absolute Error

MLP Multi-Layered Perceptron

NDCG Normalized Discounted Cumulative Gain

PMRF Pairwise Markov Random Field

PS Prediction Shift

RGCN Relational Graph Convolution Network

RNN Recurrent Neural Network

RSME Root Mean Square Error

RST Rhetorical Structure Theory

SVM support vector machine

Part I

Introduction and Background

CHAPTER 1: INTRODUCTION

1.1 Motivation

Social media and Web sources are valuable resources for public interactions about various issues, products, and services where people share information or their opinions openly. Unfortunately, such openness have also made it increasingly easy to abuse these platforms through content deception, i.e., the deliberate act intended to mislead others by falsifying information. False information is categorized into different forms, such as hoaxes in collaborative platforms, fraudulent reviews in e-commerce platforms, fake news and rumors in social media, etc., based on the information they carry and the intention behind their existence [4]. Hoax is defined as a falsehood that is deliberately fabricated to persuade readers and mimic as truth [5]. It may include information about the person, organization, or event that never existed in the real world and is created with benevolent intention as a joke or prank. Rumor represents a piece of information about an event or person whose veracity is never confirmed and circulated without ensuring the truth [6]. On the other hand, fake news and fraudulent reviews are deliberately created with malicious intentions that contain political, social, psychological, and financial connections [7, 8, 9]. Opinion spam, also known as fraudulent review, is a review with fictitious opinions that are deliberately

written to sound authentic and fake news is a low-quality news that is created to spread misinformation and mislead readers.

In this dissertation, we address two main forms of online false information: fraudulent reviews (or opinion spam) and fake news. It is frequent to have paid reviewers writing fraudulent reviews (shilling attacks) to promote or demote products or businesses in most of the e-commerce platforms as the purchase decisions of customers are highly correlated with the amount and nature of reviews [10, 11]. This eventually affects the ranking of the business and their revenue [12]. With the growth of these platforms, the prevalence of fraudulent reviews is also rising significantly from 5% to 61% [12, 13, 14, 15, 16, 17]. This behavior where malicious users create fake profiles to provide fraudulent reviews and aim to manipulate the recommender system in e-commerce platforms to promote or demote target products or simply to sabotage the system is known as shilling attack.

Likewise, the consumption of news from social media is highly increased nowadays, so is the spreading of fake news. According to the Pew Research Center [18], 64% of Americans believe that fake news causes confusion about the basic facts of current events, but often, they share news on social media without reading them and checking the validity of the content [19]. Since fraudulent reviews have directly affected the revenue of e-commerce businesses [12], visibility and trustworthiness of platforms [20] and propagation of fake news has a significant impact on political events [7], the stock market [21], terrorist activities [22, 23] and natural disaster [24], it is evident that the problem of false information is ubiquitous and alarmingly crucial to be taken care of. Despite the commitment of social media like Facebook and Twitter and review platforms like Amazon and Yelp to combat this false information, the problem still

continues to persist. Thus, in order to proactively combat false information, it is very important to understand the reason behind the success of its motive, recognize and quantify the impacts made on information systems and develop novel ways of identifying false information and actor responsible for creating and spreading them.

Having said that, false information would not be an issue if people did not fall for it. Unless we understand the motivation behind why false information is succeeding in its intent, we will not be able to find the cure. A more plausible explanation on why people fall for false information is the relative inability of humans to discern false information. Several studies have measured and highlighted that humans are poor at identifying false information [5, 25]. One striking finding across several studies is that the confirmation bias [26], lack of media literacy, and indifferent attitude of humans, together with the difficulty of discerning credibility of information, is the root cause behind the success of deception. Many researchers have studied and measured the impact of false information in terms of different strategical approaches, such as the impact of fake news is measured in terms of user engagement like the number of shares, likes, retweets, and comments. Likewise, the impact of fraudulent reviews is measured in terms of revenue growth and visibility of the business [12, 27, 20]. The sheer volume of false information being observed in online information systems is an obvious cause of concern. It highlights how prevalent false information is, underscoring the importance of knowing how to mitigate them. In order to mitigate false information and its negative impacts, it is very important to know whether the piece of information being shown is real information or not before taking any kind of step. However, because of the human inability to discern false information from truthful information, it is very crucial to develop methods that automatically

identify false information on the Web and social media. Several algorithms have been developed to detect these false information [28, 29, 30, 31, 1, 32, 33, 34, 35, 36, 37, 38, 25, 39, 40, 41, 42, 43, 44, 45]. However, these approaches are limited to binary classification of the false information overlooking the fact that the information is not merely fake or real, it can have different degrees of truthfulness and falsehood.

1.2 Overview and contributions

The goal of this dissertation is to investigate the rationale behind the success of false information and its impact and advance towards mitigation strategies to combat false information. To achieve this goal we formulate three objectives,

- **Objective 1: Investigating why deceptive information is successful.**

Given that several studies show humans are poor at identifying false information [5, 25], the comprehensive study on how humans discern false information, what are the factors that influence humans to think the information is false or real, could shed light on the theory behind the success of false information. Thus, to achieve this objective, we conduct a comprehensive study on how humans assess the credibility of news on social media which we explain in Chapter 3. There are many factors that may help users judge the accuracy of news articles, ranging from the text itself to meta-data like the headline, an image, or the bias of the originating source. In Chapter 3, we analyze the data collected from an online survey where we asked participants ($n = 175$) of various political-ideological leaning to categorize news articles as real or fake based on either article text or meta-data. We also asked participants to provide an explanation for their decision. We provided participants with the varying granularity of information as four categories: news excerpt only, news title, im-

age, and source political bias, news title and image, and news title and source bias. For this, we use political news articles from the FakeNewsNet dataset that consists of news title, excerpt, and associated image and ground truth label as fake or real. We perform quantitative and qualitative analysis on the received responses to investigate how various article elements (news title, image, source bias, and excerpt) impact users' accuracy in identifying real and fake news.

Furthermore, we use a machine-learning model to build an automated fake news detector by using the same type of information provided to the survey participants and compared human versus automated detector accuracy under the same conditions to better understand how humans and computers utilize those elements to determine real and fake news. The obtained result from our analysis reveals that automated techniques were more accurate than our human sample while in both cases the best performance came not from the article text itself but when focusing on some elements of meta-data, i.e. news title and image. We also investigate if the political ideology of participants is playing any role in their decision making and find that political-ideological leaning had little effect, though left-leaning participants were more accurate than right-leaning participants when categorizing an article as real or fake for title and image condition. We also find that adding the source bias does not help humans, but does help computer automated detectors. Our qualitative analysis of open-ended responses as to why participants identified news as fake or real revealed that the image, in particular, maybe a salient element for humans detecting fake news.

- **Objective 2: Measuring the impact of deceptive information in infor-**

mation systems. Most studies about the impacts of fake news have been conducted on real-world datasets [46], e.g., Twitter, however, the study about the impacts of fraudulent reviews thus far includes only simulated deceptive reviewers [47, 48] which fails to provide the evidence about the impacts on a real-world setting. As, the main goal of fraudulent reviews is to impact the popularity of a product in such a way that it can game the system that recommends information and products to the customers to eventually impact the revenue of the products, we study and measure the impact of fraudulent reviews on different recommender systems. Especially, collaborative filtering-based recommender systems are the most popular among the ones used in e-commerce platforms to improve user experience. However, these recommenders are more vulnerable to shilling attacks, i.e., malicious users creating fake profiles to provide fraudulent reviews to manipulate the recommender system to promote or demote target products or simply sabotage the system. Within this context, the findings presented in Chapter 4 demonstrates our empirical analysis to understand the impact of shilling attacks and quantify the impact on various well-known collaborative filtering-based recommender systems. Specifically, we analyze the robustness of five widely-used collaborative filtering-based recommendation algorithms namely - Item-Item, Probabilistic Matrix Factorization, Alternating Least Squares, Bayesian Personalized Ranking, and FunkSVD. Unlike existing works, we conduct our extensive analysis on multiple real-world datasets namely Yelp [49] and Amazon [50] data with ground truth about deceptive reviews. Trends emerging from our analysis unveil that, the performances of considered recommender systems are indeed affected by spam ratings/reviews. We then

segment users based on fairness scores as spammers, benign (mainstream and non-mainstream) users. Non-mainstream users are those whose rating patterns do not align with the majority, i.e., liking what most people dislike and vice versa. We find that recommender systems indeed are affected by fraudulent reviews and in the presence of spammers, recommender systems are not uniformly robust for all types of benign users.

- **Objective 3: Detecting trustworthy entities in the information ecosystem.** After measuring the impact of false information and analyzing how humans assess the credibility of online information, we develop several algorithms in order to automatically identify the credibility of entities in the information ecosystem (i.e. both opinion-based and fact-based). Chapter 4 pinpoints an important standpoint about the types of users in online platforms and social media. It shows that users in opinion-based platforms are beyond what existing researches have been considering i.e. classical binary fraudulent vs trustworthy rather they can be of different types based on their contributions. Keeping this in mind, in Chapter 5, we focus on identifying trustworthy reviewers from fraudulent and uninformative/unreliable reviewers, having known that reviewers may contribute to an opinion-based system in various ways, and their input could range from highly informative to noisy or even malicious. For this, we build a model, DeepTrust, that relies on a deep recurrent neural network that provides embeddings aggregating temporal information: we consider users' behavior over time, as they review multiple products. We model the interactions of reviewers and the products they review using a temporal bipartite graph and consider the context of each rating by including other reviewers' ratings of the same

items. We carry out extensive experiments on a real-world dataset of Amazon reviewers, with known ground truth about spammers and fraudulent reviews. This model outperforms several state-of-the-art models for fraudulent reviewer detection and is able to effectively learn minority classes of users whose behavior is unknown or cannot be learned from the existing traces. We also show that aggregating our model with existing fraudulent reviewer detection models further improves the performance.

This concept of multi-class scenario can also be instrumental in fact-based systems to address the problem of inferring trustworthiness degrees of entities like social media users, news publishers, and pieces of news. Therefore, in Chapter 6 and Chapter 7 we first develop models to characterize and predict fake news spreaders and fake news respectively. Then in Chapter 8 we develop a model to infer the trustworthiness degrees of entities in the news ecosystem. We study the characteristics of fake news spreaders focusing on different attributes such as user writing style, emotions, demographics, personality, social media behavior, and network features. In particular, we leverage these attributes to perform a comprehensive analysis on two different datasets, namely PolitiFact [51] and PAN [52] to investigate the patterns of user characteristics in social media in the presence of misinformation. Specifically, we study the correlation between the user characteristics and their likelihood of being fake news spreaders and demonstrate the potential of the proposed features in identifying fake news spreaders. Likewise, we learn the characteristics of fake news by reproducing the results of Horne and Adali’s work that they conducted on a small amount of data which has been the reference reading for the research community to

understand textual content differences between real and fake news. We validate their findings on larger state-of-the-art datasets with labels provided by professional journalists who have fact-checked the news, namely PolitiFact and GossipCop [53] and BuzzFeedNews [31]. We show that although most of their findings can be generalized to larger political and gossip news datasets, some of the observations are not the same as the trend of news writing is continuously evolving and uncover some new trends highlighting differences between political and gossip news domains. Altogether, we gather ample explicable understanding of the characteristics of fake news and fake news spreaders.

As we discussed in Chapter 5, we implement the concept of multi-class scenario in the fact-based system in Chapter 8 where we use the characteristics of fake news and fake news spreaders learned from Chapter 6 and Chapter 7 to infer the trustworthiness degrees of each entity in the news ecosystem. In particular, we use a supervised graph representation learning approach to model the interactions between entities and extract representation for each entity in a heterogeneous graph representing the news ecosystem. Essentially, we improve the classical Relational Graph Convolution Network (RGCN) model such that it can work with feature vectors of different dimensions and optimize on combined node-specific losses. The generated vector representation takes into account the entire characteristics of the entity as well as the relationship between each entity of the news ecosystem and is used in downstream approaches to identify the credibility degree of a news publisher, a news item, and a user jointly.

We also used the acquired knowledge on characteristics of fake news spreaders in Chapter 9 to propose an approach toward understanding the factors that

motivate a user’s decision to share news on social media. We use a diffusion of innovations theory to model and compare real and fake news sharing in social media with a focus on different levels of influencing factors consistent with the three-level framework social scientists use to characterize social and societal phenomena (micro, meso, macro). We apply that three-level approach to identify factors related to the spread of fake news as they relate to users, the structure of news items themselves, and the networks through which news is circulated.

Each of the proposed methods in all chapters except Chapter 6 (because we only focus on learning characteristics of fake news in this chapter) outperforms the existing state-of-the-art methods in the corresponding domain by overcoming their limitations.

Additional material in Appendix A presents codes generated from responses collected from participants that were used for qualitative analysis in Chapter 3, Appendix B presents the additional experiments for the analysis performed in Chapter 3. Appendix C presents the additional experiments to understand the factors that motivate a user’s decision to share real and fake news items in social media performed in Chapter 9. Appendix D presents the additional application of the model we implemented in Chapter 5 to identify depressed users in online forums.

1.3 Overarching thesis statement

The following statement summarizes the thesis aptly:

Understanding the distinguishable characteristics of entities in information systems and utilizing them to further build effective models to infer the credibility degree of entities involved in the information ecosystem.

CHAPTER 2:

BACKGROUND AND RELATED WORKS

Over the past years, serious concern about false information i.e., fake news or opinion spam, has increased due to the proliferation of user-generated content. Although the problem of opinion spam detection has been historically well studied by many researchers, fake news has gained attention in recent years, especially after the U.S. presidential elections of 2016. Many research works have been devoted towards understanding, automatically detecting, and mitigating the effects of this misinformation. In this chapter, we discuss the research works highlighting some of the plausible works conducted to understand the motivation and reason behind the success of deception, impacts caused by the spread of false information, and detect and debunk false information.

2.1 Motivation and success of false information

The most prominent reason behind the successful influence of false information is the inability of humans to discern false information. Several studies have measured and highlighted that humans are poor at identifying false information. For instance, the survey of Kumar et al. [5] observed that people are only 66% accurate at detecting Wikipedia hoaxes and concludes that both ordinary individuals and the well-trained volunteers with domain knowledge are vulnerable to the false information that mimics

genuine one. Similarly, Ott et al. [25] studied human potency in discerning fraudulent reviews by leveraging Amazon Mechanical Turk (AMT) to create fraudulent reviews for 20 popular hotels of Chicago. They introduced 160 reviews containing both fraudulent and truthful reviews to three humans and observed an accuracy ranging between 53% and 62% in identifying the fraudulent reviews. This, again, advocates that humans are poor at discerning false information. Plotkin et al. [54] followed the same technique of generating data using AMT and collected 1041 truthful and fraudulent reviews on restaurant services. They distributed the task of identifying fraudulent reviews among three groups of participants where the first group was not provided with any additional information, the second group was provided guidelines about spotting fraudulent reviews and the third group was provided with the label showing the quality of reviews. They observed that the detection accuracy of humans was 52% even in presence of cues and overall the detection accuracy of humans was 57%. Similarly, Sun et al. [55] also employed human volunteers to identify synthesized fraudulent reviews from 20 reviews with a balanced number of true and fraudulent reviews for TripAdvisor and observed human accuracy of 52%. Moreover, the advancement in deep learning techniques has also contributed to generating realistic fraudulent reviews indistinguishable from human-written reviews. Yao et al. [56] conducted a user study in order to understand whether humans are able to identify the fraudulent reviews generated using a deep neural network model trained on Yelp review data. A total of 600 reviews (set of 20 reviews for each restaurant containing 0 to 5 machine-generated fraudulent reviews and real reviews written by humans) were provided to the Mechanical Turk workers to identify fraudulent reviews where they observed the precision of 40.6% and 16.2% recall. In addition, the study also showed that humans

are more sensitive to repetitive errors meaning they were able to discern fraudulent reviews that have the same errors multiple times than reviews having small spelling or grammatical mistakes.

Likewise, the news ecosystem has also been affected by false information. People are generally awash with an overwhelming amount of news, and fake news seldom comes with warnings. That is why people find it difficult to judge news veracity correctly. A research by the Pew Research Center [57] showed that only 26% of Americans could accurately classify all provided factual statements, and only 35% could classify all provided opinion statements. This human proficiency in distinguishing false information highly depends on the readers' cognitive attributes [58], thus requiring to understand the strategies that people use to assess the credibility of the information. The study of how people assess online information's credibility requires a multidisciplinary effort as it touches information science, psychology, sociology, communication, and education. Metzger and Flanagin [59] have summarized existing research on the factors that influence people when making credibility evaluation decisions under different categories: site or source cues; author cues; message cues; receiver characteristics; and social interactions. Furthermore, the dual processing model of credibility assessment states that people tend to use two general strategies, namely *analytic* and *heuristic*, that reflect a greater and a lesser degree of cognitive rigor, respectively [60]. The *analytical* strategy suggests that people are using what is provided and what they know in order to produce a reasoned assessment. *Heuristic* implies that people use pre-existing rules that they adapt to the case at hand. This also involves a superficial evaluation of the piece of information where the user's gut feelings are often predominant.

2.2 Impact of false information

The sweeping amount of false information spread has a high potential to cause extreme negative impacts on people, society, and other crucial sectors. Prior encounter with the fact shows that the spread of the falsified information has a significant impact on political events [7], the stock market [21]. For instance, over \$130 billion in stock value was lost in minutes after the false information about President Obama's injury in the White House explosion went viral on twitter [61]. Similarly, [22, 23] and [24] shows the impact of fabricated tweets in terrorist activities and natural disaster respectively. Apart from this, there is a psychological and social impact of fake news. Repeated exposure to fake news induces deep-rooted misconceptions, making it very difficult to correct. Psychological studies show that despite clear retractions of fake information with the correct one, people's perception still remains influenced by misinformation [62] and may even increase the influence among the groups of specific ideology [63].

Many researchers have studied and measured the impact of fake news basically in terms of user engagement like the number of shares, likes, retweets, and comments. An analysis of news leading up to the 2016 election conducted by BuzzFeed, found that there was more engagement with the leading misleading news stories than real news stories [46]. This advocates that fake news can negatively impact the reliability of the entire news ecosystem. As e-commerce platforms rely on users' opinions to help potential customers make informed decisions on products and services and reduce the overload of choices, the financial growth of the business is directly associated with customer reviews. For instance, the study of Harvard Business School [12] quantifies the impact of fraudulent reviews by showing that the revenue growth of

the business in Yelp was increased by 5-9% after the addition of a one-star to the existing rating. Similarly, Lappas et al. [20] studied the impact of fraudulent reviews on the visibility of hotels from TripAdvisor, which affirm that only 50 fraudulent reviews are sufficient in boosting the rankings and pushing the hotel up into the list than any of the competitors. As per the analysis by Washington Post [27], the abundance of fraudulent reviews spuriously provides higher visibility to certain items without merit, making it very difficult to sell anything on Amazon for small businesses without engaging in the dubious act of fraudulent reviews.

2.3 Detection of false information, fake news spreaders and fraudulent reviewers

In order to mitigate the negative impacts of false information, it is very important to know whether the piece of information being shown is real information or not before taking any kind of step. However, because of the human inability to discern false information from truthful information, it is very crucial to develop methods that automatically identify false information on the Web and social media.

Several methods have been proposed to automatically classify a piece of news as real or fake [28, 29]. Many works have considered linguistic and psychological clues from news content (headline, body, image), the social network between the users and their social engagement (share, comment, and discuss given news), or a hybrid approach that considers both [30]. Potthast et. al [31] attempted to classify news as real or fake based on its style as being part of hyperpartisan news, mainstream news, or satire. This study used text-based features such as n-grams, stop words, parts of speech, and readability in determining the hyperpartisan vs. mainstream

articles. Similarly, Horne and Adali [1] considered both news body and headline to determine the validity of news based on textual features extracted. They found out that the content of fake and real news is drastically different. They also unveil that the main claim of the fake news is mostly packed into its title, indicating that the writer of the fake news tends to put much information in the headline. Pérez-Rosas et al. [32] also analyzed the news body content of seven different news domains such as education, business, sport, politics, celebrity etc. Images in news articles also play a role in misleading news detection [33, 34, 35, 36, 37]. Jin et al. [36] included image analysis in fake news detection techniques as fake images are used in news articles to provoke readers' emotional responses. Images are the most eye-catching type of content in the news; a reader can be convinced of a claim by just looking at the news's title and the image itself. Similar et al. [64] used news propagation patterns in social networks for fake news detection. In addition, several studies have been conducted to understand and characterize the users that are likely to spread fake news on social networks. Vosoughi et al. [64] revealed that the fake news spreaders had, on average, significantly fewer followers, followed significantly fewer people, and were significantly less active on Twitter. Shrestha and Spezzano showed that social network properties help in identifying active fake news spreaders [65]. Shu et al. [66] analyzed user profiles to understand the characteristics of users that are likely to trust/distrust fake news. Guess et al. [67] analyzed the user demographics as predictors of fake news sharing on Facebook and found out political-orientation, age, and social media usage to be the most relevant. Shrestha et al. [68] analyzed the linguistic patterns used by a user in their tweets and personality traits as a predictor for identifying users who tend to share fake news on Twitter data [69, 68].

In addition to the approaches using feature engineering, researchers have also been looking into deep learning approaches in order to encode information from news and social context to further harness the detection of fake news and fake news spreaders. SAFE [70] used TextCNN [71] and FakeBERT [72] used Bidirectional Encoder Representations from Transformers (BERT) to encode textual information of news content and visual. Similarly, graph-based approaches using popular Graph convolution Network (GCN) have also been utilized to encode the propagation of news on social media for the detection of fake news [73, 74]. Likewise, Giachanou et al. [75] also leveraged GCN to process the user Twitter feed in combination with features representing user personality traits and linguistic patterns used in their tweets to address the problem of discriminating between fake news spreaders and fact-checkers. Shu et al. developed a model dFEND [76] that utilized bidirectional Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU) to capture word and sentence-level representation from news articles and user comments (tweets) to detect fake news. Further, they used BERT to encode textual information from news items and implemented a two-layered Multi-Layered Perceptron (MLP) to predict fake news [77].

An abundant number of works have addressed the problem of automatically detecting fraudulent reviews. Jindal and Liu [38] first studied the problem and trained machine learning-based models using features based on the opinion content, user, and the product itself. Ott et al. [25] created a benchmark dataset by collecting real reviews from TripAdvisor and employing Amazon Mechanical Turk workers to write fake reviews and used psycholinguistic-based features and text-based features (bigrams) to identify deceptive reviews. However, Mukherjee et al. [39] found out that the linguistic and text based features is not enough to have good performances

on a larger and more realistic Yelp dataset ¹, and found that behavioral features of the user who wrote the review performed very well (86% accuracy) in identifying deceptive reviews. They also reported that the word distribution between fake and real reviews is very different in the dataset by Ott et al., while this is not true in their more realistic Yelp dataset.

As detecting fraudulent reviews from their content is difficult, researchers started focusing on the problem of detecting opinion spammers, rather than fraudulent reviews. Regarding the detection of fraudulent users (or opinion spammers) in opinion-based systems specifically, existing work can be categorized into network-based methods, behavioral-based methods, and hybrid methods combining both network and behavioral properties.

Behavioral-based methods leverage the fact that fraudulent reviewers write many, shorter, positive (4 or 5 stars) and self-similar reviews in short bursts of time [41, 42, 41, 78]. Lim et al. [44] propose ranking and supervised methods called SpamBehavior, exploiting the fact that opinion spammers target a specific set of products and their ratings deviate from the ones of benign users. Hooi et al. [45] proposed the BirdNest algorithm that detects opinion spammers according to the fact that (i) fake reviews occur in short bursts of time and (ii) fraudulent user accounts have skewed rating distributions.

Similarly, network-based algorithms assume to work with a bipartite user-item rating network. Fei et al. [42] discovered that a large number of opinions made use of a sudden burst either caused by the sudden popularity of the product or by a sudden invasion of a large number of fake opinions including some of the features of real users.

¹Yelp filters fake/suspicious reviews and puts them in a spam list. Studies found this filter to be highly accurate in detecting fraudulent reviews [40].

They used this finding to design an algorithm that applies loopy belief propagation on a network of reviewers appearing in different bursts to detect opinion spammers. Rayana and Akoglu [79] also proposed a network-based algorithm FraudEagle and applied loopy belief propagation on a user-product bipartite network with signed edges (positive or negative reviews) and considered metadata (text, timestamp, rating) to assign prior probabilities of users being spammers, reviews being fake, and products being targeted by spammers. Wang et al. [80] proposed Trustiness with three measures (the trustworthiness of the user, the honesty of the review, and the reliability of the store) to be computed on the user-product-store network. Mishra and Bhattacharya [81] proposed the Bias and Deserve algorithm for computing the trustworthiness of a node as a *bias* quantifying the tendency of the node in overestimating or underestimating the rating an item deserves (the higher the bias, the less the trustworthiness of the node). The algorithm also computes a *deserve* score for each item that takes into account the bias of the users that are ranking that item. The bias and deserve scores are computed by a pair of mutually recursive equations. Similarly, Kumar et al. [82] defined the Fairness and Goodness algorithm, which computes a fairness score for each user and a goodness score for each item.

Several other works have been proposed to detect a group of opinion spammers. For instance, CopyCatch [83] leveraged the lockstep behavior, i.e., groups of users acting together, generally liking the same pages at around the same time.

Part II

Investigating why deceptive information is successful

CHAPTER 3:

THAT’S FAKE NEWS! INVESTIGATING HOW READERS IDENTIFY THE RELIABILITY OF NEWS WHEN PROVIDED TITLE, IMAGE, SOURCE BIAS, AND FULL ARTICLES

3.1 Introduction

The problem of misleading or false information disguised as news content (colloquially called “fake news”) has drawn a great deal of renewed attention recently due to the proliferation and popularity of social media as a platform for information diffusion. Fake news has been identified as being more likely to go viral than real news, spreading both faster and wider [64]. Additionally, users are more likely to share headlines they have seen repeatedly than headlines that are novel, even when they know the headline is false [84]. Fake news has become more of an issue, given the world’s political climate and the rampant use of social media and a number of studies have sought to identify and mitigate fake news [85].

Within this context, this research seeks to better understand the accuracy of people and automatic detectors when evaluating fake news with varying granularity of information. Specifically, we seek to understand how the article title, image, source

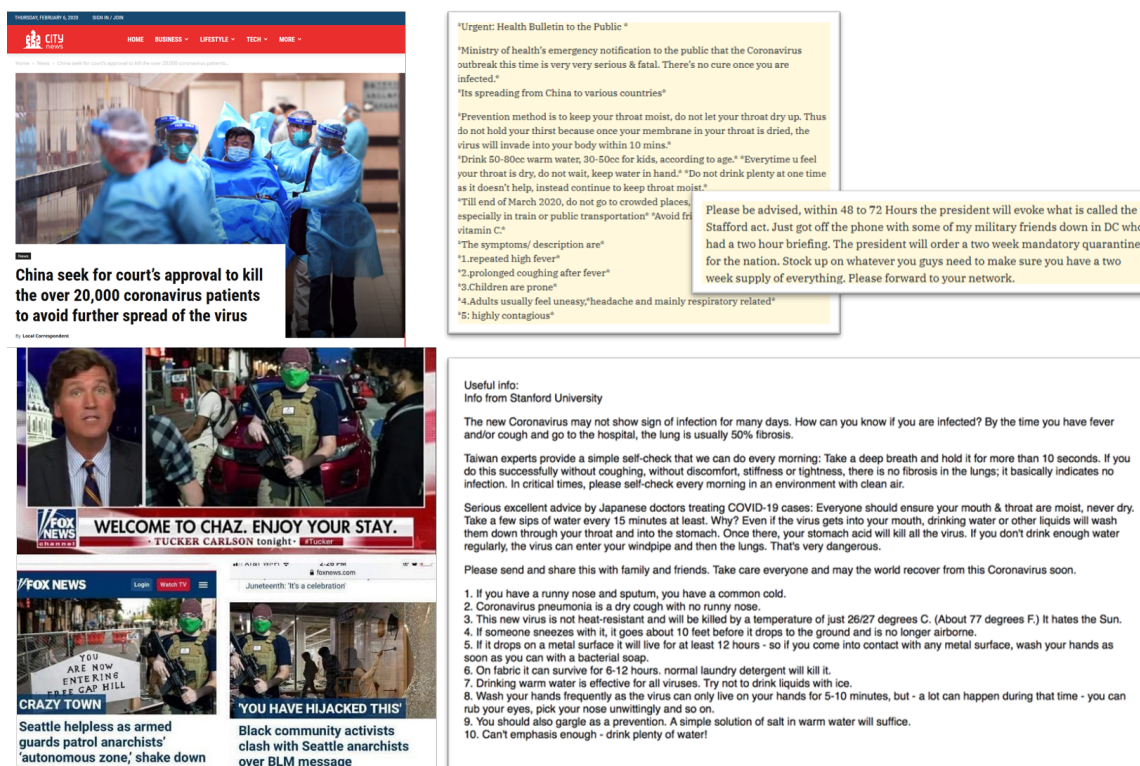


Figure 3.1: Recent examples of false social media posts (as identified by snopes.com) that illustrate use of title and image (T+I) and text excerpts (E).

bias, and text excerpt affect people's accuracy in identifying reliable or fake news. We selected these meta-data elements because this kind of information is readily available and is often shared on social media (see Figure 3.1 for recent examples).

Our investigation furthers research in this area by identifying trends among participants with regards to their ability to identify real and fake news, and initiates a discussion on how various meta-data affect people's ability to identify reliable news as compared and contrasted with automatic detection mechanisms using the same news article meta-data and content information. Specifically, we conducted an online survey where we asked participants ($n = 175$) to categorize news as real or fake and

provide an explanation for their decision. We varied the way the news was presented to the participants and considered four conditions: news excerpt only and three combinations of various meta-data elements including: (1) news title, image, and source political bias, (2) news title and image, and (3) news title and source bias. We used a combination of quantitative and qualitative research methods to compare participants' accuracy in correctly identifying news veracity and investigate the reasoning participants gave to justify their choices. Furthermore, we used machine-learning to build an automated fake news detector by using the same type of information provided to the survey participants (news excerpt and meta-data) and compared human versus automated detector accuracy under the same conditions to better understand how humans and computers utilize those elements to determine real and fake news.

Our research shows that humans are less accurate in identifying fake news when they have only the text of the article itself compared to when they rely on meta-data. In particular, participants were more accurate when they had the picture available (accuracy of 0.62 when news title and image are available versus 0.53 when only the news excerpt is available). Analysis of open-ended participant responses revealed that the professionalism of the image was a helpful heuristic that enabled more accurate judgments. Overall, political ideological leaning had little effect, though left-leaning participants were more accurate than right-leaning participants when categorizing an article as real or fake based solely on the headline plus accompanying image. Automatic detection outperformed humans on identifying fake news articles (overall best accuracy of 0.83 versus 0.62).

3.2 Related Work

The study of how people assess the credibility of online information requires a multi-disciplinary effort as it touches information science, psychology, sociology, communication, and education. Metzger and Flanagin have summarized existing research on the factors that influence people when making credibility evaluation decisions under different categories: site or source cues; author cues; message cues; receiver characteristics; and social interactions [59]. Furthermore, the dual processing model of credibility assessment states that people tend to use two general strategies, namely *analytic* and *heuristic*, that reflect a greater and a lesser degree of cognitive rigor, respectively [60]. The *analytic* strategy suggests that people are using what is provided and what they know in order to produce a reasoned assessment. *Heuristic* implies that people use pre-existing rules that they adapt to the case at hand. This also involves a superficial evaluation of the piece of information where the user’s gut feelings are often predominant.

People are generally not able to correctly judge news veracity. Specific to the news domain, research by the Pew Research Center [57] showed that only 26% of Americans could accurately classify all five provided factual statements and only 35% could classify all five provided opinion statements. Similarly, Horne et al. [86] showed that when both news title and excerpt are provided, humans’ ratings for article reliability were, on average, 6.64 for articles from reliable sources and 5.01 for articles from unreliable sources (on a 10-point scale ranging from “completely unreliable” (1) to “completely reliable” (10)). Several social and psychological theories explain why people are not able to accurately judge news veracity, including the backfire effect [87] and conservatism bias [88] which explain why it is hard for people to revise their beliefs

even when presented evidence against them. Also, people are more prone to trust information that confirms an individual's pre-existing beliefs (confirmation bias [26]) and are more likely to believe a claim simply from reading it multiple times (validity effect [89]). Information processing styles can also affect misperceptions related to news. Those who tend toward online processing (forming judgments and updating attitudes immediately when encountering new information) seem to be more open to correction of misinformation than those who tend toward memory-based processing (where information is stored and then retrieved or sampled at the time of judgment, as needed) [90].

Additionally, the dual processing model is relevant to judging the veracity of news. A recent study showed that individuals who tend toward an analytic strategy (scoring higher on the Cognitive Reflection Test) are better at distinguishing real and fake news when presented with a combination of meta-data including headline, image, byline, and source [91]. This pattern held among people on both sides of the political aisle; a lazier, more heuristic style of thinking rather than ideological leanings seemed to predict difficulty recognizing fake news.

We should not be surprised, then, to learn that users often do not invest the time to fully process information before sharing it, but may instead rely on meta-data such as headline, image, or source. For example, users of a popular social news aggregator, Reddit, can opt to rate content they see through an "upvote" or "downvote", yet research has found that roughly three-quarters of ratings occur without the user actually reading the content; indeed, most users do not read the article they vote on [92]. It is not surprising, then, that user interaction with content, including attention, rating, and engagement (e.g., commenting) are predicted by elements of the title [93].

Likewise, a study of Twitter found that for shortened bitly URLs the majority are shared without being read [19].

Cognitive biases and attentional limitations can come into play not just in judging accuracy but in the decision to share a given news story or not. For example, Pennycook et al. [91] found that accuracy judgments of a story are not strongly influenced by political leanings (though willingness to share a story is). Indeed, they found that most people claim accuracy is extremely important when deciding whether to share a news story. However, they argue that the context of social media can pull users' attention away from that value and toward a stronger weighting of other motives – such as signaling group membership or attracting followers – or a stronger influence by factors like emotional and moral valence [94]. They suggest that the sharing of fake news comes not from a conscious decision to prioritize politics over truth, but from attentional constraints. Bago et al. [95] find results consistent with this account [94]. In one condition, they measured participant accuracy in identifying fake or real news headlines where participants first made a speeded judgment of each headline while under cognitive load (i.e. distracted by a concurrent working memory task) then later deliberated on their earlier judgments under no such constraints. Under these conditions, deliberation did in fact correct previous mistakes made by participants' heuristic system when distracted, and did so regardless of the headline's concordance with political belief.

While some other recent work suggests that asking users to “take their time” and “deliberate” may have little effect on their judgments of fake news headlines [84], Pennycook et al. [91] found that merely focusing users' attention on the concept of accuracy (i.e., making it more attentionally salient) *can* reduce sharing intentions for

false content. Knowing this, social media platforms could integrate such information into their presentation layers in order to focus user attention in a way that makes their existing accuracy preference more salient and more likely to intervene in behaviors such as sharing. For example, platforms could label news that an algorithm categorizes as potentially fake, and in so doing call user attention to issues of veracity. Research by communication scholars has shown that labeling, warnings, or corrective information can weaken the effects of misinformation [96, 97, 98, 99]. However, the primary focus of these researchers was on how effective a correction was for users (e.g., a correction from the CDC versus from a friend).

Several methods have been proposed to automatically classify a piece of news as real or fake [28, 29]. These methods use features extracted from the news content and title [31, 1, 32, 85, 100], associated image [101] (but without considering image emotions or quality as we did in this research), social network context [58], and news propagation patterns in social networks [64].

Horne et al. [86] showed that AI assistance with feature-based explanations improves people’s accuracy of news perceptions while Yaqub et al. [102] examined the role of different news credibility indicators (fact-checking, news media outlets that dispute news credibility, the public, and AI systems) in decreasing the propensity to share fake news and showed fact-checking to be the most effective of the considered indicators.

Several studies have highlighted that humans are generally poorer at identifying false information in comparison with automated detectors. For instance, people were 66% accurate at detecting Wikipedia hoaxes [5] (while the computer achieved 86% accuracy) and people had an accuracy ranging between 53% and 62% in identify-

ing fake reviews [25] (whereas the computer achieved approximately 90% accuracy). While Wikipedia hoaxes and fake reviews are related, there is a difference between them and fake news. Specifically, a fake review is a dishonest individual’s opinion in a context where there is no absolute ground truth, a Wikipedia hoax is false information pretending to be true (and sometimes intended as a joke), while fake news is false or misleading information that is spread deliberately to deceive (see Molina et al. [103] for a taxonomy distinguishing fake news from other content like satire, commentary, misreporting, and so).

Ringel-Morris et al. [104] systematically manipulated elements of Twitter posts to see how those elements affected credibility assessment by users. They found that users perform poorly in accurately identifying truthfulness by content alone (where the length of a tweet is comparable in length to a news headline), and instead are influenced by shortcut information (e.g., name and image/avatar of content poster). The heuristic shortcut information – or meta-data that condenses, distills, and represents – makes it easier to identify misinformation. Whatever the reason, it is more difficult for people to ascertain whether an article is true or not from longer texts.

To understand which elements of an article drive mistakes in discerning fake news from real news, it would be helpful to compare peoples’ accuracy in judging various combinations of meta-data elements and text excerpts. A preliminary attempt in this direction can be found in Zhang et al. [105], who asked six trained readers (three critical thinking instructors and three journalism students) to use specialized tools to annotate news articles based on *content indicators* like ‘clickbait’ title and emotional tone or *context indicators* like external fact checking results, ads, layout, and impact factor of a journal. Annotators had low inter-rater agreement overall, but in both

conditions the researchers found a handful of annotated indicators that correlated with credibility scores given by a few domain experts, more for the *context* condition than the *content* condition (for example, if an article had been fact-checked as false on an external website, that predicted the domain experts finding it less credible). This hints at the possibility that text-based indicators are less reliable at discerning fake news than meta-data and/or external information, but the study used different expert annotators (university instructors vs. journalism students) using different tools in each condition, so it is hard to directly compare the conditions and does not tell us about everyday consumers of news.

The present study builds on previous work by: (1) comparing accuracy while systematically manipulating which information the user has access to (i.e. news excerpt versus various combinations of meta-data), (2) analyzing the specific reasons humans cite for classifying a news item as real or fake under those conditions, and (3) comparing human and computer accuracy at fake news detection under the same conditions (i.e., access to the same type of information).

3.3 Methods

This section describes the dataset we used for the evaluation and then presents the methods used to evaluate the accuracy of people and automated detectors (computers) when provided various news article information.

3.3.1 Dataset: FakeNewsNet

In order to conduct the evaluations, we utilized the FakeNewsNet dataset [58]. While other datasets exist (e.g., [86]), some of those assign the “ground truth” of an article at the source level (whether a source tends to produce true or fake news), whereas

the FakeNewsNet dataset contains news articles (title, excerpt, and associated image) individually labeled as real or fake by Politifact and Buzzfeed fact-checking websites. Some articles did not have all of the elements (title, excerpt, and associated image), so those were removed. This resulted in a set of 384 articles (from the full set of 422 articles). We extracted the article source bias from the MediaBias/FactCheck¹ website which assigns seven degrees of bias: extreme-right, right, right-centered, least-biased, left-centered, left, and extreme-left.

3.3.2 Evaluation by People

This study was conducted using an online “Fake News” survey delivered via Qualtrics. Through this online survey, participants were asked to judge whether news items were real or fake news and then explain the reasoning for their decision. We randomly selected 16 real and 16 fake news articles from the FakeNewsNet dataset described above in Section 3.3.1. The 16 articles were randomly selected and balanced in terms of the number of left-leaning (extreme, left, and left-center) and right-leaning news (extreme, right, right-center) for each category of news used (real or fake).

The participants responded to four different types of questions where we varied the news information provided: title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B); news text excerpt (E). Figure 3.2 shows sample survey questions for each considered condition. Each participant evaluated five articles for each of the conditions T+I, T+B, and T+I+B, and three articles for condition E². As such, each participant was exposed to a total of 18 articles. For each

¹Other datasets exist that categorize the bias of new sources, for example [106]. We checked the Spearman’s rank correlation between the MediaBias/FactCheck scores and the ones provided by [106] and obtained that they are aligned with a correlation of $\rho = 0.90$.

²Participants were presented fewer excerpt only (E) articles due to the additional length it took for them to evaluate that condition. Evaluating the title (T), title and image (T+I) and title, image and source bias (T+I+B) did not take participants very long to evaluate.



<p>Please look at the following news excerpt:</p> <p>A public high school has been accused of indoctrinating Islam into their students. Allegedly, the school has been mandating children profess the Islamic statement of faith, memorize the five pillars of Islam, as well as teaching students that the Muslims faith is stronger than a Christian or Jews. This is all according to lawsuit filed in federal court this past Wednesday. The lawsuit was filed on behalf of John and Melissa Wood with the Thomas More Law Center and the action is being taken against La Plata High School in Maryland. According to John Wood the school banished him from their property when he complained about the Islamic teachings. Richard Thompson, President of Thomas More, said, "Defendants forced Wood's daughter to disparage her Christian faith by reciting the Shahada, and acknowledging Mohammed as her spiritual leader." The Law Center commented that for non Muslims such as Christians and Jews that reciting the Shahada Islamic creed which is their statement of faith is the equivalent of converting. Spokespeople for the Charles County Public Schools have refused to comment other than to say they have not received any such lawsuit yet. However, the Principal Evelyn Arnold, Vice Principal Shannon Morris, and Charles County Board of Education were all named in the lawsuit. The lawsuit says, "During its brief instruction on Christianity, Defendants failed to cover any portion of the Bible or other non-Islamic religious texts, such as the Ten</p> <p style="text-align: center;">Excerpt</p>	<p>Please look at the following news title and image:</p> <p>Trump Silent As Police Credit A Sikh Immigrant With Capturing NYC Bomber.</p>  <p style="text-align: center;">Title + Image</p>
<p>Please look at the following news title and source bias:</p> <p>Obama Pushes One World Government.</p> <p>[Source Bias: right]</p> <p style="text-align: center;">Title + Bias</p>	<p>Please look at the following news title, image and source bias:</p> <p>A Hillary Clinton Administration May be Entirely Run by a FIGUREHEAD .</p> <p>[Source Bias: right]</p>  <p style="text-align: center;">Title + Image + Bias</p>

Figure 3.2: Example stimuli from each condition: news text excerpt (E); title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B). Participants saw 3 examples for the Excerpt condition, and 5 examples for each of the other conditions.

article in each condition, we asked participants to evaluate the veracity of the article as real or fake and provide an explanation for their judgement. For each condition, the articles were randomly assigned from the set of 32 articles we considered for our study. Different participants frequently evaluated the same articles in the same condition but co-occurrences were randomly distributed across participants due to the random assignment. The order of the presentation of the conditions was randomized

and no feedback was given to the participants throughout the experiment.

We recruited undergraduate students ($n = 175$) from a volunteer pool in general education social science courses (Psychology 101) to participate in our survey (107 F, 68 M; mean age 19.5, $SD = 2.4$). The research was approved by the university IRB. Participants were compensated with course credit (volunteering for studies being one option for a research experience requirement). Participants received no training. Participants were just asked to evaluate each article given the information presented to them. Since this was conducted online, we cannot be sure that participants did not look at external sources to help them ascertain whether the item was real or fake; however, the median completion time for the survey was about 28 minutes so extensive lateral reading [107] is unlikely.

3.3.3 Reasoning Given for Evaluations

To investigate the reasoning participants gave for classifying news articles as real or fake, we analyzed open-ended responses that were given by participants as to why they classified each item as real or fake. In doing so we used an inductive approach to generating codes [108]. Specifically, three of the authors first openly coded 160 participant responses, and independently developed codes to describe the observed participant responses on the same sample. The researchers then met and discussed each of the participant responses that were coded. They then compared, discussed, consolidated, and defined the codes that described the data. They then collectively agreed on the codes that should be used for each of the responses in this initial sample. With the codes defined, all three researchers then independently coded another subset of the data, and then met again to confirm the codes identified for each of the responses in the second sample. Codes were used to identify a response

only if at least two of the three reviewers utilized that code, and if all reviewers were in agreement on using that code during the review meeting. The purpose of this coding was to identify the prominent themes that illustrated the approaches used by participants to identify whether news items were real or fake. The results of this analysis are found in Section 3.4.2.

3.3.4 Evaluation by Computer – Automated Detector

We used the same information evaluated by our participants and utilized an automated machine learning-based fake news detector to compare people vs. machine accuracy and the predictive power of the different conditions (T+I, T+B, T+I+B, E). To build the automatic detector, we considered the whole FakeNewsNet dataset which contains 384 news (half real and half fake) to train and test with a 10-fold cross-validation logistic regression model. We considered the following features in input to the model.

To encode text data, i.e., news title and excerpt, we considered features that focus on linguistic style, text complexity, and psychological aspects such as Linguistic Inquiry and Word Count (LIWC) features and text readability measures. LIWC features are grouped into: *linguistic features* such as the average number of words per sentence, rate of misspelling, negations, as well as part-of-speech; *punctuation features* that include the kinds and frequency of punctuation; *psychological features* representing emotional, social, and cognitive processes present in the text; and *summary features* defining the frequency of words that reflect the thoughts, perspective, and honesty of the writer. Another approach is the Rhetorical Structure Theory (RST) which captures the writing style of documents [8]. However, since different studies have shown that the performance of LIWC is comparatively better than RST [58, 8],

we did not use RST in our analysis.

Readability measures how easily the reader can read and understand a text. We use popular readability measures in our analysis: Flesch Reading Ease, Flesch Kincaid Grade Level, Coleman Liau Index, Gunning Fog Index, Simple Measure of Gobbledygook Index (SMOG), Automatic Readability Index (ARI), Lycee International Xavier Index (LIX), and Dale-chall. We chose to use a topic-agnostic approach for processing the text and did not consider topic-dependent features (from the widely used bag-of-words to the most recent BERT [109] deep learning-based approach) as they are not well-suited for the dynamic environment of news where stories' topics change continuously.

To extract features from the images associated with news articles, we considered several tools including (1) the ImageNet-VGG19³ state-of-the-art deep-learning based techniques to extract features from the images [110], (2) features describing face emotions, and (3) features referring to image quality such as noise and blur detection. To capture face emotions in images, we used Microsoft Azure Cognitive Services API to detect faces in an image⁴ and extract several face attribute features. Among all the features extracted, we consider face emotion (anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise) and smile features. Each of these features ranges in [0,1] and indicates the confidence of observing the feature in the image. To capture news image quality to some extent, we computed the amount of blur in an image by using the OpenCV blur detection tool⁵ implementing a method based

³We removed the classification layer of the VGG19 model, and used the last fully connected layer of the neural network to generate a vector of latent features representing each input image. Moreover, we used PCA to reduce the number of extracted features to 10.

⁴<https://docs.microsoft.com/en-us/azure/cognitive-services/face/quickstarts/csharp>

⁵<https://www.pyimagesearch.com/2015/09/07/blur-detection-with-opencv/>

on the Laplacian Variance [111] along with noise level of face pixels provided by Microsoft Azure Cognitive Service API. We also used the article source political bias as identified by MediaBias/FactCheck as a feature to encode bias.

To avoid overfitting, from this set of features, we selected the top-30 most important title text features, the top-30 most important excerpt text features, and the top-10 most important image features via univariate feature selection. More details about the features and methods are explained in our additional work in Appendix B.

3.4 Results

Herein we present the comparisons of the conditions (varying levels of article information) when evaluated by people and a computer.

3.4.1 Comparing Conditions Evaluated by People

On average across all of the conditions, participants agreed with their assessments of real and fake news 69.8% of the time (min=67.6%, max=71.7%; with a mean standard deviation of 12.8%). Below we present the results of participants' judgments with regards to overall accuracy with respect to the various news elements they were provided, as well as a comparison by participant's ideological leaning.

Accuracy

Participants completed multiple examples per condition so we used their average accuracy per condition as the dependent variable. The summary of accuracy for each condition is given in Figure 3.3. We used a Friedman's test (non-parametric ANOVA for related samples) to compare the overall accuracy between the four conditions (T+I, T+I+B, T+B, E). There was a significant difference between con-

ditions, $X^2(3) = 14.198, p = 0.003$, so we followed up with pairwise comparisons using a Wilcoxon signed-ranks test for matched samples and Bonferroni correction for multiple comparisons. As reported in Table 3.1, we found that participants were significantly more accurate in the title+image (T+I) (accuracy of 0.621) and the title+image+bias (T+I+B) (accuracy of 0.617) conditions than in the title+bias (T+B) condition (accuracy of .559) or excerpt (E) condition (accuracy of 0.533).

Table 3.1: Comparison of accuracy between conditions: news text excerpt (E); title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B). Note: with Bonferroni correction, only a p-value < 0.0083 would be significant at a family-wise alpha level of 0.05.

	E	T+I	T+B	T+I+B
E		$z = -3.270$ $p = 0.001^*$	$z = -1.325$ $p = 0.185$	$z = -3.186$ $p = 0.001^*$
T+I			$z = -2.881$ $p = 0.004^*$	$z = -0.095$ $p = 0.924$
T+B				$z = -2.751$ $p = 0.006^*$

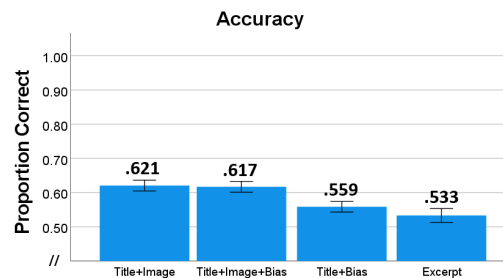


Figure 3.3: Mean accuracy for each condition. Error bars are standard error of the means (SEM).

Comparison by Ideological Leaning

Few participants self-identified as extremely left or right with regards to political ideological leaning (see Figure 3.4a), so we collapsed political ideological leaning into three groups (left-leaning $n = 58$, neutral⁶ $n = 50$, and right-leaning $n = 66$) and then compared the accuracy of those groups in each of the conditions using a Kruskal-Wallis test (non-parametric independent ANOVA). Political group did not relate to accuracy in T+B, T+I+B, or E conditions (all p 's > 0.5) but there may have been a difference in the T+I condition ($p = 0.036$, not significant after Bonferroni correction). Pairwise comparisons using a Mann Whitney U test showed that left-leaning participants were significantly more accurate than right-leaning participants in the Title+Image condition, as illustrated in Figure 3.4b (accuracy of 0.676 vs. 0.570, $p = 0.011$, significant after Bonferroni correction; the other two comparisons (left vs. neutral and right vs. neutral) were not significant).

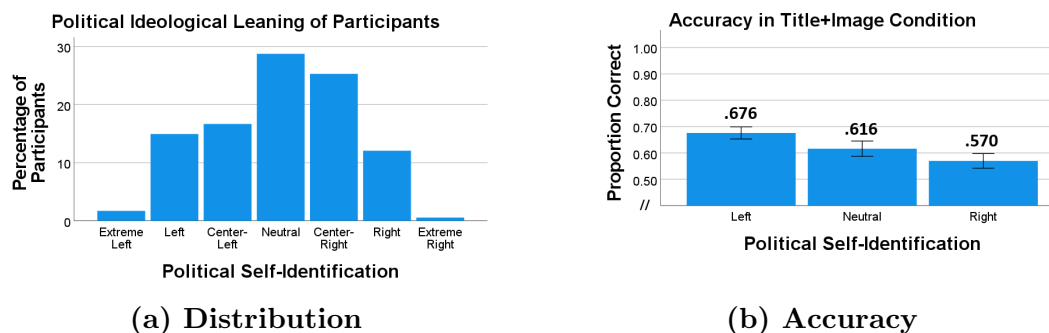


Figure 3.4: (a) Distribution of participants' self-identified political ideological leaning; (b) Accuracy in Title+Image condition for each political group. Accuracy in other conditions did not differ significantly between political groups. Error bars are SEM.

⁶Some established scales use *moderate* instead of *neutral* [112], we acknowledge this as a potential limitation, however this was not central to our analysis as the accuracy of participants was not significantly different from other groups.

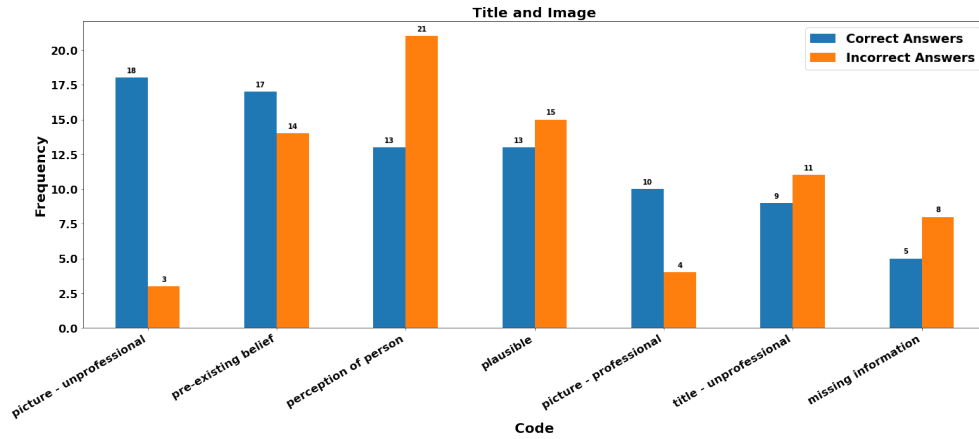
3.4.2 Reasons for Participant’s Judgement of Real or Fake

Given that the highest accuracy was in the title and image (T+I) condition and the lowest accuracy was in the text excerpt (E) condition, we analyzed open-ended responses that were given by participants as to why they classified each item as real or fake from these two conditions. The inductive approach that we used to generate and assign codes was described in Section 3.3.3. In total, the coded data comprised 320 participant responses as to why they identified the news item as real or fake. Half of these were from the title and image (T+I) condition and half from the text excerpt (E) condition.

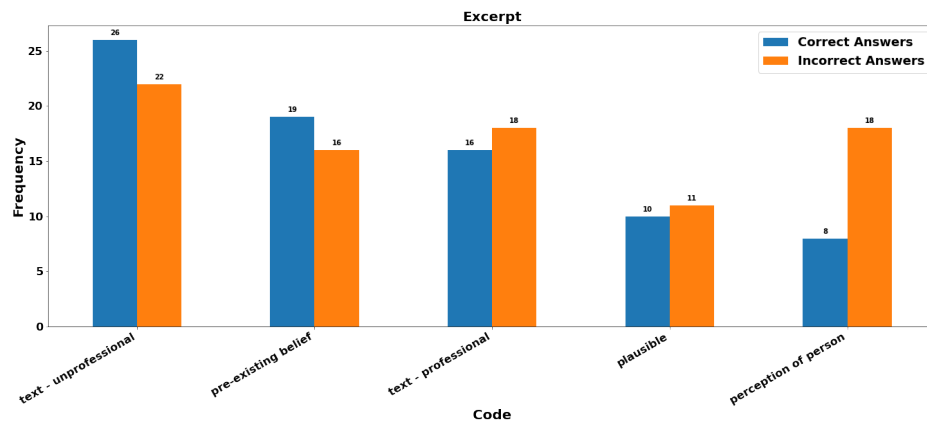
The codes included several dichotomous pairs including plausible/implausible, familiar/unfamiliar, professional/unprofessional (for various news items: title, picture, text), as well as other items including an observation surrounding the emotion the news item intended to evoke, and whether the judgment was based in part on their perception of a particular person, a pre-existing belief, that there was missing information, or whether they simply guessed or were unsure. Table A.1 in Appendix A lists the codes, a description of the code, and provides illustrative representative examples. The five most frequently used codes for the T+I and E conditions are indicated in Figure 3.5.

In the T+I condition, open-ended responses that referred to elements of the image were more often correct than incorrect in accurately labeling the news items as real or fake. For example, correct responses often came with reasoning such as: “Pictures look fake and photoshopped”, “the picture seems pretty neutral”, and “something seems off about the picture”. On the other hand, open-ended responses with more subjective elements (those relating to the respondent, like preconceptions about the

person in the news story) were more often incorrect than correct in labeling the news items as real or fake. For example, incorrect responses often came with reasoning



(a) Title + Image – Note that for this condition, there are three codes shared within the top five most frequently used codes for correct and incorrect responses. Codes are in decreasing frequency order for correct responses.



(b) Excerpt – Note that for this condition, the top five categories are the same for both the correct and incorrect responses. Codes are in decreasing frequency order for correct responses.

Figure 3.5: The five most frequent codes for the Title and Image (3.5a) and Excerpt (3.5b) conditions differentiated by when participants responded correctly and incorrectly.

such as: “Sounds like something the Clintons would do”, “Given that liberals don’t tend to be huge fans of Trump this seems like a legitimate even”, and “Obama should not say this sort of stuff”.

In the text excerpt (E) condition, the pattern in open-ended response themes was less clear, probably reflecting the greater challenge participants faced with this condition (having the lowest accuracy overall). Once again, responses that referred to preconceptions about the person in the news story tended to show up more in incorrect responses than correct ones. For example, “[Hillary] isn’t that immature to do something like that”, “Donald Trump tends to have lots of fake news”, and “I believe this is true because Hillary has been caught in a lot of issues so this doesn’t surprise me”. However, the other most common themes in open-ended responses (how professional or unprofessional the writing was, plausibility, or references to pre-existing beliefs) did not clearly correlate with accuracy.

3.4.3 Comparing Conditions Evaluated by Computer

The computer detector’s accuracy was evaluated and compared based on being provided the various news elements (title, image, source bias, and excerpt). This is analogous to the comparison performed on the human identifications in Section 3.4.1.

Accuracy

Table 3.2 reports the accuracy achieved by the above described automatic detector measured by using a 10-fold cross-validation. As can be seen, the combination of title, image, and bias features achieves the best accuracy of 0.83, while news excerpt features are the least accurate (accuracy of 0.71).⁷

⁷The focus of this work was not to build the best classifier, but rather to see how a machine does in comparison to people evaluating using a good classifier.

Table 3.2: The accuracy of a computer model in identifying real and fake news when using: excerpt (E); title and image (T+I); title and source bias (T+B); title, image, and source bias (T+I+B).

Condition	Features	Accuracy \pm STD
E	LIWC & readability on excerpt	0.71 ± 0.07
T+I	LIWC & readability on title + vgg19 & emotions & quality on image	0.78 ± 0.08
T+B	LIWC & readability on title + bias	0.81 ± 0.07
T+I+B	LIWC & readability on title + bias + vgg19 & emotions & quality on image	0.83 ± 0.05

3.5 Discussion

In this section, we discuss the results with observations regarding: human accuracy when presented with various combinations of meta-data or an excerpt, the need and value of meta-data, the reasons for participants' judgments of news veracity (real or fake), that source bias seems to not help humans in their determination, and potential reasons for the effect of political leaning in the T+I condition. Also, we discuss the comparison between human and machine evaluations, and examine and explain misclassifications (for both humans and our automated detector).

As can be seen in Figure 3.3 (humans) and Table 3.2 (computer/automatic detector), machine-learning-based techniques are dramatically more accurate than humans. This is not too surprising as automatic detectors have been shown to be more effective in identifying other forms of false information such as fake reviews and Wikipedia hoaxes [25, 5], even though Kumar et al. show that this may not necessarily be true

when both humans and computers have access to the same information (66% for humans vs. 47% for the automated detector) [5].

It is of note that accuracy is worse for both humans and the computer when looking at just the excerpt text, which indicates the importance of the other information (title, image, and resource bias) in identifying fake news. This fits with and extends the results of previous work showing that the meta-data of Twitter posts can influence reader’s perception of credibility more than the content alone [104]. Indeed, eye tracking data suggests that time spent looking at meta-data such as headline, byline, and timestamp predicts discernment of fake from real news [98]. Other work has demonstrated lasting improvement in fake news discernment when users are trained to look for problematic elements of article titles [113], hinting that automated tools may provide more assistance to users if they can point to specific elements (T+I, T+B, T+I+B, E) rather than a holistic flagging of the entire article.

According to our participants’ reasoning, an excerpt with a neutral tone, quotes, and statistics is perceived as professional, while an emotional tone is perceived as unprofessional (see “Text Professional” and “Text Unprofessional” codes in Table A.1). However, by analyzing the excerpt of real and fake news from the FakeNewsNet dataset with text-processing techniques, we did not find that, on average, the real news in our dataset had a non-emotional tone and more statistics.⁸ In analyzing the text we did find that the real news in our dataset had more quotes than the fake news on average (2.76 vs. 1.61, $p < 0.001$). Thus, we observe that fake news can more readily deceive users if it is written with a neutral tone and contains statistics and quotes. Additionally, we identified that participants’ prior perception of the people

⁸We considered emotional tone and number (as a proxy for statistics) features from Linguistic Inquiry and Word Count (LIWC).

being reported on more frequently led them to inaccurately identifying whether the news was real or fake. Interfaces could highlight that news about people can lead users to inaccurately identify news and urge them to look at all elements or highlight specific elements that could better help readers discern its veracity (e.g., the emotion found within the headline, the presence of confirmed statistics, and cited quotes).

Previous studies investigating discernment of fake and real news have presented multiple meta-data elements together (e.g., displaying headline, image, and source information for every item [91]) whereas our study begins to tease apart which elements are most relevant. Also we compared human vs. computer accuracy with varying combinations of news elements (T+I, T+B, T+I+B, E) which is also novel. Horne et al. [1] found evidence that the news title is more informative than its excerpt, but, to the best of our knowledge, there is no study comparing people's accuracy in judging combinations of news meta-data elements and excerpt.

While not conclusive from our data, including the source bias did not result in significantly higher accuracy if participants already had the title and the image. There is the possibility that reporting the news source bias does not assist people in their determination of reliability. This could be due to mistrust of the labeled bias and a tendency to over-trust sources that are concordant with their political affiliation (confirmation bias [26]). On the contrary, by adding the source bias in our automatic detector, we increased the accuracy from 0.78 (T+I) to 0.83 (T+I+B). Indeed, several studies in the field of journalism have theorized a correlation between the political bias of a publisher and the trustworthiness of the news content it distributes [114, 115].

Like Pennycook et al. [116], we did not find a large effect of political ideological leaning on accuracy, with the exception of a possible difference in the T+I condition.

While limited to only one condition, this is an important case to investigate further as many social media shares include the title and the image associated with the article. Lazer et al. suggest fake news may be more of an issue for those who are right-identified [117], and that conservatives are more suspicious of fact-checking sites.

Carraro et al. suggest attentional mechanisms are different in liberals versus conservatives, such that negatively valenced information draws the attention of, and impairs the performance of, conservatives more than liberals [118, 119]. This may be one explanation as to why right-leaning participants appear to show worse accuracy in the T+I condition. If the salient or influential aspects of fake news images come from valence (especially negative valence), that could diminish processing in right-leaning participants when fake news images are present.

Beyond apparently undervaluing the source bias as explained above, we noted in Section 3.4.2 that people mistakenly judge the veracity of news when they focus on more subjective elements (those relating to the respondent, like preconceptions about the person in the news story). On the other hand, the machine focuses more on objective elements of the news and is not influenced by reader’s biases. To better understand where the automatic detector made mistakes, we used LIME⁹ – a state-of-the-art technique [120] – to explain the reasons for the false positive and false negative instances. Like our qualitative analysis for humans, we applied this technique to the title and image (T+I) and excerpt (E) conditions. In the T+I condition, the machine mistakenly classified real news as fake (false positive) because the use of punctuation

⁹LIME is a technique that explains single instances by creating an interpretable representation that is locally faithful to the classifier. For each instance to explain, LIME computes a local linear classifier and uses the feature weights of the local classifier to assign an importance to the features. The most important features are the ones who explain the label assigned by the global classifier to the given instance.

(parentheses and dash), negations, and male related words in the title was more similar to the style of fake news (fewer parentheses, dash, negations, and male related words than other real news in the considered dataset). Also, the machine mistakenly classified fake news as real (false negative) because the use of exclamations, stop words, religion, death, sexual, and tentative related words in the title was similar to the style of real news (fewer use of exclamations, religion, death, sexual, and tentative related words and more use of stop words than fake news in the considered dataset). We did not observe a clear pattern in the explanation of false positives and false negatives when we considered the excerpt (E) condition, which is similar to how participants' open-ended responses in this condition did not correlate with their accuracy.

Moreover, we also investigated whether humans and the automatic detector made the same mistakes. We considered a piece of news as human mistake if $\leq 60\%$ ¹⁰ of the participants who answered that piece of news got its credibility right, while automatic detector's mistakes are given by false positives and false negatives. We found that, in the T+I condition, out of all the mistakes humans made, 37.5% of them were also mistakenly classified by the automatic detector, while, for the E condition, out of all the mistakes humans made, 38.1% of them were also made by the automatic detector. Thus, in both conditions, the majority of humans' mistakes can be corrected by assisting uncertain or confused humans with an automated detector.

¹⁰We considered humans' accuracy in the range 50% to 60% as a condition where humans are confused or undecided, hence we included it as a mistake.

3.6 Limitations

One potential limitation of our study is that the automated detector was trained on 384 articles whereas the analogous “training” for human participants would be an unknown and heterogeneous amount of past experience with many years of news items. Thus, the comparison between AI and human performance can never be a ‘fair’ apples-to-apples comparison despite both having access to the same types of information in a given condition (e.g., title+image). However, in no way does it undermine the usefulness of machine learning models to assist humans in discerning real and fake news, even if machine learning models solve the problem differently than a human brain.

One minor limitation is that our political leaning question used “neutral” in the central position of the political leaning scale instead of the more frequently used “moderate”. We acknowledge this may have misled participants. It is possible some respondents just assumed it was similar to moderate (as in the question it was shown between “left-centered” and “right-centered”). However, others may have interpreted it to mean not ideological. This was not central to our analysis however, as the accuracy of participants in this category was not significantly different from the one in other groups.

Furthermore, older readers are more likely to share fake news [121], so many studies have focused on older users, whereas the present work focuses on a younger sample. Our results may or may not generalize to older individuals; but young people use social media at a higher rate than older individuals [122] so this subgroup will likely become increasingly important to study. Indeed, our results show that even a relatively young sample is far from accurate at discerning fake from real news.

3.7 Conclusion & Future Work

In this paper, we present our findings from a study of people’s ability to determine whether news is reliable or fake when given varying combinations of basic information (title, image, source bias, and/or an excerpt from the article). The results show that participants are less accurate when they have only an excerpt and more accurate when provided a title and image, and that for the T+I condition left-leaning participants may be more accurate than right-leaning participants. Our qualitative analysis of responses as to why participants identified news as fake or real revealed that participants’ perception of the person being reported on more often misled them, than helped them in accurately identifying the news as real or fake. While overall people were less accurate than an automated detector in identifying the reliability of news articles, both humans and automated identification was improved when provided meta-data (e.g., title, image, source bias).

Implications of this work could guide designers to focus their attention (and that of their users) to various elements of social media posts. Automated detectors could utilize some of the categories identified in our qualitative analysis that helped humans more correctly identify fake news. To help users better identify fake or real news, interfaces can draw users attention to text professional/unprofessional and image professional/unprofessional. Additionally, platforms may want to prioritize labeling and providing assistance to humans with regards to excerpts since they are the hardest for users to accurately identify as real or fake. This research can inform how computers may be able to train people how to identify the elements of fake news. Beyond just adding a “disputed” label, automated systems could identify or annotate specific credibility indicators [105] (e.g., emotionality of the headline, relevant elements of the

image) to help train readers what to look for. Alternatively, platforms could institute a hybrid approach whereby potentially misleading news is signalled to users during their browsing, and the user can choose to expand the labeling to see which elements the algorithm identifies as indicative of real or fake news and its confidence.

Future work will utilize a larger sample size of participants and giving them a larger variety of articles, allowing us to take a closer look at other predictors of accuracy in fake news detection (e.g., media diet and additional demographic information) and measuring additional depending variables (e.g., likelihood to share a news article). Also, as people were able to use the non-professionalism of the image (too emotional, photoshopped, etc.) to make more accurate judgments, we will further investigate using these features to increase the accuracy of automated detection algorithms on larger datasets. Additionally, we will consider the open-ended explanations collected from our participants to train an algorithm which is able to explain to users why they are mistakenly judging a piece of news.

Part III

Measuring the impact of deceptive information in information systems

CHAPTER 4: AN EMPIRICAL ANALYSIS OF COLLABORATIVE RECOMMENDER SYSTEMS ROBUSTNESS TO SHILLING ATTACKS

4.1 Introduction

The effect of shilling attacks on recommender systems, where malicious users create fake profiles so that they can then manipulate algorithms by providing fake reviews or ratings, have been long studied [123, 124, 125]. So far, recommender system researchers have: (1) Characterized and modeled recommender system shilling attacks (where malicious users insert fake profiles to manipulate recommendations), (2) Defined new metrics to quantify the impacts of these attacks on commonly used recommender systems, and (3) Applied a *detect + filtering* approach to mitigate the effects of spammers on recommendations. Nevertheless, we observe from the literature that the analysis thus far has focused on assessing the robustness of recommender systems via *simulated attacks* [47, 48]. Unfortunately, there is lack of evidence on what is the impact of fake reviews or fake ratings in a *real-world* setting.

In this chapter, we present an analysis conducted to understand the influence of

fraudulent reviews on the recommendation process in real-world scenarios. We do this through a study of known datasets with gold standards in different domains and several commonly-used recommendation algorithms. Specifically, we utilize data from two widely-used e-commerce platforms, Yelp! and Amazon. As a preliminary analysis, we considered probabilistic MF [126] to analyze its robustness to shilling attacks [127]. Motivated from the findings we observed in our analysis, we extended our analysis to other recommendation algorithms. Among various recommendation algorithms, we consider collaborative filtering-based approaches as they are the most efficient and popular recommenders in such platforms. Thus, we focused our exploration on the robustness of these algorithms to shilling attacks.

The main contribution of this chapter is two-fold. First, we analyze the performance of widely used five collaborative filtering-based algorithms in presence of spammers and compared them when spammers are removed. By doing so we seek to answer whether shilling attacks affect the robustness of the considered recommender algorithms. Second, we investigate if there is a specific user group (non-mainstream users) that are affected more than others (mainstream users) by spammers.

Our results are validated by an empirical evaluation using classical measures for evaluating predictive and top-N recommendation strategies. We show that RMSE scores decrease and NDCG@5 scores increase when removing spammers in the majority of the considered algorithms and datasets. This serves as indication that the performance of considered collaborative-filtering based recommender algorithms are indeed affected by spam ratings/reviews. Further, a deep investigation to quantify the effects of spammers on recommendations received by certain groups of users lead us to conclude that, for the Yelp! datasets removing spammers improves the predictive

ability (RMSE) of all the considered recommender algorithms regardless of the type of users, i.e., mainstream or non-mainstream. In the case of Amazon datasets, we observed a trend where removing spammers lessen the predictive ability for mainstream users based on RMSE whereas improves for non-mainstream users according to both RMSE and NDCG@5. Therefore, non-mainstream users whose rating behavior does not align with the majority of the users are the most affected ones by spam ratings for Amazon datasets. Overall, we observed that 25%-29% of benign users in Amazon datasets are users who would not be equally satisfied by recommenders affected by shilling attacks. Thus, recommender algorithms are not uniformly robust for all types of benign users in presence of spammers ratings/reviews.

The rest of this chapter is organized as follows. In Section 4.2, we summarize related work; we then outline the dataset, algorithms and evaluation strategies used in our empirical analysis in Section 4.3. In Section 4.4 we report on our results and, finally, conclusions are drawn in Section 4.5.

4.2 Related work

Collaborative filtering-based recommender systems are widely used to provide recommendations to users in opinion-based systems, yet they are vulnerable to shilling attacks [128, 124]. These attacks consist of fake user profiles injected into the system with the goal of providing spam ratings or reviews to promote or demote specific products. While some Shilling attacks promote the recommendations of certain attacked items (referred to as the push attack), others might demote the predictions that are made for attacked items (referred to as the nuke attack) [47, 129]. Previous work has defined several attack strategies including random, average, bandwagon, love/hate, segmented, and probe attacks. These strategies differ in the way fake profiles choose

filler items, i.e., other rated items chosen beyond attacked items to camouflage the fraudulent behavior. More sophisticated attacks have been recently proposed [128], for instance the one by Fang et al. [130] looks at how to choose filler items to recommend an attacker-chosen targeted item to as many users as possible. In the field of machine learning, many efforts have been devoted over the years to develop techniques for automatic detection of such fraudulent profiles, the techniques presented in [131] and [132] are among the most recent ones. In the field of recommender systems, researchers have focused on studying the effects of such shilling attacks mainly on collaborative filtering-based recommenders since early 2000s [133, 47] and developed strategies to make such algorithms more robust to shilling attacks [124]. We highlight, for examples, outcomes of the research conducted by Seminario and Wilson [134, 135] who explicitly look at power user and power items attacks, i.e., attacks targeting influential users and items, respectively, within collaborative filtering-based recommender systems.

Most recently, the concept of differential privacy have been explored to make matrix factorization based collaborative filtering recommender algorithms more robust [136]. The vulnerability of deep-learning-based recommender systems to shilling attacks has been studied in [137]. In particular, Lin et al. introduce a framework that consider complex attacks aimed towards specific user groups. On a different perspective, Deldjoo et al. [138] explore dataset characteristics to explain an observed change in the performance of recommendation under shilling attacks.

Our work add to this body of knowledge by exploring the robustness of collaborative recommender systems to shilling attacks by using real-world data with spam reviews ground truth, as opposed to attack simulation and investigating if some user are more

vulnerable than others.

4.3 Experimental Settings

As previously stated, our goal is to analyze commonly-used memory-based and model-based collaborative filtering-based recommendation algorithms robustness to shilling attacks using a number of datasets with ground truth on spam reviews. In the rest of this section, we describe the experimental protocol for our analysis.

4.3.1 Datasets

For analysis purposes, we rely on four datasets (described below) produced based upon data from two well-known e-commerce platforms: Yelp! and Amazon.

Dataset	Users	Items	Ratings	Spammers
<i>YH</i>	5,027	72	5,857	14.92%
<i>YR</i>	34,523	129	66,060	20.25%
<i>AB</i>	167,725	29,004	252,056	3.57%
<i>AH</i>	311,636	39,539	428,781	4.12%

Table 4.1: Details on the datasets considered for our analysis.

Yelp! We consider Yelp! reviews from two domains: hotels (**YH**) and restaurants (**YR**) [49]. Yelp filters fake/suspicious reviews and puts them in a spam list. A study found the Yelp filter to be highly accurate [139] and many researchers have used filtered spam reviews as ground truth for spammer detection [140, 141]. Spammers, in our case, are users who wrote at least one filtered review. We removed users who rated same products multiple times and reviews with 0 rating.

Amazon We also consider Amazon reviews from two domains: beauty (**AB**) and health (**AH**) [50]. In this case, we define ground truth based on helpfulness votes

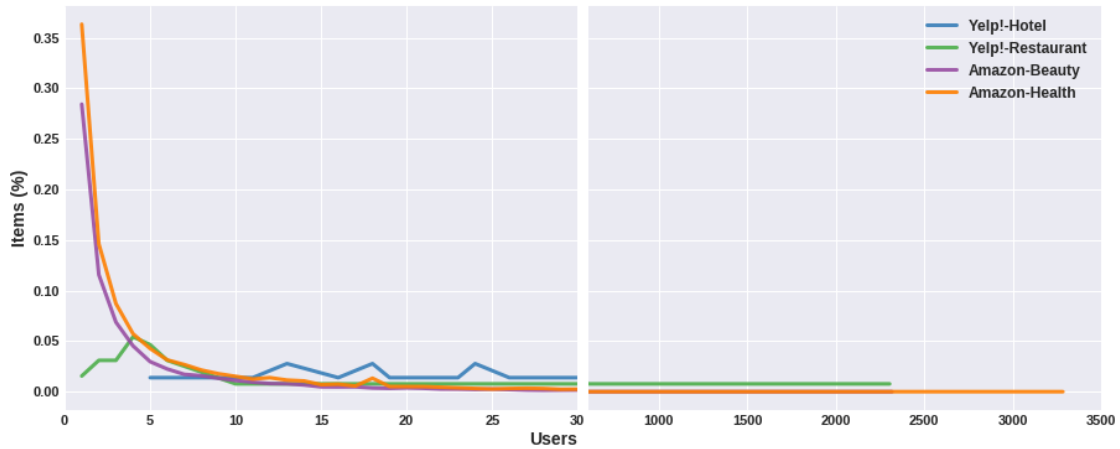


Figure 4.1: Rating distribution across the four datasets considered in our analysis.

following the approach suggested by [131] and based on the findings provided by Fayazi et al. [142]. Thus, we treat as a spammer every user who wrote at least one spam review. We define a review as spam if the rating is 4 or 5 and the helpfulness ratio is ≤ 0.4 .

We provide descriptive statistics for the four datasets in Table 4.1. It is important to note that rating distribution is not similar across the datasets. As illustrated in Figure 4.1, rating trends from AH are dissimilar to the other counterparts, with a vast number of users rating only 1 item. Moreover, the rating distribution of benign users vs. spammers on attacked products (i.e., products receiving at least one spam review) is captured in Figure 4.2. From this figure it emerges that benign users and spammer counterparts exhibit similar rating behaviour in YR, AB, and AH, whereas in the case of YH, spammers noticeably assign ratings of ‘1’ more often than benign counterparts, the opposite is true for ratings of ‘4’.

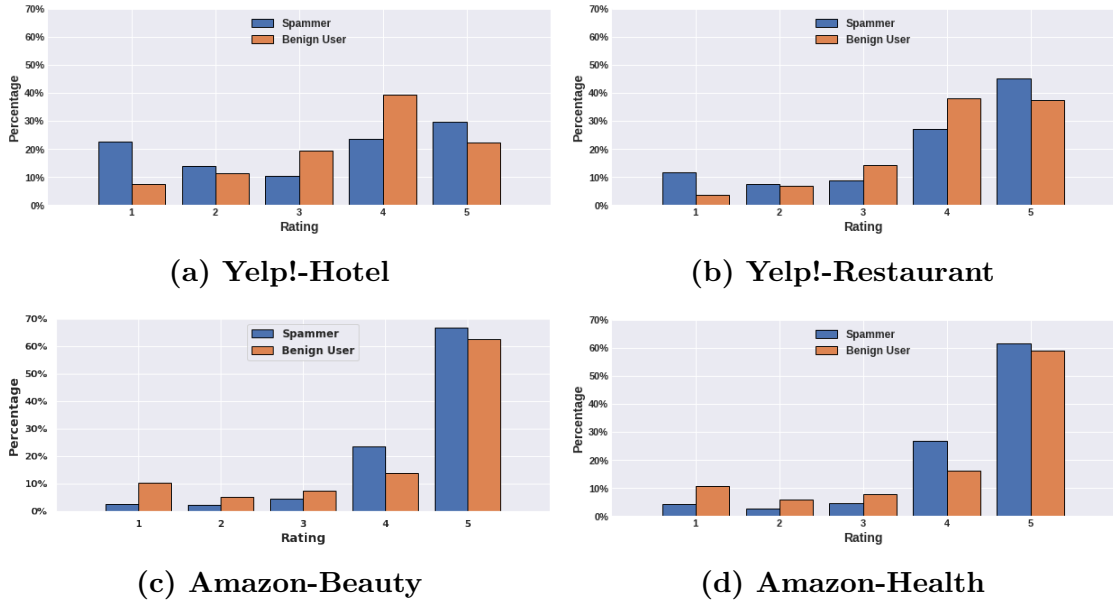


Figure 4.2: Rating distribution for attacked products by spammers and benign users.

4.3.2 Algorithms

We focus our analysis on well-known and widely-used collaborative filtering-based recommendation algorithms implemented using Lenskit for python [143], with the exception of Probabilistic Matrix Factorization, for which we relied on the implementation provided by Mnih et al. [126].

Item-item [144] is the popular item-based collaborative filter algorithm. It utilizes an item-item matrix to determine the similarity between the target item and other items (neighbors). We used this algorithm with 20 neighbors and cosine similarity as similarity measure.

Probabilistic Matrix Factorization (PMF) [126] is a commonly-used latent factor based recommendation algorithm. Specifically, probabilistic matrix factoriza-

tion decomposes the sparse user-item matrix to low-dimensional matrices with latent factors to generate recommendation. We used this algorithm with 40 latent factors and 150 iterations. This algorithm is known from its accuracy, scalability and dealing with sparsity.

Alternating Least Squares (ALS) [145] is a matrix factorization based algorithm designed to improve recommendation algorithms performance in large-scale collaborative filtering problem. This algorithm gain recognition following its success on the Netflix Challenge [146, 147]. In our case, we consider 40 latent factors, 5 damping factors and 150 iterations for our experiment.

Bayesian Personalized Ranking (BPR) [148] is a rank-based matrix factorization algorithm, with 40 latent factors, 5 damping factors and 150 iterations for our experiment. Note that as top-N recommendation algorithm, i.e., based on rating information is in the form of implicit feedback [148, 149], BPR scores items, but does not produce rating predictions. Thus, we are forced to exclude BPR from RMSE-based analysis discussed in Section 4.

FunkSVD [150] is the well-known gradient descent matrix factorization technique with 40 latent features and 150 training iterations per feature.

4.3.3 Evaluation Framework

By following the classical evaluation framework for shilling attacks on recommender systems [47], we measured the performances on the original dataset (with spammers) and when we remove all the spammers (shilling attack), using well known performance metrics. In all cases, we performed 5-fold cross-validation. We tested whether differ-

ences in the metric values with and without spammers were statistically significant by using a paired t-test.

Metrics. For assessment, we turn to *Root Mean Square Error (RSME)* and *Normalized Discounted Cumulative Gain (NDCG)*, which are classical measures for evaluating predictive and top-N recommendations. We also consider measures explicitly defined to quantify the impact of spammer attacks on recommenders: Prediction Shift (PS), which captures the average absolute changes in predicted ratings for attacked items and Hit Ratio (HR), which considers if attacked items are promoted to user top-n recommendations (cf. Burke et al. [47] for formal definitions of these metrics).

We first examined recommender performance by considering all users in the respective datasets. We then segmented users into *fairness* categories as computed by the Fairness and Goodness algorithm described in the next paragraph in order to allow for more in-depth explorations. By segmenting users based on fairness scores, it is possible to identify mainstream and non-mainstream users. The latter, are those whose rating patterns do not align with the majority, i.e., liking what most people dislike and vice versa [151].

Fairness and Goodness The Fairness and Goodness algorithm (F&G) [152] provides a measure for capturing user rating behavior. While many measures exist for this task [131], we chose to use F&G as Serra et al. [132] recently show it to be the best measure to identify trustworthy users in opinion-based systems. F&G computes a fairness score for each user and a goodness score for each item. Specifically, the *fairness* $f(u)$ of a user u is a measure of how fair or trustworthy the user is in rating items. Intuitively, a ‘fair’ or ‘trustworthy’ rater should give an item the rating that it

deserves, while an ‘unfair’ one would deviate from that value. In the case of benign users, the latter could be the case of an uninformative or non-mainstream user. The *goodness* $g(i)$ of an item i specifies how much users in the system like the item and what its true quality is. Fairness and goodness are mutually recursively computed as:

$$f(u) = 1 - \frac{1}{|out(u)|} \sum_{i \in out(u)} \frac{|W(u, i) - g(i)|}{R} \quad (4.1)$$

$$g(i) = \frac{1}{|in(i)|} \sum_{u \in in(i)} f(u) \times W(u, i) \quad (4.2)$$

where $W(u, i)$ is the rating given by the user u to the item i , $out(u)$ is the set of ratings given by user u , $in(i)$ is the set of ratings received by item i , and $R = 4$ in this case which corresponds to the maximum error in a five-star rating system. Thus, the goodness of an item is given by the average of its rating where each rating is weighted by the fairness of the rater, while the fairness of a user considers how much the ratings a user gives are far from the goodness of the items. The higher the fairness, the more trustworthy the user is. Fairness scores of the user lies in the $[0, 1]$ interval and goodness scores lie in the $[1, 5]$ interval.

4.4 Results and Discussion

In this section, we present our experimental evaluation of five recommendation algorithms on four datasets of different domains. We discuss the effect of shilling attacks on recommendations offered to users in *real-world* scenarios, as opposed to *simulated* attacks. Specifically, we aim to answer the following research questions:

RQ1 Do spammer’s ratings impact recommendations?

Dataset	Algorithm	RMSE		NDCG@5		HR@5		PS
		W Spammers	W/o Spammers	W Spammers	W/o Spammers	W Spammer	W/o Spammer	Attacked Items
YH	Item-Item	1.33	1.32	0.104	0.105	0.031	0.032	0.087
	PMF	1.125	1.124	0.57	0.57	0.0217	0.0216	0.119
	BPR			0.023	0.030	0.022	0.022	0.159
	ALS	1.028	1.020	0.041	0.043	0.0218	0.0217	0.143
	FunkSVD	1.023	1.019	0.034	0.032	0.022	0.022	0.129
YR	Item-Item	1.181	1.179	0.0073	0.0075	0.013	0.013	0.119
	PMF	1.040	1.037	0.56	0.56	0.014	0.014	0.122
	BPR			0.088	0.087	0.013	0.013	0.160
	ALS	0.971	0.970	0.049	0.051	0.013	0.013	0.148
	FunkSVD	0.993	0.981	0.012	0.013	0.0136	0.0137	0.138
AB	Item-Item	0.95	0.95	0.295	0.299	0.000140	0.000149	0.130
	PMF	0.912	0.905	0.552	0.553	0.000051	0.000051	0.121
	BPR			0.801	0.828	0.00023	0.00024	0.124
	ALS	0.802	0.802	0.265	0.264	0.0003	0.0003	0.116
	FunkSVD	0.637	0.644	0.028	0.032	0.0064	0.0061	0.11
AH	Item-Item	1.151	1.154	0.290	0.294	0.00013	0.00012	0.105
	PMF	1.053	1.051	0.518	0.519	0.000036	0.000036	0.101
	BPR			0.794	0.827	0.00023	0.00024	0.104
	ALS	0.952	0.952	0.198	0.204	0.00033	0.00032	0.10
	FunkSVD	0.994	0.933	0.070	0.067	0.00298	0.00296	0.10

Table 4.2: Performance analysis using different metrics on datasets with and without spam. Statistically significant differences are shaded in gray, $pvalue \leq 0.001$.

RQ2 Who is really affected by spammers?

By investigating recommender algorithm performance in the presence of spammers as well as when spammers are removed, the first question enables us to gauge the shilling attacks effect on the robustness of the considered recommendation algorithms. For the second question, we used the fairness metric to determine mainstream and non-mainstream users and quantify the effect of spammers on recommendations received by non-mainstream users.

4.4.1 Do spammers ratings impact recommendations?

To answer RQ1, we consider the performance of the recommender algorithms yielded on four different datasets, as reported in Table 4.2. It comes across from the reported scores that removing spammers indeed leads to lower RMSE scores, i.e., better predictions. Previous works have shown PS values ranging from 0.5 to 1.5 when shilling

attacks are simulated [153]. However, we observe very low values in real-world scenarios: in our case, considered, PS ranges from 0.087 to 0.160, which we argue might not be enough to promote or demote products attacked by the spammers. When looking at algorithm performance from a top-N recommendation standpoint, from reported NDCG@5 we see that, often, NDCG@5 scores tend to increase when removing spammers. This means that users' preferred items are more likely to appear within the top-5 recommendations when spammers are excluded. Unfortunately, improvement is not always meaningful, i.e., from Table 4.2 we see that improvements are not always significant, specially on YH. We anticipated lower HR@5 scores when excluding spammers—we assumed fewer attacked items would be promoted among the top-5 recommendations. Instead, we see similar trends among HR@5 results as those observed for NDCG@5. In other words, for YH and YR performance is comparable regardless of the presence of spammers (i.e., differences in performance are not significant); for AB and AH we see significant differences in performance.

Overall, we can say that, in theory, the performance of collaborative filtering-based recommender algorithms is affected by spammers' ratings/reviews. This is particularly noticeable for predictive recommenders (i.e., all algorithms yielded significant differences across the datasets). In practice, however, performance improvements are in their majority barely perceptible. This leads us to question whether algorithm robustness is reflected by average metrics like RMSE or NDCG. In the end, looking at recommender performance as a whole may not clearly quantify how much spammers are able to deceit recommenders and more importantly, if there are specific user groups that are affected more than others. With this in mind, we conduct a more thorough analysis with the aim of understanding if aforementioned differences in per-

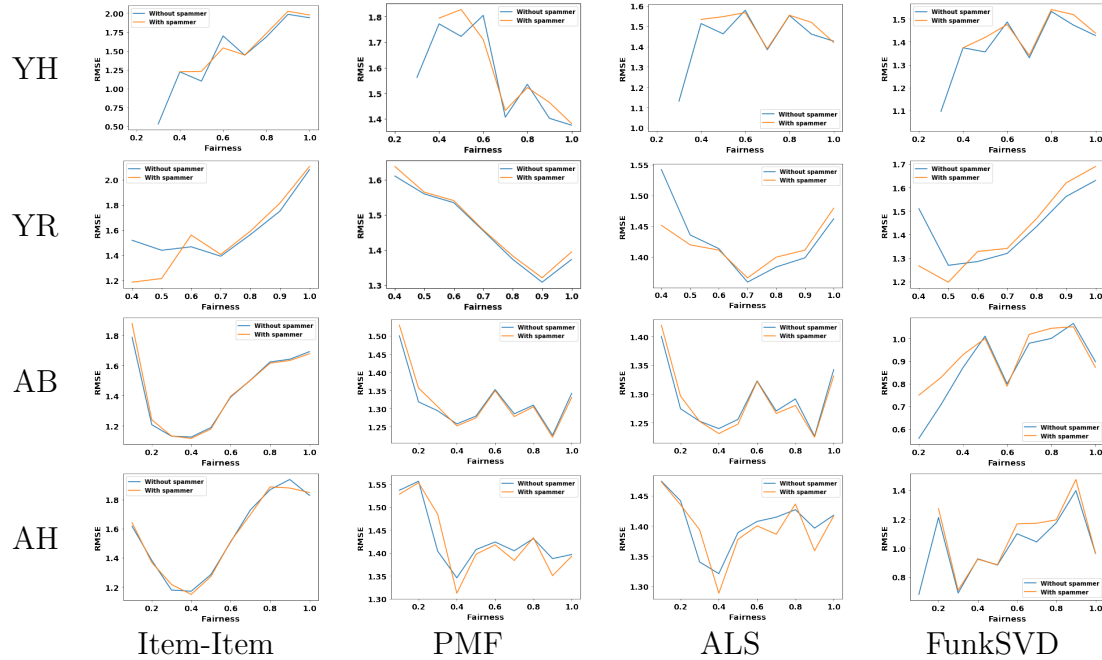


Figure 4.3: RMSE differences across fairness range.

formance are more pronounced among certain types of users (i.e., non-mainstream ones).

4.4.2 Who is really affected by spammers?

To better understand which users are really affected by spammers, we analyzed users based on their fairness: the ability of a user to rate a product according to what it deserves. It is worth noting, however, that the rating a product deserves often aligns with what the majority of benign users (mainstream users) think about that product, as mainstream users often outnumber non-mainstream and spam users. We investigate trends according to RMSE, NDCG@5, and hit ratio. As noted in the prior subsection, prediction shift values were small, so we excluded this metric from our analysis).

Figure 4.3 illustrates how the RSME varies according to the fairness of benign

Dataset	Algorithm	[0-0.1]	(0.1-0.2]	(0.2-0.3]	(0.3-0.4]	(0.4-0.5]	(0.5-0.6]	(0.6-0.7]	(0.7-0.8]	(0.8-0.9]	(0.9-1]
YH	#Benign Users	1	6	10	347	443	368	1281	550	882	389
	Item-Item									$W > W/o^{**}$	
	PMF							$W > W/o^{**}$		$W > W/o^{**}$	
	ALS					$W > W/o^{**}$		$W > W/o^{**}$		$W > W/o^{**}$	$W > W/o^{**}$
	FunkSVD						$W > W/o^*$		$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$
YR	#Benign Users	0	0	2	269	6502	4898	4580	7450	1889	2071
	Item-Item									$W > W/o^{**}$	
	PMF					$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$
	ALS					$W/o > W^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$
	FunkSVD					$W/o > W^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W > W/o^{**}$
AB	#Benign Users	1716	5352	13914	27947	25755	18632	16354	15353	11860	24861
	Item-Item		$W > W/o^{**}$	$W > W/o^{**}$		$W/o > W^{**}$			$W/o > W^{**}$		
	PMF		$W > W/o^*$		$W > W/o^*$					$W/o > W^{**}$	$W/o > W^{**}$
	ALS	$W > W/o^*$	$W > W/o^{**}$	$W > W/o^{**}$	$W/o > W^{**}$	$W/o > W^{**}$			$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$
	FunkSVD									$W > W/o^{**}$	
AH	#Benign Users	2574	6718	23273	46289	56630	41021	32512	30435	22363	36978
	Item-Item	$W > W/o^{**}$	$W/o > W^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W/o > W^{**}$	$W/o > W^{**}$				
	PMF		$W/o > W^{**}$	$W > W/o^{**}$	$W > W/o^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$
	ALS	$W/o > W^*$	$W/o > W^*$	$W > W/o^{**}$	$W > W/o^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$	$W/o > W^{**}$
	FunkSVD						$W > W/o^*$				

Table 4.3: Statistically significant RMSE differences between recommendations without spammers (W/o) and with spammers (W) across different user fairness ranges (* means $pvalue < 0.03$ and ** means $pvalue \leq 0.01$). Cases where removing spammers reduces the RMSE are shaded.

users; for ease of readability, we highlight statistically significance differences in performance when spammers are excluded in Table 4.3. We start by observing that, regardless of the algorithm for both Yelp! datasets and with just one exception (YR, ALS and FunkSVD, (0.4 – 0.5]), removing spammers reduces the RSME for all users. For the Amazon datasets, instead, when the user fairness is greater than 0.4, removing spammers increases the RMSE for all users, for each algorithm. We posit these results could be due to the the rating distributions of spammers vs. benign users, across these two platforms. As previously shown in Figure 4.2, spammer and benign users are more similar in Amazon than Yelp!, with the majority of ratings being 4 and 5. Therefore, removing spam could cause the recommender to lose information from mainstream users. On the other end, when fairness is less than or equal to 0.4 among Amazon users, in most cases where the difference is statistically significant, i.e., 14 out of 19 cases, removing spammers enables algorithms to avoid noise signals and thus perform better for these users (lower RMSE). Note that there are more cases in AH than AB (4 out of 11 vs. 1 out of 8) where removing the spammers is not

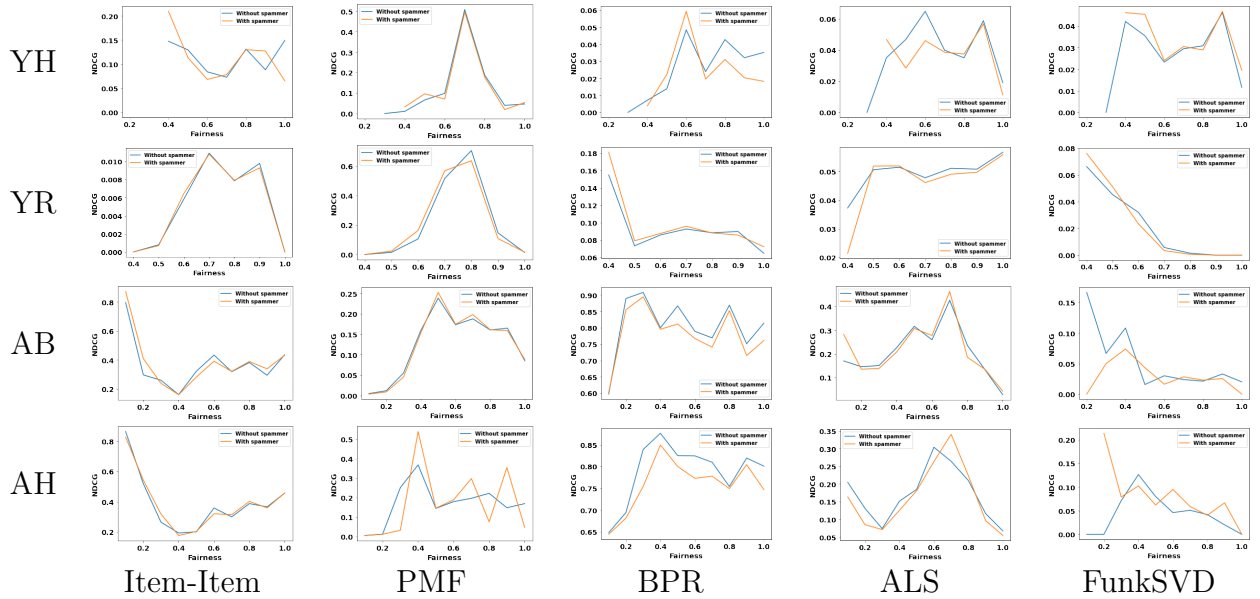


Figure 4.4: NDCG differences across fairness range.

beneficial for non-mainstream users. This could be due to the fact that, as shown in Figure 4.1, AH data is more sparse than other datasets, making the process of generating recommendations more difficult for most of the users in such a setting, independently of the presence of spam.

When we look at trends for NDCG@5, Figure 4.4 and Table 4.4 show that the quality of the generated recommendations seldom improves on the Yelp! datasets for users having fairness greater than 0.5, whereas for the Amazon datasets the value of NDCG@5 is higher when spammers are removed in the majority of the cases (28 out of 47) and independently of the user type.

Overall, our analysis reveals that removing spammers helps in reducing the number of attacked items that hit the top-5 recommendations for all the users in all the datasets (see HR@5 analysis in Table 4.5).¹ Moreover, removing spammers in Yelp!

¹For brevity, we exclude a figure akin to those complementing Tables 4.3 and 4.4.

Dataset	Algorithm	(0-0.1]	(0.1-0.2]	(0.2-0.3]	(0.3-0.4]	(0.4-0.5]	(0.5-0.6]	(0.6-0.7]	(0.7-0.8]	(0.8-0.9]	(0.9-1]
YH	Item-Item										
	PMF										
	BPR					W > W/o**					W/o > W**
	ALS										
	FunkSVD					W > W/o**					
YR	Item-Item										
	PMF					W > W/o**	W > W/o**	W > W/o**	W/o > W**	W/o > W**	
	BPR							W > W/o**			
	ALS							W > W/o**			
	FunkSVD							W/o > W**	W/o > W**		
AB	Item-Item		W > W/o**	W > W/o**	W/o > W**			W/o > W**	W/o > W**	W > W/o**	W > W/o**
	PMF		W/o > W**	W/o > W**		W > W/o**	W > W/o**	W > W/o**	W > W/o**	W/o > W**	
	BPR				W/o > W**	W/o > W**	W/o > W**	W/o > W**		W/o > W**	
	ALS										
	FunkSVD					W/o > W**	W > W/o**	W > W/o**			
AH	Item-Item			W > W/o**	W > W/o**	W/o > W**	W/o > W**	W/o > W**	W > W/o**	W > W/o**	W > W/o**
	PMF			W/o > W**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W/o > W**	W > W/o**	W > W/o**
	BPR		W/o > W**	W/o > W**	W/o > W**	W/o > W**	W/o > W**	W/o > W**	W/o > W**	W/o > W**	W/o > W**
	ALS		W/o > W**		W/o > W**						
	FunkSVD								W > W/o**		

Table 4.4: Statistically significant NDCG@5 differences between recommendations without spammers (W/o) and with spammers (W) across different user fairness ranges (* means $pvalue < 0.03$ and ** means $pvalue \leq 0.01$). Cases where removing spammers increases the NDCG@5 are shaded.

Dataset	Algorithm	(0-0.1]	(0.1-0.2]	(0.2-0.3]	(0.3-0.4]	(0.4-0.5]	(0.5-0.6]	(0.6-0.7]	(0.7-0.8]	(0.8-0.9]	(0.9-1]
YH	Item-Item										
	PMF					W/o > W**					W > W/o**
	BPR					W/o > W**					W > W/o**
	ALS					W/o > W**					W > W/o**
	FunkSVD					W/o > W**					W > W/o**
YR	Item-Item						W > W/o**	W > W/o**			
	PMF						W > W/o**	W > W/o**			W > W/o**
	BPR						W > W/o**	W > W/o**			
	ALS						W > W/o**	W > W/o**			W > W/o**
	FunkSVD						W > W/o**	W > W/o**			
AB	Item-Item	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**
	PMF	W > W/o**		W > W/o**		W/o > W**	W/o > W**	W/o > W**			
	BPR	W > W/o**	W > W/o**	W > W/o**		W > W/o**	W > W/o**	W > W/o**			
	ALS		W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**		W > W/o**	W > W/o**
	FunkSVD	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**		W > W/o**	W > W/o**	W > W/o**
AH	Item-Item	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**		W > W/o**	W > W/o**	W > W/o**
	PMF	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**			W > W/o**
	BPR		W > W/o**	W > W/o**		W/o > W**		W > W/o**			
	ALS		W > W/o**	W > W/o**	W > W/o**	W > W/o**	W > W/o**	W/o > W**		W > W/o**	W > W/o**
	FunkSVD	W > W/o**	W > W/o**		W > W/o**	W > W/o**	W > W/o**	W > W/o**			

Table 4.5: Statistically significant HR@5 differences between recommendations without spammers (W/o) and with spammers (W) across different user fairness ranges (* means $pvalue < 0.03$ and ** means $pvalue \leq 0.01$). Cases where removing spammers reduces the HR@5 are shaded.

is beneficial for all the users when considering predictive performance of algorithms (based on RSME); for Amazon, top-N algorithms are better (according to NDCG@5) among mainstream users, who see more tailored recommended items in their top-5 item list when spammers are removed from the system. Also, we see performance improvement in terms of both RMSE and NDCG@5 scores for Amazon non-mainstream

users, i.e., the ones with low fairness, hence the ones whose ratings are very far from the ones of the majority of the users. Non-mainstream users affected by spammers represent 29% (resp. 25%) of benign users in AB (resp. AH). In a real-world scenario, these percentages would translate into hundreds of thousands of users who would not be equally satisfied by recommenders that are not robust to shilling attacks.

4.5 Conclusions

In this chapter, we have taken a deeper look into how shilling attacks affect on recommender systems in a real world scenario. For this, we conducted an in-depth exploration on the performance of five well-known collaborative filtering algorithms on four different datasets.

We saw similar trends among performance of predictive and top-N recommenders: users are exposed to better recommendations when spammers are excluded (RQ1). This highlights the importance of further research of spammer detection and robust recommender systems. At the same time, we question if the small differences in performance (albeit statistically significant) would be evident to recommender systems' users and whether metrics considered for assessment which aggregate performance for all users could obfuscate users who are more deeply affected by spammers. This lead us to explore differences in performance between mainstream vs. non-mainstream users (RQ2). We saw that Amazon non-mainstream users are the ones most affected by spam ratings according to both RMSE and NDCG@5.

Based on the findings emerging from the analysis presented in this chapter, it follows that future work will be devoted to looking at other types of recommender algorithms, beyond those collaborative filtering-based, to see if the trends we have observed in our analysis remain. Moreover, we plan to also test the effectiveness

of adversarial training for recommender systems [125] under the real-world attacks considered in this chapter.

Part IV

Detecting trustworthy entities in the information ecosystem

CHAPTER 5:

DEEPTRUST: AN AUTOMATIC FRAMEWORK TO DETECT TRUSTWORTHY USERS IN OPINION-BASED SYSTEMS

5.1 Introduction

Opinion-based systems rely on users' collective opinion to rank or rate products, services, or even other users' qualifications or qualities (e.g., editors, programmers, micro-task workers). Such a crowdsourced approach allows for transparency and enables informed choices for other users interested in learning about certain reviewed items (or services). Underpinning such a system is an inherent expectation of trust on the participants' willingness and commitment to compile reliable and unbiased reviews accounting for their own experience with the item being reviewed. Although this expectation is generally met, research has consistently shown how these platforms are polluted by unreliable reviews that are either fraudulent, uninformative or inaccurate [49, 142, 154].

Malicious actors use underground Internet sites to recruit fake reviewers, who are given strict guidelines on the type of review to write, so as to generate relatively high-quality reviews - that are difficult to ascertain from authentic reviews [49]. This

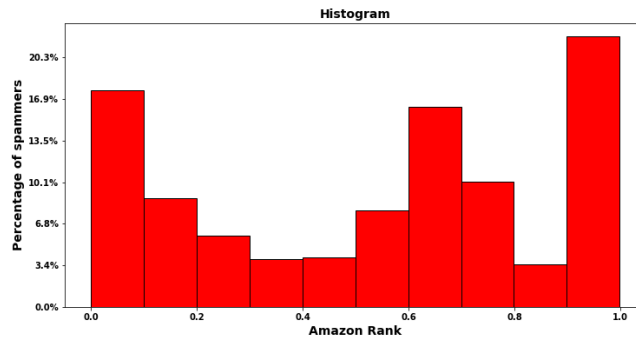


Figure 5.1: Distribution of the Amazon rank among fraudulent reviewers in the dataset from Section 5.5.

makes the detection of rogue reviews a non-trivial task. Existing user-led endorsement features offered by these platforms (e.g., “Is this review helpful?” in Amazon) are often unable to single out bogus reviews. Figure 5.1 shows the distribution of the Amazon rank for fraudulent (or opinion-spam) users on a selected dataset of fraudulent reviewers (see Section 5.5). Amazon ranks each reviewer according to peer-rated helpfulness. The lower the rank, the better the reviewer is. A reviewer’s rank is determined by the overall helpfulness of all their reviews, factoring in the number of reviews they have written, and, more recently, a review is written, the greater its impact on rank ¹. As Figure 5.1 shows, while the majority of the fraudulent users have a high rank, some of them are sophisticated enough to fuel the ranking system and camouflage as top-reviewers.

Opinion-based platforms have strong interests in identifying the best (and worst) contributors of their sites, to block or filter fraudsters, and to provide incentives to reviewers who contribute with honest and accurate information. In an effort to improve the detection of trustworthy individuals within opinion-based systems, in

¹<https://www.amazon.com/gp/customer-reviews/guidelines/top-reviewers.html>

this chapter, we focus on classifying reviewers based on their behavioral patterns and feedback they received from other peer reviewers. Particularly, we model the problem of detecting *trustworthy* reviewers as a multi-class classification problem, wherein the possible classes of users include fraudulent, unreliable (or uninformative), and trustworthy. Here, by trustworthy, we refer to individuals who positively contribute to the opinion-based system by means of productive content and, therefore, whose reviews can be trusted as informative and useful. Fraudulent reviewers are instead malicious in nature, in that truly uploaded to affect the ranking of a product or a seller’s reputation. For instance, let us consider user u_2 in Figure 5.2. This user is fraudulent as she/he is trying to demote p_2 , which is a generally liked product and promote p_3 , which other users consider a bad product. Finally, unreliable reviewers are users whose reviews are “noisy” in that they are not informative or of generally poor quality (e.g., inaccurate or generic). Further, as some reviewers may not have sufficient historical records to ascertain their nature reliably, we also consider a class of “unknown” users, whose true behavioral patterns is not well-supported by data. We classify such users based on their limited history or information.

We note that expanding from the classic binary classification of trustworthy/untrustworthy (or malicious) reviewers to a multi-class setting gives rise to an interesting and challenging problem. Untrustworthy reviewers may be rooted by a variety of motives, and be either perceived as uninformative or unreliable, or be actually malicious. Hence, natural language processing may not be sufficient nor accurate.

Our proposed approach accounts for users’ behavior over time, as they review multiple products by means of temporal embeddings. We model the interactions of reviewers and the products they review using a temporal review sequence and

consider the context of each interaction by including other reviewers’ interactions with the same items.

Precisely, we propose DeepTrust, an unsupervised temporal user embedding model (i.e., it does not require any label about the category of the user to be learned) that is able to extract latent features for each user automatically. Given a certain user u , these features take into account the entire temporal evolution of *all* the posted reviews from all users who reviewed the same products of the user u . In summary, the main strengths of this approach include:

- The entire historical sequence of the reviews can be reconstructed given the embedding features. Thus, we do not suffer from information loss as other existing methods that rely on aggregation of user information;
- Since for each user we also consider the reviews of their peers, the obtained sequences of reviews are usually large enough to allow the neural network embedding to be trained and produce significant embedding features. This allows us to classify users who reviewed a few products and whose history is, therefore, hard to leverage.

We report the results of our approach on a large dataset of Amazon reviews with fraudulent reviewer ground truth [142]. Our results show a drastic improvement in the detection of fraudulent reviewers as compared to related approaches. In addition, DeepTrust can detect trustworthy, uninformative, and fraudulent users with an F1-measure of 0.93. Also, we drastically improve on detecting fraudulent reviewers (AUROC of 0.97 and average precision of 0.99 when combining DeepTrust with the F&G algorithm) as compared to REV2 state-of-the-art methods (AUROC of 0.79 and

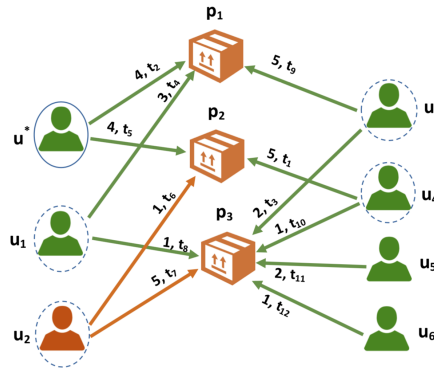


Figure 5.2: Sample user-item bipartite rating network ($t_i \leq t_j$ if $i \leq j$).

average precision of 0.48). Moreover, we show that DeepTrust performances do not decrease in the case of cold start users, and DeepTrust overperforms all the baselines approaches.

5.2 Related Work

Knowing the trustworthiness or reputation of a node u in opinion-based systems allows other peers to assign the right value to u 's judgments. Typically, the trustworthiness of a node is computed as a global trust value taking into account the interactions of a node with other nodes or items in the system [155, 154]. Specifically to opinion-based systems where users provide their opinions of items or products and not of other users as, for instance, in Amazon, several works have been proposed that have the common denominator of computing a trustworthiness score for each user and a goodness score for each item reflecting the rating the item actually deserves. These algorithms assume to work with a bipartite user-item rating network (cf. Figure 5.2 for an example).

Mishra and Bhattacharya [156] proposed the Bias and Deserve (BAD) algorithm

for computing the trustworthiness of a node as a *bias* quantifying the tendency of the node in overestimating or underestimating the rating an item deserves (the higher the bias, the less the trustworthiness of the node). The algorithm also computes a *deserve* score for each item that takes into account the bias of the users that are ranking that item. The bias and deserve scores are computed by a pair of mutually recursive equations.

Similarly, Kumar et al. [157] defined the Fairness and Goodness (F&G) algorithm, which computes a fairness score for each user and a goodness score for each item. Specifically, the *fairness* of a user is a measure of how fair or trustworthy the user is in rating items. Intuitively, a ‘fair’ or ‘trustworthy’ rater should give an item the rating that it deserves, while an ‘unfair’ one would deviate from that value. The latter could be the case of a fraudulent user who is trying to promote (resp. demote) a bad (resp. good) item or a user that is in good faith but unreliable or uninformative. The *goodness* of an item specifies how much users in the system like the item and what its true quality is. Fairness and goodness are mutually recursively computed. Specifically, the goodness of an item is given by the average of its rating where each rating is weighted by the fairness of the rater, while the fairness of a user considers how much the ratings a user gives are far from the goodness of the items. The higher the fairness, the more trustworthy the user is.

Recently, Kumar et al. [158] proposed the REV2 algorithm, which is an extension of the F&G algorithm where they compute a fairness score for each user, a goodness score for each item and a reliability score for each rating as they argue that fraudulent users can also give reliable rating to increase their reputation and fair users can sometimes give unreliable rating, as in case of the class of unreliable or uninformative

users we want to detect. Again, fairness is a measure of how trustworthy a user is, and a fair user is one that assigns reliable ratings that are close to the goodness of the items. REV2 computes fairness, goodness, and reliability by using a set of mutually recursive equations where they also combine user behavioral properties computed via the BirdNest algorithm [159].

Trustiness [160] is another algorithm similar to the above-mentioned ones that computes a trustworthiness score for each user, an honesty score for each item, and a reliability score for each rating.

Regarding the detection of fraudulent users (or opinion spammers) in opinion-based systems specifically, existing work can be categorized into network-based methods, behavioral-based methods, and hybrid methods combining both network and behavioral properties. BAD, F&G, and Trustiness can be used to detect fraudulent users as well, and they can be categorized as network-based algorithms. FraudEagle [161] is another network-based algorithm that models the user-item bipartite rating network as a Markov Random Field and computes an anomaly score for each user that is used to identify the opinion spammers. They assume that honest (resp. fraudulent) reviewers are more likely to give positive ratings to good (resp. bad) products and honest (resp. fraudulent) reviewers are more likely to give negative ratings to bad (resp. good) products.

Behavioral-based methods leverage the fact that fraudulent reviewers write many, shorter, positive (4 or 5 stars) and self-similar reviews in short bursts of time [49, 162, 163, 164]. SpamBehavior [165] proposes ranking and supervised methods exploiting the fact that opinion spammers target a specific set of products and their ratings deviate from the ones of benign users. BirdNest [159] detects opinion spammers

according to the fact that (i) fake reviews occur in short bursts of time and (ii) fraudulent user accounts have skewed rating distributions.

Among hybrid methods, SpEagle [140] extends FraudEagle by combining both the user-review-product graph and metadata such as text, timestamps, and ratings to detect fraudulent users, reviews, and targeted products. REV2 combines both network and behavioral properties (by incorporating the user BirdNest anomalous score) and is the state-of-the-art algorithm in detecting fraudulent users in opinion-based systems. *In this chapter, we extensively compare with REV2 (and other algorithms presented in this section) and show that our proposed DeepTrust user embedding technique significantly outperforms REV2 and other algorithms under different settings. Also, DeepTrust can identify uninformative reviewers, which are not considered in prior work, to avoid they are mistakenly classified as fraudulent users, and addresses the cold start user problem.*

5.3 DeepTrust User Embedding

In this section, we describe DeepTrust, a deep-learning-based approach to extract user features from their temporal review sequence in an unsupervised way. An “embedding” is a technique to transform an input sequence into a k -dimensional vector. Once the embedding is obtained for each user, its vectorial representation can be used as features in input to machine learning algorithms. In this chapter, we use the computed user embedding to determine if a user belongs to one of these categories: *trustworthy, unreliable or uninformative, or fraudulent.*

Let $U = \{u_1, \dots, u_m\}$ be the set of users active in the opinion-based system, and $P = \{p_1, \dots, p_l\}$ be the set of products being reviewed by users in U . We denote by R the set of all reviews generated by users in U for products in P . Each review

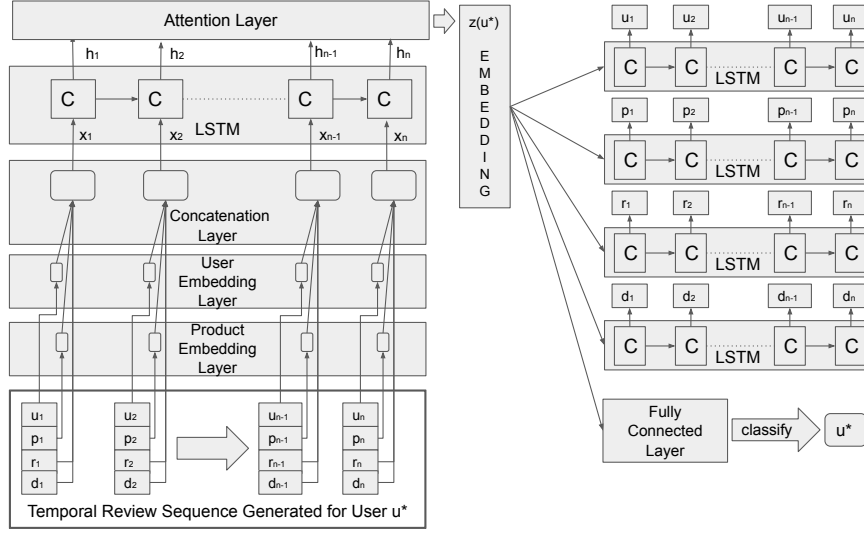


Figure 5.3: DeepTrust architecture.

$re \in R$ is represented by a 4-tuple $re = (u, p, r, t)$ where $u \in U$ is the reviewer, $p \in P$ is the product being reviewed, $r \in \{1, 2, 3, 4, 5\}$ is the five-star scale rating assigned by u to p , and t is the review timestamp. Given a user $u \in U$, we define the set of products reviewed by u as $pr(u) = \{p | (u, p, -, -) \in R\}$. Given a product $p \in P$, the set of users who reviewed p is defined as $us(p) = \{u | (u, p, -, -) \in R\}$.

In order to define an embedding describing the behavior of a user $u^* \in U$ in the opinion-based system, we consider, in addition to their reviews, also the reviews from other users in the system that can have potentially shaped u^* 's opinions or that can help in identifying an anomalous or fraudulent behavior. We call these reviews the *context* of the user u^* . For instance, users' opinion on a given product may change as they read other users' reviews, or their ratings may change slightly based on other reviewers' ratings of the same or similar products. Also, some reviews written by u^* can affect the opinion of other users. Moreover, we also consider reviews made

by other users *after* u^* 's reviews as they can help uncover fraudulent behavior. For instance, opinion spammers may be engaged to promote a new (but not good) product p , which initially is described only by spam reviews with star ratings 4 or 5. Later, benign users start reviewing the product p and giving low ratings. These future reviews are helpful to recognize the opinion spammers.

Accordingly, our embedding will consider in input a sequence of temporally ordered reviews for each user. Given a user u^* the sequence consists of the set of all the reviews given by u^* plus the reviews given by other users to the set of products reviewed by u^* (as they can be potentially related to the behavior of u^*). Then, we define the concept of a temporal review sequence as follows.

Definition 1 (Temporal Review Sequence). *Given a user $u^* \in U$, let $re(u^*) = \{(u, p, r, t) \mid (u, p, r, t) \in R, p \in pr(u^*)\}$ be the subset of reviews that describe the rating behavior of u^* in comparison to the ratings that other users give to the products rated by u^* .² The temporal review sequence*

$$tres(u^*) = \langle (u_0, p_0, r_0, t_0, d_0), (u_1, p_1, r_1, t_1, d_1), \dots, (u_n, p_n, r_n, t_n, d_n) \rangle$$

for the user u^* is the set of reviews in $re(u^*)$ ordered by the timestamp such that

- $(u_i, p_i, r_i, t_i) \in re(u^*)$ for each $i \in \{1, \dots, n\}$
- $t_{i-1} \leq t_i$ for each $i \in \{1, \dots, n\}$

²Because of data limitation, as discussed in Section 5.5, behavioral analysis is limited to users' rating activities. However, our proposed technique can be easily extended in case other behavioral data is available.

$$\bullet d_i = \begin{cases} 0, & i = 0 \\ t_i - t_{i-1}, & i > 0 \end{cases} \quad \square$$

It is worth noting that two reviews $(u_{i-1}, p_{i-1}, r_{i-1}, t_{i-1})$ and (u_i, p_i, r_i, t_i) that occur at the same time ($t_{i-1} = t_i$) will be recognized by $d_i = 0$. Moreover, we define the *context* of the user u^* as the subsequence of reviews in $re(u^*)$ not written by u^* , i.e., $ctx(u^*) = \{(u, p, r, t) \mid (u, p, r, t) \in R, p \in pr(u^*), u \neq u^*\}$.

Example 1. *Let us consider the sample user-item bipartite rating network shown in Figure 5.2. We have three products $P = \{p_1, p_2, p_3\}$ reviewed by seven users $U = \{u^*, u_1, u_2, u_3, u_4, u_5, u_6\}$. Each edge is labeled with rating and review timestamp. For user u^* , the temporal review sequence that describes their behavior is*

$$tres(u^*) = \{(u_4, p_2, 5, t_1), (u^*, p_1, 4, t_2), (u_1, p_1, 3, t_4), (u^*, p_2, 4, t_5),$$

$$(u_2, p_2, 1, t_6), (u_3, p_1, 5, t_9)\}$$

since p_1 and p_2 are the products reviewed by user u^ and u_1, u_2, u_3, u_4 are other users reviewing same products. The context for user u^* is given by*

$$ctx(u^*) = \{(u_4, p_2, 5, t_1), (u_1, p_1, 3, t_4), (u_2, p_2, 1, t_6), (u_3, p_1, 5, t_9)\}$$

□

5.4 DeepTrust Architecture

We compute a set of latent features describing the user behavior that can be used as input to machine learning algorithms to classify users as trustworthy, unreliable, or

fraudulent. We learn the features from the input user temporal review sequence in an unsupervised way as we aim to be generic and not constrained on the particular task we are going to use the features for.

Figure 5.3 illustrates DeepTrust, our proposed deep-learning architecture to compute the user embeddings. Given as input a temporal review sequence of variable length $tres(u^*)$, DeepTrust maps the sequence into a fixed-size vectorial representation $z(u^*) \in \mathbb{R}^k$. For each element $(u_i, p_i, r_i, t_i, d_i)$ in the sequence $tres(u^*)$, DeepTrust associates the IDs of the user u_i and product p_i with their latent representations $e(u_i)$ and $e(p_i)$ (embeddings). These operations are performed by the “Product Embedding Layer” and the “User Embedding Layer” of the neural network.

Note that the embeddings associated with each product and user will be determined during the training phase, i.e., they will be trained with the entire network. Once the IDs are converted into embeddings, the “Concatenation Layer” will concatenate, for each element $(u_i, p_i, r_i, t_i, d_i)$ in the sequence, the embedding of the user $e(u_i)$, the embedding of the product $e(p_i)$, the rating r_i , and the time elapsed since the previous review (delta) d_i . We denote by x_i the above concatenation. Moreover, to improve the training convergence time of our network, we scaled with a standard scaler all the rating r_i and all the deltas d_i . Once the concatenated representation x_i is obtained for each element at time i in the temporal review sequence, the sequence $\langle x_1, \dots, x_n \rangle$ is passed through a Long Short Term Memory (LSTM) recurrent neural network [166].

Figure 5.4, adapted from [2], describes the architecture of a single LSTM cell that outputs the next state h_t by taking in input the previous state h_{t-1} and the next symbol x_t . The operations done by the single LSTM cell C are described by the

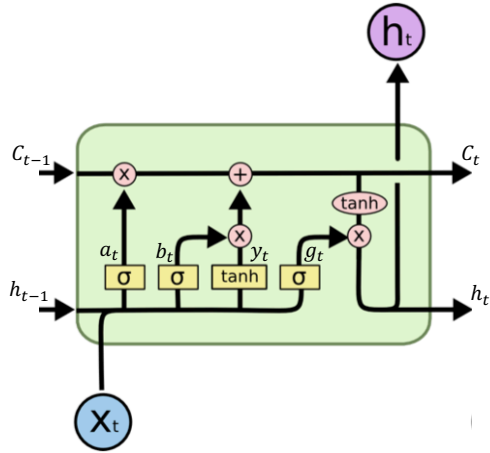


Figure 5.4: Description of an LSTM cell C . Figure adapted from [2].

following equations:

$$a_t = \rho(W_a \cdot [h_{t-1}, x_t])$$

$$b_t = \rho(W_b \cdot [h_{t-1}, x_t])$$

$$y_t = \tanh(W_y \cdot [h_{t-1}, x_t])$$

$$g_t = \rho(W_g \cdot [h_{t-1}, x_t])$$

$$c_t = c_{t-1} \cdot a_t + b_t \cdot y_t$$

$$h_t = \tanh(c_t) \cdot g_t$$

where the W_a, W_b, W_y and W_g are the weights representing the LSTM cell C and the entire LSTM neural network.

The LSTM outputs, for each element of the sequence, a vectorial representation h_i representing the sub-sequence till the element i . We then merge the vectorial representation sequence $\langle h_1, \dots, h_n \rangle$ via a soft attention layer that produces a unique fixed-size vectorial representation for the temporal review sequence. The attention

[167] is a mechanism to discover parts of the sequence that are more relevant for describing user behavior and weight them more when computing the user embedding. The attention layer takes as input the vectorial representation sequence $\langle h_1, \dots, h_n \rangle$ from the LSTM and returns the output $z(u^*) = \tanh(W_c[co; h_n])$ where

- $W_c \in \mathcal{R}^{2|h_n| \times |h_n|}$ is a set of weights to learn, and
- co is the attention context vector obtained by the weighed mean of all h_i vectors, i.e., $co = \sum_{i=1}^n c_i^s \cdot h_i$.

The weights $\{c_1^s, \dots, c_n^s\}$ to compute the attention context vector are obtained by using (1) a unique fully connected layer that, applied singularly to each h_i , produces a single value q_i , and (2) a softmax layer taking in input all $\{q_1, \dots, q_n\}$ and finally outputting the $\{c_1^s, \dots, c_n^s\}$.

The output $z(u^*)$ represents the embedding of the user u^* that summarizes their entire temporal review sequence.

Given the embedding $z(u^*)$, the next part of our neural network works on the reconstruction of the input temporal review sequence $tres(u^*)$. The embedding $z(u^*)$ is passed through four different LSTMs that reconstruct the original sequences of all the sub-part of each review: the sequence of user IDs, the sequence of product IDs, the sequence of product ratings, and the sequence of deltas.

We use a softmax layer to reconstruct product and user sequences, and add to the global loss function the cross-entropy loss for each element of the sequence. The reconstruction for the rating and the deltas is done by adding to global loss function the mean square loss for each element of the sequence. This ensures that the embedding $z(u^*)$ stores all the information contained in $tres(u^*)$. In addition, we

pass $z(u^*)$ through a fully connected layer that identifies the specific user u^* who is generating the temporal review sequence. By training this entire neural network (i.e., minimizing the global loss function), the output is an embedding $z(u)$ for each user $u \in U$.

5.5 Dataset

We carry out our experimental evaluation on an Amazon dataset from Fayazi et al. [142]. The dataset includes a large candidate set of potential deceptive reviews, reviewers, and targeted products. Deceptive reviews are retrieved by identifying products that were targeted for manipulation in underground crowdsourcing platforms (e.g., RapidWorkers.com, ShortTask.com, and Microworkers.com). These platforms pay workers to post a review on a target site (e.g., Amazon, Yelp). The dataset also includes samples of reviews of other products performed by suspected deceptive reviewers, along with their profile information (i.e., reviewers whose reviews appear in the targeted items). A review is labeled as deceptive if (i) the review was associated with a product which was targeted by a crowdsourced malicious task; and if (ii) the review had a high rating and was posted within a few days after the task was posted. Otherwise, it is labeled as a legitimate review.

For each reviewer, attributes such as helpfulness ratio (or reviewer helpfulness), number of helpful/unhelpful votes, and Amazon rank are included. Amazon users can provide feedback on other users' reviews by voting if a given review is helpful or not. Thus, given a reviewer u , their *helpful ratio* is given by the number of helpful votes divided by the total number of helpful/unhelpful votes given to all the reviews of this user. The Amazon rank, as explained earlier in the Introduction, is assigned by taking into account the helpfulness of the reviewer and the recency of the votes.

Thus, when reviewers receive enough helpful/unhelpful votes (we consider a threshold of 20 votes in this chapter), their helpfulness score can be seen as a collective measure of user trustworthiness (when it is high) or user unreliability (when it is low).

We filtered out reviewers who did only one review and products that have only one review. The resulting dataset includes 94.8K reviews, 14.1K reviewers, and 22K products. We further split reviewers into four classes:

- **Fraudulent users:** users who are marked as opinion spammers (fraudulent) in the dataset.
- **Trustworthy users:** users who are not fraudulent and who have at least 20 helpful votes and a helpfulness ratio ≥ 0.75 .
- **Unreliable or uninformative users:** users who are not fraudulent and who have at least 20 helpful votes and a helpful ratio ≤ 0.25 .
- **Unknown users:** users who are not fraudulent and have less than 20 helpful votes or the helpfulness ratio is between 0.25 and 0.75. We classify them as unknown as there is not enough evidence in the helpfulness data to reliably assign them a classification label.

Table 5.1 shows the breakdown of labels in each of the above classes. As we discuss in the next section, the obvious class imbalance shown here is taken into account by adding appropriate weights to our learning models.

5.6 Experiments

In this section, we report an extensive experimental evaluation of our proposed DeepTrust user embedding model and compare its performance against several state-of-

Table 5.1: Number of users for each class in the Amazon dataset.

Class	Num. of users
Fraudulent	846
Trustworthy	5,322
Unreliable	91
Unknown	7,872

the-art algorithms.

5.6.1 Experimental Settings

We consider three main settings in our experiments: (1) a multi-class problem where we classify users as trustworthy, fraudulent, or unreliable/uninformative; (2) a binary classification problem where we detect *fraudulent* users (vs. trustworthy and unreliable/uninformative), and (3) a binary classification problem where we classify trustworthy vs. untrustworthy (fraudulent and unreliable/uninformative) users. In all the experiments, unknown users are included in the user-item rating network used for computing the temporal review sequence we use to learn the DeepTrust features and for computing the baselines, but unknown users are not used as instances in the classification tasks. However, in Section 5.6.5, we tackle the problem of classifying the “unknown” users in the dataset into one of the remaining three possible categories and correlate the computed labels with the Amazon rank for additional insights.

We report results for classification with a Random Forest model. We also tested other classification algorithms, including Logistic Regression and Support Vector Machine, but Random Forest resulted in overall best performance.

We used class weighting to deal with class imbalance. Class weighting is a way to learn from an unbalanced dataset where the classification imposes, during training, a penalty proportionally inverse to the class distribution on the model for making

Table 5.2: Precision, recall, and F1-measure results of detecting trustworthy, fraudulent, and unreliable users with DeepTrust and comparison with related work.

Algorithm	Precision	Recall	F1-measure
DeepTrust	0.93	0.93	0.93
DeepTrust w/o context	0.81	0.54	0.58
BAD [156]	0.81	0.40	0.43
F&G [157]	0.90	0.89	0.89
REV2 [158]	0.82	0.64	0.69
Trustiness [160]	0.72	0.26	0.32

Table 5.3: Precision, recall, and F1-measure for DeepTrust combined with related work for detecting trustworthy, fraudulent, and unreliable users.

Algorithm	Precision	Recall	F1-measure
DeepTrust	0.93	0.93	0.93
DeepTrust + BAD	0.93	0.93	0.93
DeepTrust + F&G	0.95	0.95	0.94
DeepTrust + REV2	0.93	0.93	0.92
DeepTrust + Trustiness	0.93	0.93	0.93

classification mistakes. We performed 10-fold cross-validation for all reported experiments.

In regards to evaluation measures, we report weighted precision, recall, and F1-score in the case of multi-class classification. For binary classification, in addition to reporting the above measures to allow comparison, we also calculate the Area Under the Receiver Operating Characteristics (AUROC) and Average Precision (AvgP). The best results are highlighted in bold in the tables.

5.6.2 Detecting Trustworthy, Unreliable, and Fraudulent Users

We tested our DeepTrust user embedding on the problem of classifying users as trustworthy, unreliable, and fraudulent. Table 5.2 reports the classification results accord-

ing to precision, recall, and F1-measure for DeepTrust (with and without context) and several state-of-the-art approaches. Our baselines include methods to compute trustworthiness scores for users in opinion-based systems. Specifically, we compare with Bias and Deserve (BAD) [156], Fairness and Goodness (F&G) [157], REV2 [158], and Trustiness [160].

As we can see, our DeepTrust proposed embedding technique consistently outperforms all the other approaches. Among the competitors, F&G achieves the best performance. DeepTrust improves over F&G by 3% in precision and 4% in recall and F1-measure. Moreover, as we can see from the table, the DeepTrust performance significantly drops when removing the context from our user sequences (i.e., not considering the reviews from other users on the set of products reviewed by the given user). This further motivates our choice of considering the user context when computing our embedding.

Next, we combine DeepTrust with any of the existing methods to see if we can further improve our method’s performance. To combine DeepTrust with another method, we consider our embedding features and the predictive user features of the other method together in input to the Random Forest classifier. For instance, to combine DeepTrust with REV2, we added to our features the user fairness scores computed by REV2. Table 5.3 reports comparative results. We see that DeepTrust+F&G yields the best combination, which further improves DeepTrust achieving both precision and recall of 0.95 and F1-measure of 0.94.

Addressing Cold Start Users. We also tested DeepTrust on the problem of classifying users with short or no history (cold start users). In our dataset, we define these “cold-start users” as the users who completed less than four reviews. To per-

Table 5.4: Precision, recall, and F1-measure results of detecting trustworthy, fraudulent, and unreliable *cold start* users with DeepTrust and comparison with related work.

Algorithm	Precision	Recall	F1-measure
DeepTrust	0.92	0.92	0.91
DeepTrust w/o context	0.84	0.34	0.40
BAD [156]	0.02	0.12	0.04
F&G [157]	0.89	0.87	0.88
REV2 [158]	0.82	0.51	0.58
Trustiness [160]	0.001	0.03	0.001

Table 5.5: Precision, recall, and F1-measure for DeepTrust combined with related work for detecting trustworthy, fraudulent, and unreliable *cold start* users.

Algorithm	Precision	Recall	F1-measure
DeepTrust	0.92	0.92	0.91
DeepTrust + BAD	0.92	0.92	0.92
DeepTrust + F&G	0.94	0.94	0.93
DeepTrust + REV2	0.91	0.92	0.91
DeepTrust + Trustiness	0.92	0.92	0.91

form this experiment, we tested only on cold start users in each test of the 10-fold cross-validation. Results are reported in Table 5.4 for DeepTrust and baselines, and in Table 5.5 for the combination.

We see from the results that DeepTrust performance is pretty stable, seemingly due to the user context in the formulation that also allows addressing the cold start user problem. When we compare results from Tables 5.2 and 5.4, we note that that DeepTrust achieves precision, recall, and F1-measure always above 0.91 for *both* general users and cold start ones. Further, we note that not knowing the context results in lower recall performance for the cold start users than for all the users (0.34 vs. 0.54). Specifically, by looking at the individual class recall values, we

observe that the recall drastically drops from 0.48 to 0.27 for the class of helpful users. This is because cold start users have just a limited number of reviews, similarly to many fraudulent users. Consequently, benign cold start users can be misclassified as fraudulent. Context information, on the other hand, helps our model with additional information on the ratings of other users to overcome the problem of having a few reviews. This improves the recall.

Baseline methods perform worse than DeepTrust (as expected) and, similarly to what observed before, combining DeepTrust with F&G further improves overall performance by 2% in precision, recall, and F1-measure (see Table 5.5).

5.6.3 Detecting Fraudulent Users

For most online platforms, the most damaging category of users is that of fraudulent users who spoil the community posts with fake content. Accordingly, we test our embedding on its ability to detect *fraudulent* users specifically. As this is a binary classification problem, we report Average Precision and AUROC scores in addition to precision, recall, and F1-Measure. Moreover, we compare our proposed DeepTrust with five state-of-the-art algorithms specifically defined for fraudulent user detection in opinion-based systems, namely FraudEagle [161], Bias and Deserve (BAD) [156], SpamBehavior [165], ICWSM’13 [49], and REV2 [158]. We chose these algorithms as they are the top-five best-performing algorithms according to the experiments done for supervised classification in [158] on a similar Amazon dataset. Moreover, we also included the Fairness and Goodness (F&G) algorithm in the comparison as it was not included in the experimental evaluation performed in [158]. Results are shown in Table 5.6. As we can see, DeepTrust significantly outperforms all the competitors according to all the measures considered on the task of detecting fraudulent users,

Table 5.6: Classification result for fraudulent user detection: precision, recall, F1-measure, AUROC, and average precision (AvgP).

Algorithm	Precision	Recall	F1	AUROC	AvgP
DeepTrust	0.94	0.95	0.94	0.94	0.88
DeepTrust w/o context	0.84	0.59	0.64	0.72	0.31
REV2 [158]	0.84	0.67	0.71	0.79	0.48
BAD [156]	0.85	0.51	0.56	0.69	0.25
FraudEagle [161]	0.85	0.64	0.69	0.77	0.35
SpamBehavior [165]	0.88	0.87	0.88	0.84	0.57
ICWSM'13 [49]	0.89	0.88	0.88	0.86	0.59
F&G [157]	0.92	0.91	0.91	0.91	0.71
DeepTrust+F&G (Best combination)	0.96	0.96	0.96	0.97	0.99

especially in terms of average precision for fraudulent user detection (+17%). Also in this setting, dropping the context information from the computation of our user embedding decreases the performance. Among the competitors, F&G is the best method according to all performance measures. As we can see from the last row of Table 5.6, when we combine DeepTrust+F&G we further improve: F1-measure of 0.96, AUROC of 0.97, and average precision of 0.99 (an improvement of 5% in F1-measure, 6% in AUROC and 28% in average precision as compared to F&G performances). All the other combinations of DeepTrust with baselines achieve worse results than DeepTrust+F&G as reported in Table 5.8 in the Appendix.

5.6.4 Classifying Trustworthy vs. Untrustworthy Users

We now analyze the ability of DeepTrust in detecting trustworthy vs. untrustworthy users (fraudulent and unreliable/uninformative) via a binary classification problem.

We aim to identify trustworthy users in opinion-based systems so that other users in the platform can rely on their reviews when buying products. In comparing DeepTrust with other works, we consider all the algorithms for trustworthiness and fraudulent user detection used in Sections 5.6.2 and 5.6.3. Results are reported in Table 5.7. Also in this setting, DeepTrust with the user context outperforms all the competitors achieving an F1-measure of 0.93, an AUROC of 0.92, and an average precision of 0.97.

As we can see from the last row of Table 5.7, when we combine DeepTrust with F&G (which also in this case is the best performing baseline), we further improve the classification: F1-measure of 0.95, AUROC of 0.94, and average precision of 0.98. All the other combinations of DeepTrust with baselines achieve worse results than DeepTrust+F&G as reported in Table 5.9 in the Appendix.

5.6.5 Classifying Unknown Users

Finally, we carried an additional experiment attempting to classify users' whose status is "unknown" (see Section 5.5). We trained our model using trustworthy, unreliable, and fraudulent users and use as test-set the unknown users.

We used our best feature set, i.e., DeepTrust+F&G ³, for training a Random Forest classifier. In order to interpret the quality of our labels, since no other ground truth is available, we relied on Amazon Ranking, and specifically the rank assigned to users. Figure 5.5 shows the unknown users ordered by the Amazon rank (on the x-axes) along with our prediction: trustworthy users are shown in green, unreliable in yellow, and fraudulent in red. Since the lower the rank, the better the reviewer is, we expect to see in Figure 5.5 that a higher frequency of trustworthy users are on the

³We considered the user fairness feature from F&G.

left (low rank), unreliable users are mostly in the middle, and fraudulent users are on the right side of the figure (high rank).

As the figure shows, our prediction follows this pattern very closely. In fact, the top-ranked users are correctly predicted as trustworthy, and the majority of untrustworthy users (fraudulent and unreliable users) appear on the right side. Specifically, 74% of the predicted fraudulent users and 80% of the users predicted as unreliable have a rank higher than 4,000.

Observe that, within our predicted fraudulent and unreliable/ uninformative users, some of them (9 fraudulent and 15 unreliable users) rank relatively high, between 1,500 and 2,500. This is suggestive of a possible bias in either our results or in the ranking system itself. While it is not possible to point to a specific error on either side, we note that while Amazon ranks are accepted as strong indicators of the quality of reviewers (and this is consistent with our findings), Amazon ranking method is known to be vulnerable to biases, and fraudulent users may reach high ranks (see Figure 5.1). We speculate that some of the anomalies in our findings are examples of such users' ability to climb the ranking system.

5.7 Discussion

This work contributes to the state-of-the-art on trustworthiness detection in opinion-based systems. Our approach is able to detect both trustworthy and malicious users and leverages review traces of users across products. Despite its strengths, our approach is not privy of limitations. We summarize some open issues in what follows.

- *Trustworthiness*: In the context of opinion-based systems, a trustworthy reviewer is a user with a record of well-perceived reviews by readers or testers/users of the items reviewer commented on. Hence, trust in recommender systems (or

Table 5.7: Classification result for trustworthy vs.untrustworthy user detection: precision, recall, F1-measure, AUROC, and average precision (AvgP).

Algorithm	Precision	Recall	F1	AUROC	AvgP
DeepTrust	0.93	0.95	0.93	0.92	0.97
DeepTrust w/o context	0.82	0.58	0.63	0.71	0.89
REV2 [158]	0.82	0.68	0.70	0.77	0.92
BAD [156]	0.83	0.52	0.56	0.68	0.88
FraudEagle [161]	0.82	0.64	0.66	0.73	0.91
SpamBehavior [165]	0.87	0.85	0.86	0.83	0.94
ICWSM'13 [49]	0.87	0.86	0.86	0.84	0.94
F&G [157]	0.90	0.91	0.90	0.88	0.96
Trustiness [160]	0.74	0.61	0.65	0.60	0.85
DeepTrust+F&G (Best combination))	0.95	0.95	0.95	0.94	0.98

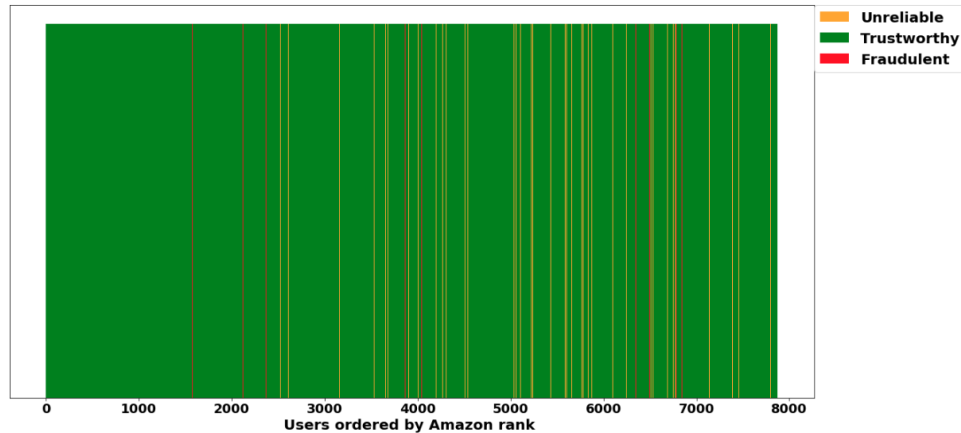


Figure 5.5: Classification of unknown users as trustworthy (green), unreliable (yellow), and fraudulent (red) and correlation with the Amazon rank.

opinion-based systems) is sometimes defined as “competence” or “confidence,” to distinguish it from conventional trust notions wherein trust is based on prin-

cipals' identities and credentials [168, 169, 170]. As the definition is often not robust, fraudulent users or users whose behavior changes over time may mistakenly be labeled as trustworthy.

- *Limitations of method:* We have assumed opinion-based systems where the rater and the item being rated are two different entities, e.g., in the case of Amazon, we have users rating products. However, there are other types of opinion-based systems where users rate other users. As an example, in Bitcoin trade networks, users rate the level of trust they have in other users [157]. The definition of temporal review sequence we have given in this work does not fit the case of user-user opinion-based systems as a user u should be considered fraudulent according to how badly she/he is judged by other benign users, rather than how u judges other users. Thus, we should consider u 's incoming edges from the user-user rating network to build the temporal review sequence of user u rather than the outgoing edges, as in the case of the user-item rating network. Further, the notion of context should be adapted to the case of user-user opinion-based systems. We plan to investigate this case as future work.
- *Evolution of users:* We have assumed that users do not change status, i.e., they are either malicious or not, with no possible state change. That is to say, the temporal sequence does not reveal an evolving pattern and users are labeled as trustworthy (or otherwise) regardless of the incidence of fake reviews (e.g., if a user posts one fake review she/he is labeled untrustworthy or fraudulent, even if other reviews were actually authentic). This may create some noise in the temporal sequences, in addition to being unrealistic. A soft label approach may be needed to better account for users' changing of behavior.

Table 5.8: Precision, recall, F1-measure, AUROC, and average precision (AvgP) for DeepTrust combined with related work for detecting fraudulent users.

Algorithm	Precision	Recall	F1	AUROC	AvgP
DeepTrust	0.94	0.95	0.94	0.94	0.88
DeepTrust + BAD	0.95	0.95	0.94	0.95	0.89
DeepTrust + F&G	0.96	0.96	0.96	0.97	0.99
DeepTrust + REV2	0.94	0.94	0.94	0.95	0.88
DeepTrust + FraudEagle	0.93	0.93	0.93	0.93	0.86
DeepTrust + SpamBehavior	0.93	0.94	0.93	0.92	0.85
DeepTrust + ICWSM'13	0.94	0.94	0.93	0.93	0.87
DeepTrust + Trustiness	0.95	0.95	0.94	0.96	0.89

Table 5.9: Precision, recall, F1-measure, AUROC and average precision (AvgP) of combining DeepTrust with related work for detecting trustworthy users.

Algorithm	Precision	Recall	F1	AUROC	AvgP
DeepTrust	0.93	0.95	0.93	0.92	0.97
DeepTrust + BAD	0.93	0.93	0.93	0.93	0.98
DeepTrust + F&G	0.95	0.95	0.95	0.94	0.98
DeepTrust + REV2	0.93	0.93	0.93	0.92	0.97
DeepTrust + FraudEagle	0.93	0.93	0.93	0.92	0.97
DeepTrust + SpamBehavior	0.93	0.94	0.93	0.92	0.97
DeepTrust + ICWSM'13	0.94	0.94	0.93	0.93	0.98
DeepTrust + Trustiness	0.93	0.93	0.93	0.93	0.98

5.8 Conclusions

In this chapter, we proposed a supervised approach to identify trustworthy reviewers in an opinion-based system. We presented the problem of detecting trustworthy reviewers as a multi-class classification problem, wherein users may be fraudulent, unreliable or uninformative, or trustworthy. We address the problem by means of a temporal user embedding based on a deep recurrent neural network. We auto-

matically learn relevant features from the input user temporal review sequence in an unsupervised way and use these features for classifying users into trustworthy, unreliable, or fraudulent. Our proposed approach outperforms existing methods under different settings and is able to effectively learn minority classes of users whose behavior is unknown or cannot be learned from the existing traces.

We also implemented our approach of generating psycho-linguistic embedding in an unsupervised way in other domains. Specifically, we leverage the approach in identifying depressed users in online forums as explained in Appendix D.

CHAPTER 6:

**TEXTUAL CHARACTERISTICS OF NEWS
TITLE AND BODY TO DETECT FAKE NEWS:
A REPRODUCIBILITY STUDY**

6.1 Introduction

Social media and online news sources have become the major source of news diet for the increasingly large population instead of traditional media. In 2019, the Pew Research Center reported that more than half (55%) of American adults consume news from online platforms often or sometimes, which is 8% increase since 2018 [171]. With its increase in popularity, social media have also been proven to be an effective platform for fake news proliferation due to its lower cost and convenience of further sharing [30], which has attracted the attention of researchers, making it a global topic of interest. Several studies have been carried out to determine the validity of news relying on linguistic cues derived from the readability and lexical information of the news content [32, 1, 31].

Horne and Adali [1] conducted a study to understand and analyze the associated language patterns of the title and content of fake news. This chapter has gained a lot of attention by the research community, with over 200 citations according to

Google Scholar, and became the reference reading to understanding textual content differences between real and fake news. Horne and Adali witnessed that the general assumption about fake news that it is written to camouflage with real news and deceive the reader who does not care about the news sources' veracity is actually not true. In fact, they found the fake news is more similar to satire than to real news, and the focus of fake news is on users who are unlikely to read beyond the title. This sheds light on the necessity of research to understand the significant difference between the title of fake and real news separately from the news body content to mitigate the possible diffusion of the fake news. However, these claims were established based on a small data used in which labels were assigned according to the credibility of the news source, instead of fact-checking, which does not consider the fact that a news source can have mixed credibility and publish both real and fake information.

Thus, we decided to reproduce the paper by Horne and Adali [1] to validate their findings on larger state-of-the-art datasets with labels provided by professional journalists who have fact-checked the news, namely PolitiFact and GossipCop [53] and BuzzFeedNews [31]. Because the news trends continuously evolve, we analyze, similarly to Horne and Adali, news text (from body and title) by focusing on linguistic style, text complexity, and psychological aspects of the text, rather than topic-dependent representations of documents (e.g., [109]). In addition, we expanded the set of emotion features considered in the original paper to explore this aspect of the text further, given that Ghanem et al. [9] recently showed emotions play a key role in detecting false information. We also compare the classification performance of different classifiers beyond linear SVM (the only model used in [1]), and we discuss textual differences between two news domains, namely political and gossip news.

Our experiments confirm most of the original paper’s findings regarding title and body feature differences between fake and real news, e.g., fake political news packs a lot in the title. However, differently from Horne and Adali, we found that fake titles contain more stop words than real titles. When using linear SVM to classify fake vs. real news, we confirm that title features outperform body features, but we observe the opposite results if we consider a non-linear and more expressive classifier such as Random Forest.

Furthermore, we show new patterns that were not present in the paper by Horne and Adali, namely fake news title and body express more negative emotions and sentiment than real news, and real news articles are more descriptive than fake news ones. Also, we highlight some differences between two different news domains: political and gossip. For instance, among stylistic, psychology, and complexity features in the news title, psychology features are the most important group of features for gossip news, while the most important group for political news is the one containing stylistic features. This shows how gossip news titles tend to be more persuasive than other news domains.

6.2 Overview of the paper by Horne and Adali

In this section, we provide an overview of the approach, features, and findings by Horne and Adali [1].

6.2.1 Approach

Horne and Adali conducted a content analysis to study fake news by analyzing three small datasets: (i) a dataset (DS1) created by BuzzFeed leading to the 2016 U.S. elections which contains 36 real news stories and 35 fake news stories; (ii) a dataset

(DS2) created by using Zimdars' list of fake and misleading news websites [172] and fact-checking website like snopes.com [1], containing 75 stories for each category: real, fake and satire sources; (iii) a dataset (DS3) containing 4000 real and 233 satire articles from a previous study [173]. During the experiments, they considered features from both news body and title for determining the veracity of news and comparing real news vs. fake news vs. satire.

6.2.2 Features

This research focused on three groups of features, including stylistic features (syntax, text style, and grammatical elements measured by 2015 Linguistic Inquiry and Word Count (LIWC) [174] and the Python Natural Language Toolkit Part of Speech tagger [175]), complexity features to capture details about how complex the article or title is (e.g., words per sentence, syntax tree depth determined by the Stanford Parser and readability level of text), and psychological features to capture emotional (positive/negative), social, and cognitive processes incorporated in news body or title computed by using the LIWC tool. Sentiment analysis was done through SentiStrength [176].

Feature Selection and Analysis. The goal of feature selection is to avoid overfitting and increase generalizability. Because the datasets were small and the features generated were large, Horne and Adali performed feature selection by leveraging the one-way ANOVA test for those normally distributed features and the Wilcoxon rank-sum test for those that did not pass the normality test. This feature selection concluded with the selection of top 4 features for news body (number of nouns, lexical diversity (TTR), word count, and number of quotes) and news title (percentage of

stop words, number of nouns, average word length, and Flesh-Kincaid Grade Readability Index).

Besides, they also used the above mentioned statistical tests to uncover statistically significant feature value differences among news with different labels (fake, satire, and real). If the value of a feature was higher (on average) for real news articles as compared to fake news articles, they denoted this by $R > F$ (and $F > R$ vice versa). We used the same notation while reproducing this experiments in Tables 6.2 and 6.3.

6.2.3 Observation and Evaluation

Horne and Adali’s findings show how real news is different from fake and satire news and that fake news and satire have a lot in common across several dimensions. Regarding real vs. fake news (which is the scope of our reproducibility paper), they found that:

- (f1) fake news articles tend to be shorter in terms of content, but use repetitive language,¹ smaller words, less punctuation, and fewer quotes (these results is consistent between datasets DS1 and DS2);
- (f2) fake news articles require a lower educational level to read, use fewer analytic words, use more personal pronouns and adverbs, but fewer nouns (this result is not consistent between datasets DS1 and DS2 and it is less significant);
- (f3) fake titles are longer, contain shorter words, use more all capitalized words,

¹Repetitive language is measured by using the Type-Token Ratio (TTR) which is the number of unique words in the document by the total number of words in the document. A low TTR means more repetitive language, while a high TTR means more lexical diversity. Horne and Adali claim fake news has more repetitive language but show the opposite result in their paper, i.e., TTR is on average higher for fake than real news (cf. Table 4 in [1]), indicating more lexical diversity for fake than real news. Our results confirms more lexical diversity for fake news as shown in Table 6.2.

fewer stop words, and fewer nouns overall but more proper nouns (these results is consistent between datasets DS1 and DS2);

- (f4) titles are a strong differentiating factor between fake and real news. They performed a binary classification of real vs. fake news separately on news body content and title on dataset DS2. They used the top 4 features from the feature selection process to run a linear SVM model with 5-fold cross-validation. The classification results show 71% accuracy for news body content and 78% accuracy for the title. Thus, they argued that the title is more important in predicting fake vs. real news, and the title and the body of the news should be analyzed separately.

6.3 Reproducibility

In this section, we describe in detail our attempt to reproduce and generalize findings (f1)-(f4) shown by Horne and Adali in their paper [1].

6.3.1 Datasets

There is generally limited availability of large scale benchmarks for fake news detection, especially where the ground truth labels are assigned via fact-checking, which is a time-consuming activity. FakeNewsNet [53] and BuzzFeedNews [31] are the only publicly available datasets having fact-checked labels. Thus, in this chapter, we use these datasets to conduct our study.

FakeNewsNet: PolitiFact and GossipCop.

FakeNewsNet consists of two datasets, PolitiFact and GossipCop, from two different domains, i.e., politics and entertainment gossip, respectively. Thus, we used these

Table 6.1: Size of datasets used in our study.

Dataset	# Total News	# Fake News	# Real News
PolitiFact	838	378	460
BuzzFeedNews	1,561	299	1,262
GossipCop	19,759	4,734	15,025

two datasets separately in our study. Each of these datasets contains details about news content, publisher information, and social engagement information. We only used news content information in this chapter.

PolitiFact contains news with known ground truth labels collected from the fact-checking website PolitiFact.² After cleaning the dataset from missing news bodies or titles, we obtained a total of 838 news articles, 378 fake and 460 real.

The GossipCop dataset contains fake news collected from GossipCop³, which is a fact-checking website for entertainment stories and real news collected from E!Online,⁴ a trusted media website for entertainment stories. After cleaning the dataset from missing news bodies or title, we obtained a total of 19,759 news articles, 4,734 fake and 15,025 real.

BuzzFeedNews Dataset.

The BuzzFeedNews dataset contains news regarding the 2016 U.S. election published on Facebook by nine news agencies. This dataset⁵ contains 1,262 articles that are mostly true, 212 that are a mixture of true and false, and 87 that are false, after cleaning the dataset from missing news bodies or titles. Ground truth is derived from

²<https://www.politifact.com/>

³<https://www.gossipcop.com/>

⁴<https://www.eonline.com/ap>

⁵The BuzzFeedNews dataset is available at <https://zenodo.org/record/1239675#.X5riw0JKgXA>

professional journalists at BuzzFeed who have fact-checked the news in the dataset. As also done in the other datasets, we considered false news and news with a mixture of true and false as fake news and mostly true news as real news.

6.3.2 Features

This section describes the set of features we used in the chapter to analyze real vs. fake news. In our implementation, we consider features similar to Horne and Adali [1], namely stylistic features, text complexity features, and psychology features. These features are computed for both the title and body text of the news.

Stylistic Features.

We used the subset of LIWC features that represent the functionality of text, including word count (WC), words per sentence (WPS), time orientation (e.g., focus on past (focuspast) and focus on future (focusfuture)), number of personal (I, we, you, she/he – one feature each) and impersonal pronouns, number of quantifying words (quant), number of comparison words (compare), number of exclamation marks (exlam), number of negations (negate), e.g., no, never, not, number of swear words (swear), number of online slang terms (netspeak), e.g., lol, brb, number of interrogatives, e.g., how, what, why (interrog), number of punctuation symbols (allPunc), number of quotes (quote).

Regarding the part of speech features, we used the Python Natural Language Toolkit part of speech (POS) tagger to compute the number of nouns (NN), proper nouns (NNP), personal pronouns (PRP), possessive pronouns (PRP\$), Wh-pronoun (WP), determinants (DT), Wh-determinants (WDT), cardinal numbers (CD), adverbs (RB), interjections (UH), verbs (VB), Adjective (JJ), past tense verbs (VBD),

gerund or present participle verbs (VBG), past participle verbs (VBN), non-3rd person singular present verbs (VBP), and third-person singular present verbs (VBZ).

This stylistic group of features also includes the upper case word count (all caps) and percent of stop words (per_stop).

Psychology Features.

We computed these features by using the LIWC tool and include the number of analytic words (analytic), insightful words (insight), causal words (cause), discrepancy words (discrep), tentative words (tentat), certainty words (certain), differentiation words (differ), affiliation words (affil), power words, reward words, risk words, personal concern words (work, leisure, religion, money, home, death – one each), anxiety-related words (anx), emotional tone words (tone), and negative (negemo) and positive (posemo) emotional words. This group of features also includes positive (pos) and negative (neg) sentiment metrics as computed by the VADER sentiment analysis tool [177]. We also investigated the importance of features describing emotions expressed through the text, as Ghanem et al. [9] recently showed emotions play a key role in deceiving the reader and can successfully be used to detect false information. Thus, in addition to some emotion features provided by the LIWC tool (as described above), we computed additional emotion features such as anger, joy, sadness, fear, disgust, anticipation, surprise, and trust by using the Emotion Intensity Lexicon (NRC-EIL) [178] and the approach proposed in [179].

Complexity Features.

The complexity of text in natural language processing depends on how easily the reader can read and understand a text. We used popular readability measures as complexity features in our analysis: Flesh Kincaid Grade Level (FK), Gunning Fog Index (GI), Simple Measure of Gobbledygook Index (SMOG). Higher scores of these readability measures indicate that the text is easier to read. This group of features also includes lexical diversity or Type-Token Ratio (TTR) and the average length of each word (avg wlen).

6.3.3 Analysis

Considering all the features from each group, we have a total of 68 features, which can still be too many for the size of the considered datasets (PolitiFact, BuzzFeedNews, and GossipCop) to perform a real vs. fake news articles classification. Therefore, we used the same statistical tests (ANOVA and Wilcoxon rank-sum) used by Horne and Adali to perform feature selection and analysis. For each dataset, features are sorted by F-value in descending order to determine the importance, and only features where the two averages (real vs. fake) were significantly different according to the statistical test (p -value < 0.05) were considered. Among these features, we selected a number of features up to the square root of the training set size (rule of thumb) for both news body and title to feed the classification algorithm.

Instead of just using the linear SVM classifier as done by Horne and Adali, we compared the performances of different classification algorithms, namely Logistic Regression (LR) classifier with L2 regularization, linear Support Vector Machine (SVM), and Random Forest (RF), with default parameters. As the datasets we considered

are not balanced, we used class weighting to deal with class imbalance, stratified 5-fold cross-validation, and results are reported by using AUROC and average precision (AvgP).

Table 6.2: Features that differ in body of news content. ($p < 0.05$).

Features	PolitiFact	BuzzFeed	GossipCop	Features	PolitiFact	BuzzFeed	GossipCop
allPunc	$R > F$	$R > F$	$R > F$	analytic	$F > R$	$R > F$	$R > F$
exclam	$F > R$	$F > R$	$F > R$	quote	$F > R$	$R > F$	$F > R$
tone	$R > F$	$R > F$	$R > F$	WC	$R > F$	$R > F$	$R > F$
WPS		$R > F$	$R > F$	affect		$F > R$	$R > F$
affil	$R > F$		$F > R$	cause		$F > R$	$F > R$
certain		$F > R$	$F > R$	all caps	$R > F$	$R > F$	$R > F$
differ	$R > F$	$F > R$	$F > R$	discrep	$R > F$	$F > R$	$F > R$
FK		$R > F$		focusfuture			$F > R$
GI		$R > F$	$F > R$	i			$R > F$
insight		$F > R$		interrog			$R > F$
leisure	$F > R$		$R > F$	TTR	$F > R$	$F > R$	
money	$R > F$			negate		$F > R$	$F > R$
netspeak			$R > F$	JJ	$R > F$	$R > F$	$R > F$
RB	$R > F$	$R > F$		CD	$R > F$	$R > F$	$R > F$
DT	$R > F$	$R > F$	$R > F$	UH	$R > F$		
NN	$R > F$	$R > F$	$R > F$	NNP	$R > F$	$R > F$	$R > F$
PRP	$R > F$	$R > F$	$R > F$	PRP\$	$R > F$	$R > F$	
VBD	$R > F$	$R > F$	$R > F$	VBG	$R > F$	$R > F$	
VBN	$R > F$	$R > F$		VBP	$R > F$	$R > F$	$R > F$
VBZ	$R > F$	$R > F$		VB	$R > F$	$R > F$	$R > F$
WP	$R > F$	$R > F$	$R > F$	WDT	$R > F$	$R > F$	$R > F$
per_stop	$F > R$	$F > R$	$F > R$	power		$R > F$	$R > F$
quant	$R > F$			relig	$F > R$	$F > R$	$R > F$
reward			$R > F$	risk			$F > R$
sheshe	$F > R$		$F > R$	SMOG		$R > F$	$F > R$
swear	$F > R$	$F > R$		tentat		$F > R$	$F > R$
we	$R > F$		$R > F$	avg wlen		$R > F$	
work	$R > F$	$R > F$		you	$R > F$	$F > R$	$R > F$
compare		$R > F$		focuspast	$F > R$		$F > R$
neg	$F > R$	$F > R$	$F > R$	surprise	$F > R$		
disgust	$F > R$	$F > R$	$F > R$	negemo	$F > R$	$F > R$	$F > R$
pos	$R > F$		$R > F$	fear	$F > R$	$F > R$	
posemo	$R > F$		$R > F$	anx	$F > R$	$F > R$	$F > R$
sadness	$F > R$	$F > R$	$F > R$	anger		$F > R$	$F > R$
trust			$F > R$	joy			$F > R$

Table 6.3: Features that differ in the title of news content. All differences are statistically significant ($p < 0.05$).

Features	PolitiFact	BuzzFeed	GossipCop	Features	PolitiFact	BuzzFeed	GossipCop
WC	$F > R$	$F > R$		avg wlen	$R > F$	$R > F$	$F > R$
quote	$F > R$	$F > R$	$F > R$	allPunc	$R > F$		$F > R$
exclam	$F > R$	$F > R$	$F > R$	tone	$R > F$	$R > F$	$R > F$
WPS	$F > R$	$F > R$	$R > F$	affect	$F > R$		$R > F$
affil			$F > R$	compare	$F > R$		$R > F$
differ			$F > R$	discrep	$F > R$		$F > R$
focusfuture	$F > R$		$F > R$	focuspast	$F > R$	$F > R$	
insight		$F > R$		interrog			$R > F$
leisure			$R > F$	TTR	$F > R$		$F > R$
money		$R > F$		negate			$F > R$
netspeak	$R > F$		$R > F$	JJ		$R > F$	$R > F$
UH			$F > R$	GI	$F > R$		$F > R$
FK	$F > R$		$F > R$	SMOG	$F > R$		$F > R$
analytic		$R > F$	$R > F$	all caps	$F > R$	$F > R$	
NN		$R > F$	$R > F$	NNP	$F > R$	$F > R$	$F > R$
PRP	$F > R$	$F > R$		PRP\$	$F > R$	$F > R$	$R > F$
DT			$R > F$	RB	$F > R$		$F > R$
VBD	$F > R$			VBG	$F > R$		$F > R$
VBN	$F > R$			VBP	$F > R$	$F > R$	
VBZ	$F > R$	$R > F$		VB	$F > R$		$F > R$
WP		$F > R$		per_stop	$F > R$	$F > R$	
quant			$R > F$	relig	$F > R$	$F > R$	
reward			$R > F$	risk			$F > R$
work	$R > F$	$R > F$		i	$F > R$		$R > F$
you			$R > F$	shehe	$F > R$	$F > R$	
CD			$R > F$	fear	$F > R$	$F > R$	$F > R$
neg	$F > R$	$F > R$	$F > R$	sadness	$F > R$	$F > R$	$F > R$
surprise	$F > R$	$R > F$		anger	$F > R$	$F > R$	$F > R$
negemo	$F > R$		$F > R$	trust	$R > F$		$R > F$
disgust	$F > R$	$F > R$	$F > R$	pos			$R > F$
posemo			$R > F$	anx			$F > R$
joy			$R > F$				

6.3.4 Results

Feature Statistical Analysis. We start our analysis by checking whether Horne and Adali’s findings (f1), (f2), and (f3) reported in Section 6.2.3 are confirmed in the three larger datasets we considered, namely PolitiFact and BuzzFeed (political news

Table 6.4: News title vs. news body features for detecting fake news on the PolitiFact, BuzzFeedNews, and GossipCop datasets: stylistic, psychology, and complexity features. Best results for both news title and body are in bold. Best overall results between news title and body are shaded.

	PolitiFact		BuzzFeedNews		GossipCop	
Features	AUROC	AvgP	AUROC	AvgP	AUROC	AvgP
News body (SVM)	0.583	0.466	0.614	0.257	0.623	0.327
News body (LR)	0.855	0.809	0.728	0.351	0.703	0.437
News body (RF)	0.911	0.878	0.785	0.417	0.782	0.630
News Title (SVM)	0.833	0.804	0.669	0.317	0.588	0.309
News Title (LR)	0.849	0.813	0.787	0.423	0.663	0.380
News Title (RF)	0.867	0.823	0.812	0.424	0.715	0.490

datasets), and GossipCop (gossip news dataset). To analyze these findings we refer to the results reported in Table 6.2 for news body text and Table 6.3 for news title.

Regarding finding (f1) (cf. Table 6.2), we confirm that fake news articles have a shorter content (WC) and use less punctuation (allPunc) than real news articles in all the three datasets we considered, and fake political articles have more lexical diversity (TTR) than real political articles. Our analysis does not allow us to generalize the finding that fake news articles use smaller words (avg wlen) and fewer quotes (true in BuzzFeedNews, but not in Politifact and GossipCop).

Regarding finding (f2) (cf. Table 6.2), we can generalize the finding that fake news articles use fewer analytic words (true in BuzzFeedNews and GossipCop). We found that fake news articles require a lower educational level to read (as measured by FK, GI, and SMOG readability indexes) only in one dataset (BuzzFeedNews) while the opposite trend holds for GossipCop dataset; the use of more personal pronouns (PRP), adverbs (RB), and proper nouns (NNP) in fake news articles is not confirmed in our analysis. We observe fake titles containing more proper nouns (NNP) in all

Table 6.5: News title vs. news body features for detecting fake news on the PolitiFact, BuzzFeedNews, and GossipCop datasets: same four features as in Horne and Adali [1] – NN, TTR, WC, and Quote for news body and FK, NN, per_stop, and avg_wlen for title. Best results for both news title and body are in bold. Best overall results between news title and body are shaded.

	PolitiFact		BuzzFeedNews		GossipCop	
Features	AUROC	AvgP	AUROC	AvgP	AUROC	AvgP
News Body (SVM)	0.544	0.445	0.678	0.292	0.500	0.232
News Body (LR)	0.754	0.663	0.691	0.297	0.534	0.251
News Body (RF)	0.861	0.803	0.708	0.342	0.631	0.42
News Title (SVM)	0.649	0.531	0.713	0.342	0.528	0.250
News Title (LR)	0.643	0.530	0.716	0.342	0.530	0.251
News Title (RF)	0.735	0.612	0.706	0.330	0.582	0.332

the three datasets considered.

Regarding finding (f3) (cf. Table 6.3), we confirm that fake titles have more proper nouns (NNP) than real titles in all the three datasets we considered and have fewer nouns (NN) in BuzzFeedNews and GossipCop. Also, we confirm that fake political titles are longer (WC and WPS), use more capitalized words (all caps) (they also use more possessive pronouns – PRP\$), and contain shorter words (avg_wlen). Our analysis does not confirm the fact that fake titles contain fewer stop words (per_stop). Similarly, we observe that fake news articles contain more stop words.

Furthermore, our results in Tables 6.2 and 6.3 highlight new patterns that were not present in the analysis performed by Horne and Adali. Specifically, we found that real news articles use a more positive tone and more nouns (NN), determinants (DT), wh-determinants (WDT), verbs (VB), past tense verbs (VBD), Wh-pronouns (WP), and adjectives (JJ) in all the three datasets considered. This indicates that real news articles are more descriptive than fake news articles. Also, fake news titles

Table 6.6: Feature group ablation for news title and body when the best classifier (Random Forest) is used on the PolitiFact, BuzzFeedNews, and GossipCop datasets. Best results for both news title and body are in bold.

	PolitiFact		BuzzFeedNews		GossipCop	
Features	AUROC	AvgP	AUROC	AvgP	AUROC	AvgP
News Body						
Stylistic (RF)	0.882	0.838	0.753	0.382	0.752	0.590
Psychology (RF)	0.723	0.662	0.681	0.319	0.713	0.509
Complexity (RF)	0.804	0.708	0.630	0.285	0.000	0.000
News Title						
Stylistic (RF)	0.819	0.729	0.805	0.433	0.634	0.365
Psychology (RF)	0.791	0.691	0.645	0.320	0.651	0.407
Complexity (RF)	0.583	0.486	0.555	0.257	0.553	0.287

and bodies use more exclamation marks (exclam) than real news titles (true in all the three datasets considered).

In addition, we observe that fake titles express more negative emotions (anger, sadness, fear, and disgust) and negative sentiment (neg) than real titles consistently across all the three considered datasets. This pattern is also true for fake news body. In contrast, real titles tend to express more positive emotions (trust, posemo, joy) and positive sentiment (pos), but this is less consistent across datasets. When selecting information, people have a sensitivity to negative information [180]. This negativity bias induces people to pay more attention to negative news, hence fake news titles, bodies, and even associated images [181] express negative emotions to be catchier and circulate more among people.

Furthermore, there are some differences between political and gossip news. We found that fake political news articles have more religion-related words (relig) than real political news articles, while fake gossip news articles have fewer religion-related

words; fake political news titles contain shorter words (avg wlen), and more words per sentence (WPS) and possessive pronouns (PRP\$) than real political news titles, while this is the opposite for gossip news titles.

Real vs. Fake News Classification. Finding (f4) by Horne and Adali claims that title features are more informative (i.e., achieve higher accuracy) than news body features in classifying fake vs. real news with a linear SVM. Table 6.4 shows our classification results by comparing three classifiers, and when we used a number of features up to the square root of the training set size. We observe that when we consider the linear SVM classifier, finding (f4) is confirmed, i.e., AUROC and average precision scores are higher for the title than the news body. However, Random Forest is the best classifier for both news body and title and outperforms linear SVM. When we consider Random Forest as the classifier, finding (f4) is reversed, i.e., AUROC and average precision scores are higher for news body than news title (this is true for two out of three of the datasets considered). We observe a similar trend also when we consider only the four features chosen by Horne and Adali to perform the classification (see results reported in Table 6.5). Of course, considering more than four features as we did in Table 6.4 results in better AUROC and average precision in all the three datasets.

Thus, our experiments reveal that whether or not the title is more informative than the news body depends on the chosen classifier. A non-linear classifier such as Random Forest has higher expressive power and outperforms linear SVM. Thus, if we choose the best classifier, namely Random Forest, finding (f4) does not hold in the larger datasets we considered. Having more information helps the Random Forest classifier to increase classification performances.

In addition, we performed feature ablation by feature group (style, psychology, and complexity) when the best classifier (Random Forest) is used. Results are reported in Table 6.6. We observe that stylistic features are the most important features in both title and news body for political news. For gossip news, stylistic features are the most important news body features, while psychology features are the most important features in title. Interestingly, this validates the definition of gossip as “small talk” that is originated from evolutionary psychology and has the basic intent to share information about third persons to indulge people in some discussion. Also, the reason people like gossip is because it is tempting and fun. Thus, the news title of gossip stories are written with more psychological words like tone and affect, e.g., “Angelina Jolie Can’t Get Over Heartbreak Of Losing Brad Pitt — Real Reason For Fury, Says Source” to catch readers attention even though the body text is not that engaging.

6.4 How to Reproduce our Experiments

For reproducibility propose, we made our code available in a GitHub repository.⁶ Because we did not directly collect the datasets, we are not uploading them in our repository, but we provide instructions on finding and downloading them. In our repository, we make our code available for extracting the features that are considered in this chapter, including complexity, stylistic and psychology features extracted using NLTK part-of-speech, VADER Sentiment Analyser and the Emotion Intensity Lexicon (NRC-EIL),⁷ except LIWC features as the LIWC tool has proprietary dictionaries whose licence should be purchased. LIWC features can be computed in two

⁶<https://github.com/shresthaanu/ECIR21TextualCharacteristicsOfFakeNews>

⁷The NRC-EIL lexicon should be downloaded at <https://www.saifmohammad.com/WebPages/AffectIntensity.htm>

ways: (1) by using the software tool to compute the features, or (2) by downloading the dictionary provided by the tool for which we have provided code to extract features using the dictionary. In addition, we also provide code for the statistical test performed in this chapter to reproduce Tables 6.2 and 6.3. Likewise we also provide code for the classification to reproduce Tables 6.4, 6.5 and 6.6.

6.5 Conclusions

In this chapter, we reproduced the study by Horne and Adali [1] of the relative importance of news body and title in detecting fake news. We extended their experimental setting by using larger real and fake news datasets with ground truth at the news level, considering additional features describing emotions expressed through the text, comparing different classification algorithms, and highlighting differences between political and gossip news domains. Our experiments have shown that some of the original paper’s observations are not the same as the trend of news writing is continuously evolving. For instance, the finding that the news title is more informative and plays an important role in discerning the news’s veracity is confirmed if we use the same classifier, linear support vector machine (SVM), as in [1], but using a non-linear classifier such as Random Forest reverses the finding. Finally, we provide evidence that fake news title and body attract readers’ attention with more negative emotions and sentiment, while real news articles are more descriptive.

CHAPTER 7:

CHARACTERIZING AND PREDICTING FAKE NEWS SPREADERS IN SOCIAL NETWORKS

7.1 Introduction

Online social media platforms such as Facebook and Twitter have drastically changed the landscape of news consumption and the pattern of information flow in the past decade. The majority of the population relies on social media for news on important events, breaking news, and emergencies. According to Pew Research Center 71% of American adults ever get news through social media in 2020 [182]. With the increase in its popularity, social media has significantly transformed the way of creating news content, user interactions, and engagement, reshaping the traditional medium to whole new information ecosystems[183]. Individuals in social media actively participate in creating and sharing news items due to its ease of use, lower cost, and convenience of further sharing [184, 30]. This shift of the news paradigm has led to an unprecedented transformation in both news quality and quantity that users encounter in social media, reducing the credibility of news articles and eventually fostering the production and dissemination of misinformation.

Indeed, the rapid spread of fake news has become a concerning problem in online social networks in recent years. Research has found that fake news is more likely to go

viral than real news, spreading both faster and wider [64] and people engage more with fake news than real news [46]. Moreover, the worrisome amount of fake news widely spreading over social media can negatively influence users' opinions creating threats on public health [185], emergency management and response [186, 24], election outcomes [187], and is responsible for a general decline in trust that citizens of democratic societies have for online platforms [188]. Surprisingly, bots are equally responsible for spreading real and fake news, and human activity causes the considerable spread of fake news on Twitter [64, 66] as people are generally not able to accurately identify which news item is fake and which is real [189]. Thus, fake news is successful mainly because people are not able to disguise it from truthful information [57, 86] and often share news online without even reading its content [19]. Also, even if people recognize news as fake, they are more likely to share it if they have seen it repeatedly than the news that is novel [84].

Thus, identifying fake news spreaders in social networks is one of the key aspects to mitigate misinformation spread effectively. Examples of strategies that could be implemented include assisting fake news spreaders with credibility indicators to lower their fake news sharing intent [102], and mitigation campaign, e.g., target the most influential real news spreader to maximize the spread of real news [51]. However, less is known about the characteristics of fake or real news spreaders.

Therefore, in this article, we seek to understand the characteristics of fake news spreaders focusing on different attributes such as user writing style, emotions, demographics, personality, social media behavior, and network features. In particular, we leveraged these attributes to perform a comprehensive analysis on two different datasets, namely PolitiFact [51] and PAN [52] to investigate the patterns of user char-

acteristics in social media in the presence of misinformation. We hypothesize that users likely to share fake news hold specific patterns based on these attributes which are different from real news spreaders. To the best of our knowledge, some of the features we considered, such as user stress, needs, values, and tweeting behavior, have not been analyzed before. Furthermore, we investigate to what extent these features can be used to identify users who are likely to share fake news by addressing the problem as a binary classification task.

Our analysis unveils some interesting characteristics of fake news spreaders across the two datasets considered. Specifically, our results show that:

- The majority of users under 18 or over 40 may tend to share more fake than real news.
- Female users may tend to be more fake news spreaders than male users.
- The political orientation of a fake news spreader is more likely to coincide with the source's political bias of the majority of circulating fake news items.
- Fake news spreaders (1) have newer accounts, (2) spend, on average, less time between two consecutive tweets, and (3) tend to tweet more at night.
- Fake news spreaders tend to express more negative emotions and stress in their tweets than real news spreaders.
- Fake news spreaders are estimated to be more extroverted and less neurotic than real news spreaders.
- Classification results using our proposed features outperform the results of baseline approaches with n-grams in both datasets. Specifically, we show that our

proposed features can identify fake news spreaders with an average precision of 0.99 on the PolitiFact dataset (vs. 0.96 achieved by the best baseline) and 0.79 on the PAN dataset (vs. 0.78 achieved by the best baseline).

- Emotions and personality features are strong predictors of fake news spreaders in all the considered datasets.

The article is organized as follows. Section 7.2 summarizes related work, Section 7.3 describes the dataset we used in this article, Section 7.4 presents our proposed features to characterize and classify fake news spreaders, Section 7.5 presents the user characteristics patterns that we found by analyzing the considered datasets, Section 7.6 reports on our experimental evaluations and, finally, conclusions are drawn in Section 7.9.

7.2 Related Work

Several studies have been conducted to understand the characteristics of users that are likely to contribute to spreading fake news on social networks. Vosoughi et al. [64] revealed that the fake news spreaders had, on average, significantly fewer followers, followed significantly fewer people, and were significantly less active on Twitter. Moreover, bots tend to spread both real and fake news, and the considerable spread of fake news on Twitter is caused by human activity. Shrestha and Spezzano showed that social network properties help in identifying active fake news spreaders [65]. Shu et al. [66] analyzed user profiles to understand the characteristics of users that are likely to trust/distrust fake news. They found that, on average, users who share fake news tend to be registered for a shorter time than the ones who share real news and that bots are more likely to post a piece of fake news than a real one, even though users

who spread fake news are still more likely to be humans than bots. They also show that real news spreaders are more likely to be more popular and that older people and females are more likely to spread fake news. Guess et al. [67] also analyzed user demographics as predictors of fake news sharing on Facebook and found out political-orientation, age, and social media usage to be the most relevant. Specifically, people are more likely to share articles they agree with (e.g., right-leaning people tended to share more fake news because the majority of the fake news considered in the study were from 2016 and pro-Trump), seniors tend to share more fake news probably because they lack digital media literacy skills that are necessary to assess online news truthfulness. The more people post on social media, the less they are likely to share fake news, most likely because they are familiar with the platform and they know what they share.

Shrestha et al. [68] analyzed the linguistic patterns used by a user in their tweets and personality traits as a predictor for identifying users who tend to share fake news on Twitter data [69, 68]. Likewise, Giachanou et al. [75] proposed an approach based on a convolutional neural network to process the user Twitter feed in combination with features representing user personality traits and linguistic patterns used in their tweets to address the problem of discriminating between fake news spreaders and fact-checkers.

Ma et al. [190] went beyond the user and news characteristics and analyzed the characteristics of diffusion networks to explain users' news sharing behavior. They found opinion leadership, news preference, and tie strength to be the most important factors at predicting news sharing, while homophily hampered news sharing in users' local networks. Also, people who are driven by gratifications of information seek-

ing, socializing, and status-seeking were more likely to share news on social media platforms [191].

Moreover, creating hashtags has been widely used to organize campaigns, sharing information and opinion about events and news stories on social media. These hashtags have also been used to draw attention and enhance the topic’s visibility, eventually causing its wide spread over social media. Both individuals and news organizations have capitalized on this feature of social media via the massive use of political hashtags to increase readership and user engagement [192]. This target turns true and amplifies if a user shares the piece of news with partisan affiliation [193]. Thus, the political orientation of a user can provide additional cues about the user being a fake news spreader or not.

As compared to previous work, which has been mainly done on the PAN 2020 dataset [52], this article addresses the problem of characterizing and predicting users that are keen to spread fake news on an additional larger dataset with more reliable ground truth extracted from FakeNewsNet [51]. We consider several groups of topic-agnostic features, including new features that have not been used in previous work, such as behavioral features, stress, needs, and values, to profile and predict fake news spreaders on two datasets and evaluate the relative importance of the considered groups of features. We also highlight feature patterns that are common to both datasets.

7.3 Datasets

This section describes the datasets we used to carry out our experiments, namely the PAN 2020 and PolitiFact (FakeNewsNet) datasets. The size of these datasets is shown in Table 7.1.

PAN 2020 Dataset. The first dataset we consider is the one provided by the PAN CLEF¹ 2020 shared task on profiling fake news spreaders on Twitter [52]. The dataset has been collected in two languages, namely English and Spanish, and consists of a balanced train and test set for each language. For each considered language, the training set includes 300 Twitter users and 100 tweets for each user from their Twitter feed, resulting in 30,000 English tweets and Spanish 30,000 tweets. The test set contains 200 users in each language and 100 tweets from their feed for each user, resulting in 20,000 English tweets and 20,000 Spanish tweets. In this article, we have considered only the English dataset and combined the train and test set together in a unique (balanced) dataset.

In the PAN 2020 dataset, users that shared fake news in the past are labeled as fake news spreaders and real news spreaders, otherwise. However, it is worth noting that, because the dataset is GDPR compliant², users are labeled as “class 0” or “class 1,” and the authors of the dataset did not disclose which one of the two labels corresponds to the class of users who are fake news spreaders. In this article, we assumed one of the two labels to identify fake news spreaders according to feature patterns that result similar to the ones of the PolitiFact dataset. These patterns are described in Section 7.5.

PolitiFact (FakeNewsNet). The FakeNewsNet dataset consists of two datasets, PolitiFact and GossipCop, from two different domains, i.e., politics and entertainment gossip, respectively [51]. Each of these datasets contains details about news content, publisher, social engagement information, and user social network. In this article, we

¹PAN CLEF (<https://pan.webis.de/>) is a well-known forum that focuses on applying text mining for user profiling.

²<https://www.privacyshield.gov/article?id=European-Union-Data-Privatization-and-Protection>

only used the PolitiFact dataset, which contains news with known ground truth labels collected from the fact-checking website PolitiFact³ where journalists and domain experts fact-checked the news items as fake or real. We decided not to use the GossipCop dataset because in our previous work [194] we found that gossip news is quite different than political news; hence we focused our attention on the same news domain as the other dataset we considered, i.e., the PAN 2020 dataset. Overall, the considered PolitiFact dataset contains 295,469 users (after removing self-claimed bot accounts) sharing 701 news items via tweets and retweets. As this dataset only provides ground truth for news, we computed the labels for the users (fake news spreader or real news spreader) as explained here below. First, we filtered out those users who had shared the same news item multiple times, and then we selected only those users who had shared at least eight unique news items. We manually analyzed the profiles of users who shared the same news item multiple times and found that they were bots; hence we excluded them from our analysis as research has shown that false news spreads more than the truth because of humans, not bots [64]. Next, the resulting group of 1,046 users is labeled as fake news spreaders or real news spreaders as follows: (1) a user is a fake news spreader if at least 60% of the news items they shared are fake, or (2) a user is a real news spreader if at least 60% of the news items they shared are real.⁴ We labeled 648 users as fake news spreaders and 398 as real news spreaders. Moreover, we retrieved additional user data as follows. For each user who did not have enough tweets, i.e., more than 100 words among all their tweets combined, we crawled all tweets posted one month prior to his first tweet creation

³<https://www.politifact.com/>

⁴One limitation of this labeling approach is that we may not catch fake news spreaders who camouflage themselves as real news spreaders through their news sharing behavior.

Dataset	# Fake News Spreader	# Real News Spreader
PAN 2020	250	250
FakeNewsNet (PolitiFact)	648	398

Table 7.1: Datasets and statistics.

time in our dataset. These additional tweets were utilized to generate personality features and political orientation.

7.4 Features

This section describes the features we analyzed to characterize and classify fake news spreaders in the two datasets considered. Specifically, we study users according to six user features groups: demographics, Twitter behavior and network, emotions, personality, readability, and writing style. Text-based features such as emotions, personality, and readability are computed on the document resulting from the concatenation of all the user tweets. To have a more accurate estimation of user emotions, personality, readability, and writing style, retweets are excluded when computing these features.

7.4.1 Demographics

The first group of features we consider deals with user demographics, including age, gender, and political orientation. Previous work has shown how these features influence users' news-sharing behavior. For instance, Reis et al. [195] show that white and male users potentially share more news on Twitter. Differently, Shu et al. [66] analyzed user profiles to understand the characteristics of users that are likely to trust/distrust fake news and propagate them on Twitter. They also show that older people and females are more likely to spread fake news.

Demographic features are often not explicitly available on social media platforms. Therefore, as detailed in the following, we used machine learning -based methods to infer such attributes in the PolitiFact dataset users. However, as the required metadata and hashtags are not available for the PAN 2020 dataset, we were not able to compute demographics for this dataset.

Age and Gender. We utilized m3inference [196], a deep-learning-based system trained on Twitter data, to infer user demographic characteristics. Based on the available metadata such as username, screen name, description, and profile image, it predicts the *gender* of the user as male or female, *age* of the user grouped in four categories (≤ 18 , 19–29, 30–39 and ≥ 40) and whether the given account is handled by an *organization* or not. We utilized only two characteristics (age and gender) for both types of users for our analysis. The m3inference has been shown to have an F1 score of 0.918 for gender prediction and 0.522 for age prediction [196].

Political Orientation. As the political ideology can provide additional cues about profiling fake news spreaders, we computed a polarization score to identify their political leaning. We used the method defined by Hemphill et al. [197] where a polarization score (*#polar score*) for each user is defined by using the hashtags from the user tweets to estimate their political ideology. Each of those hashtags is scored according to how political figures with known party affiliation use them. Specifically, we implemented the *#polar score* as follows. As a political figure dataset, we used the dataset provided by Chamberlain et al. [198] which contains tweets collected from Jan 04, 2007, to Jan 03, 2019, and authored by in-office U.S. Congress members during that time period [198]. Then, we classified each politician as Republican or Democrat by using

a TF-IDF vector representation of their tweet hashtags as input features to a binary classifier. We experimented with different classifiers, including support vector machine, logistic regression, extra tree classifier, and random forest with 5-fold stratified cross-validation using class weighting to deal with class imbalance. Random Forest resulted in being the best classifier with 0.69 AUROC and 0.67 average precision, providing us with good confidence in using those hashtags to estimate user political orientation.

Then, we generated Chi-Squared scores for each hashtag, and we leveraged these scores as a polar dictionary to assign polarization scores to the users in the dataset we considered (i.e., PolitiFact). Each hashtag in the tweet is looked up in the polar dictionary, and Chi-Squared scores of matching hashtags are averaged across the entire hashtags included in user tweets defined as polarization score for that user. A positive polarization score indicates that the user tends to incline towards right-leaning political orientation, and a negative score indicates left-leaning political orientation.

7.4.2 Behavioral-based features

This group of features measures the tweeting/sharing behavior and engagement of the users and consists of the following features:

Insomnia index. We analyzed the user tweeting behavior within the day (24 hours).

We divided the time into day and night and considered the ‘night’ window as ‘9PM-6AM’ and the ‘day’ window as ‘6:01AM-8:59PM’ (we used the local time of the user), and analyzed the normalized difference between the number of tweets shared during these time windows for each user as in [199, 200].

Weekend index. Similarly to the insomnia index, we computed the normalized difference in the number of tweets on weekdays and weekends.

Time elapsed. Average time elapsed between two consecutive tweets of the user.

Account duration. The duration (in the number of days) of the account since it is registered.

7.4.3 Network-based features

Vosoughi et al. [64] have shown that fake news spreaders had fewer followers and followed fewer people than real news spreaders. Thus, in this article, we computed the Twitter follower to following (TFF) ratio as in [66] to measure user connectivity in the Twitter social network. TFF is computed by using the following formula

$$TFF = \frac{\#Follower + 1}{\#Following + 1}$$

which indicates the ratio of the number of followers to the number of followings of the user. The greater the ratio, the higher the popularity of the user.

7.4.4 Emotions

Fake news is deliberately induced with emotionally charged words to influence public opinion and affects the vulnerabilities of people by triggering their sentiments such as anger, fear, and distrust towards the event, person, and organization. Moreover, Ghanem et al. [9] recently showed emotions play a key role in detecting fake news. Therefore, we computed emotion features such as anger, joy, sadness, fear, disgust, anticipation, surprise, and trust by using the Emotion Intensity Lexicon (NRC-EIL) [178] and happy, sad, angry, don't care, inspired, afraid, amused, and an-

noyed using Emolex⁵. We started by cleaning tweets by expanding contraction words, correcting misspellings and grammatical mistakes using LanguageTool⁶, replacing negated words with their WordNet antonym, removing stop words, and lemmatizing the words. Next, we computed feature vectors using the approaches proposed by Milton et al. [179, 201]. Specifically, each word is looked up in both emotion dictionaries, and the associated affect values of matching words are extracted. Next, we normalized the scores of each emotion category by the total number of emotions retrieved from a tweet to generate an emotion vector. In case the same emotion was present in both lexicons, e.g., sad in NRC-EIL and sadness in Emolex, we considered the average of the two computed values.

Stress. Along with these emotions (i.e., positive and negative emotions), frustrations, worries, and irritations, which are the characteristics of stress expressed through the language used in the user feed, can also progressively accelerate the spread of fake news. Thus, we incorporate a stress feature computed using the lexical dictionary, a Stress Word Count Dictionary created by Wang et al. [202] as the LIWC tool lacks this category. To compute this feature, we concatenated all the tweets by each user to form a single document per user. We removed words like 'RT,' 'Via,' and '&' for each document.

7.4.5 Personality

The IBM Watson Personality Insights service uses linguistic analytics to infer individuals' intrinsic personality characteristics, including Big Five personality traits, Needs, and Values, from digital communications such as social media posts. The

⁵<https://sites.google.com/site/emolexdata/>

⁶<https://pypi.org/project/language-tool-python/>

tool can work for different languages, including English and Spanish. In our case, we concatenated all the user tweets in a unique document to compute their personality characteristics.

The features computed by this service are detailed in the following (we considered the raw scores provided by the service):

Big Five. The Big Five personality traits, also known as the five-factor model (FFM) and the OCEAN model, are a widely used taxonomy to describe people's personality traits [203]. This taxonomy's five basic personality dimensions are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. For each personality dimension, IBM Watson Personality Insights also provides a set of additional six facet features. For instance, agreeableness' facets include altruism, cooperation, modesty, morality, sympathy, and trust.

Needs. These features describe a user's needs as inferred by the text they wrote and include excitement, harmony, curiosity, ideal, closeness, self-expression, liberty, love, practicality, stability, challenge, and structure.

Values. These features describe the motivating factors that influence a person's decision-making. They include self-transcendence, conservation, hedonism, self-enhancement, and openness to change.

These features ranges from 0 to 1. In terms of how precise is the IBM Watson Personality Insights service, the official documentation ⁷ reports an average Mean Absolute Error (MAE) for the English language of 0.12 for the Big Five dimensions, 0.12 for the Big Five facets, 0.11 for Needs, and 0.11 for Values. The reported average

⁷<https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-science#researchPrecise>

MAE scores are based on a dataset containing user Twitter feeds from between 1500 and 2000 participants for all characteristics and languages.

7.4.6 Readability

Readability measures the complexity of the text, and when computed from text written by the user (tweets in our case), it also represents which level of text complexity a user can understand. To determine that, we used popular readability measures in our analysis:

Flesh Reading Ease

Flesh Kincaid Grade Level

Coleman Liau Index

Gunning Fog Index

Simple Measure of Gobbledygook Index (SMOG)

Automatic Readability Index (ARI)

Lycee International Xavier Index (LIX)

Dale-chall Score

Flesch scores range from 0 to 100. Higher scores of Flesch reading-ease indicate that the text is easier to read, and lower scores indicate difficulty to read. Coleman Liau Index depends on the characters of the word to measure the understandability of the text. The Gunning Fog Index (that generates grade level between 0 and 20), Automatic Readability Index, SMOG Index, Flesh Kincaid Grade Level are algorithmic heuristics used for estimating readability, that is, how many years of

education is needed to understand the text. Finally, Dale-Chall’s readability test uses a list of words well-known by the fourth-grade students (easily readable words) to determine the difficulty of the text. We use this group of 8 readability features to measure the complexity of a user’s writing style.

7.4.7 Writing Style

This set of features captures the writing style of the tweets authored by the same user. Specifically, we computed the average number of certain words, items, and characters per user tweet, which includes the average number of (1) words, (2) characters, (3) lowercase words, (4) uppercase words, (5) lowercase characters, (6) uppercase characters, (7) stop words, (8) punctuation symbols, (9) hashtags, (10) URLs, (11) mentions, and (12) emojis and smileys. Also, we considered the (13) percentage of user tweets that are a retweet and (14) the percentage of user tweets that are a sharing of breaking news; we considered a tweet sharing breaking news if the keyword ‘breaking’ or ‘breaking news’ was appearing in the tweet text. All but features (13) and (14) are computed by removing retweets from the user feed.

7.5 User Characterization

This section presents the main patterns characterizing users who spread fake news that we found by analyzing the features described in the previous section on the two considered datasets. However, as the PAN 2020 dataset only provides 100 tweets per user, and user profile meta-data and timestamps are not included, and hashtags are blurred, we were not able to compute demographic, behavioral, and network features for this dataset. All the feature differences discussed in this section are statistically significant with a p -value < 0.05 (ANOVA or Wilcoxon rank-sum according to the

data distribution).

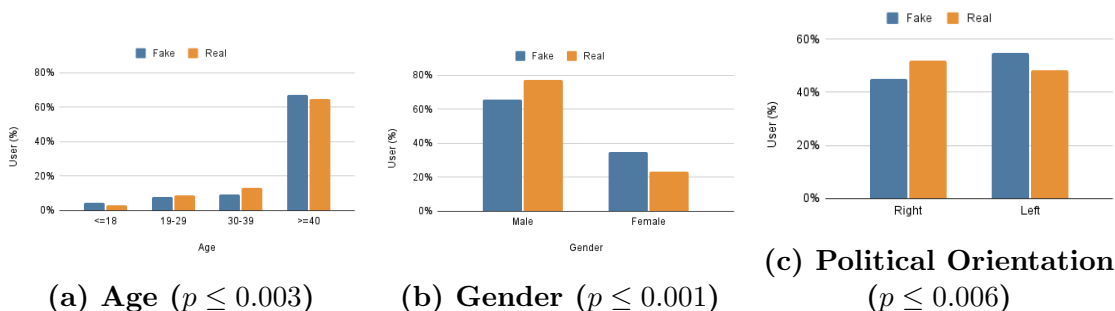


Figure 7.1: Distribution of user demographics on the PolitiFact dataset.

7.5.1 Demographics

Demographics have been shown to be predictors of fake news spreaders [67]. Figure 7.1 shows the distribution of age, gender, and political orientation on the PolitiFact dataset. Here, we observe that among users who have been predicted to be under 18 or over 40, the majority of them tend to share more fake news than real one. The trend is the opposite for users whose age is predicted to be in the age range of 19-39. While previous work has shown that people over 65 tend to share more fake news than the younger generations (age range 18-64), the sharing behavior of users under 18 has not been investigated. Here we observe that these users, together with the ones over 40, may be the most vulnerable to fake news, which is somehow aligned with previous findings. The majority of teenagers are, in fact, unable to assess the credibility of the information that floods their devices [204, 205], while seniors are not as adept as the younger generation in assessing online news veracity [206]. Regarding the role of gender in user sharing behavior, we observe in Figure 7.1b that users whose gender is predicted to be female tend to be more fake news spreaders than male users in the considered dataset. One possible explanation could be that

female users may be less interested in political news and, consequently, less informed and then more vulnerable on these topics [207, 208]. Even if the presented findings for age and gender seem to be somehow aligned with previous research, it is worth noting that these user attributes have been automatically inferred by using a tool whose accuracy is not perfect; hence some errors may have been introduced. Also, the age groups are highly unbalanced, and the last group (≤ 40) is very broad and diverse compared to the other ones. Hence our findings may not be general but just limited to the (not very large) considered dataset.

Figure 7.1c shows the distribution of fake and real news spreaders according to their political orientation. We see that, in the PolitiFact dataset, left-leaning users are more likely to be fake news spreaders than right-leaning users. Guess et al. [67] have shown that, in 2016, conservatives were more likely to share articles from pro-Trump fake news domains than liberals or moderates because those news items were aligned with their beliefs, and the majority of fake news items that were circulating were right-leaning. What we observe in the PolitiFact dataset is not in contradiction with this finding. To show that, we gathered the source bias of the news items present in this dataset from the MediaBias/FactCheck website⁸ and found that the majority of these news items came from left-leaning sources and were tweeted much more than news coming from right-leaning sources (9,435 tweets about news from left-leaning sources vs. 3,408 tweets about news from right-leaning sources). Thus, we also observe in the PolitiFact dataset that the political orientation of a fake news spreader

⁸mediabiasfactcheck.com. The website's main goal is to educate the public on media bias and deceptive news practices. This website contains a comprehensive list of news sources, their bias, and their credibility of factual reporting scores. Here, the publisher's political bias is defined by using seven degrees of bias: *extreme-right*, *right*, *right-centered*, *neutral*, *left-centered*, *left*, and *extreme-left*.

is more likely to coincide with the one of the sources of the majority of circulating fake news items (left-leaning in this case).

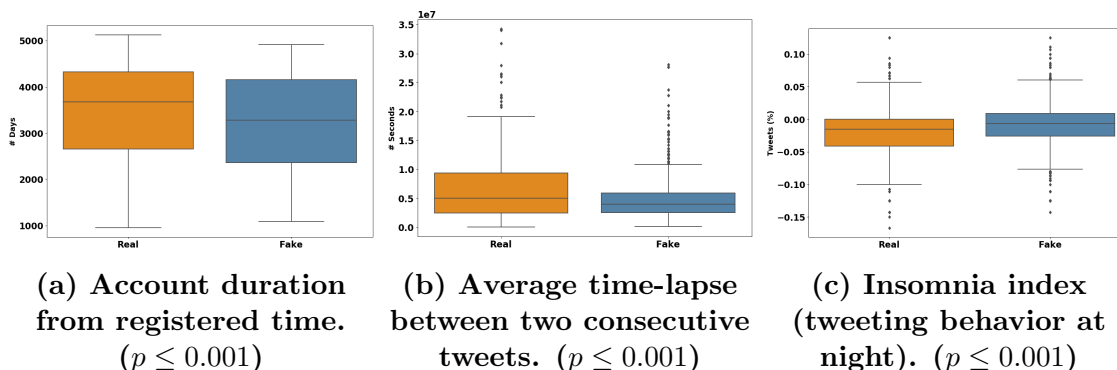


Figure 7.2: Boxplots of user behavioral features on the PolitiFact dataset.

7.5.2 User Behavior

The presence of timestamps in the PolitiFact dataset allows us to investigate fake news spreaders tweeting behavior. Figure 7.2 shows the box plots of the considered behavioral features on such dataset. Here, we observe that fake news spreaders (1) have newer accounts, (2) spend, on average, less time between two consecutive tweets, and (3) tend to tweet more at night (higher insomnia index) than real news spreaders.

Thus, fake news spreaders are users who are newer to the platform (we are not considering bot accounts) and may be less expert about its functionalities/usage, and who tend to tweet more frequently, perhaps to increase their social capital. Also, their higher nighttime online activity may be connected with the presence of a higher stress condition for fake news spreaders, as shown in section 7.5.4.

7.5.3 User Network

Figure 7.3 shows the distribution of the average Twitter follower to following (TFF) ratio on the PolitiFact dataset. We observe how non-fake news spreaders are much

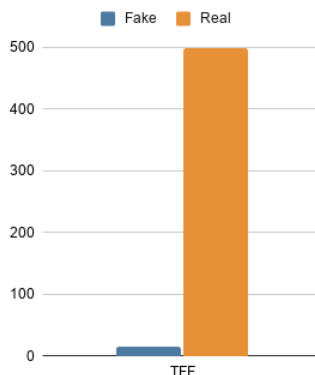


Figure 7.3: Average Twitter follower to following (TFF) ratio on the PoliFact dataset. The difference is statistically significant ($p \leq 0.001$)

more popular (they have around 500 times more followers than following, on average) than fake news spreaders. Thus, users with lower TFF may tend to spread fake news more to increase their popularity on Twitter. For instance, users may know a news item is fake and spread it anyway because it is funny or of interest to user’s friends and hence generate engagement among Twitter followers. Another motivation could be that a user with a low TFF is new to the platform and is not familiar with its features, hence may mistakenly share fake news.

7.5.4 User Emotions

Figure 7.4 shows the radar charts of user emotions while Figure 7.5 shows a comparison of user stress levels on both the considered datasets. We notice that, in both cases, fake news spreaders tend to express more negative emotions (fear, sadness, disgust, and angry) and stress in their tweets than real news spreaders (all p -values are ≤ 0.001). Conversely, non-fake news spreaders are happier and more inspired, but also more afraid (all p -values are ≤ 0.001). Being induced by negative bias, people generally pay more attention to negative news [180, 194]. Hence fake news spreaders

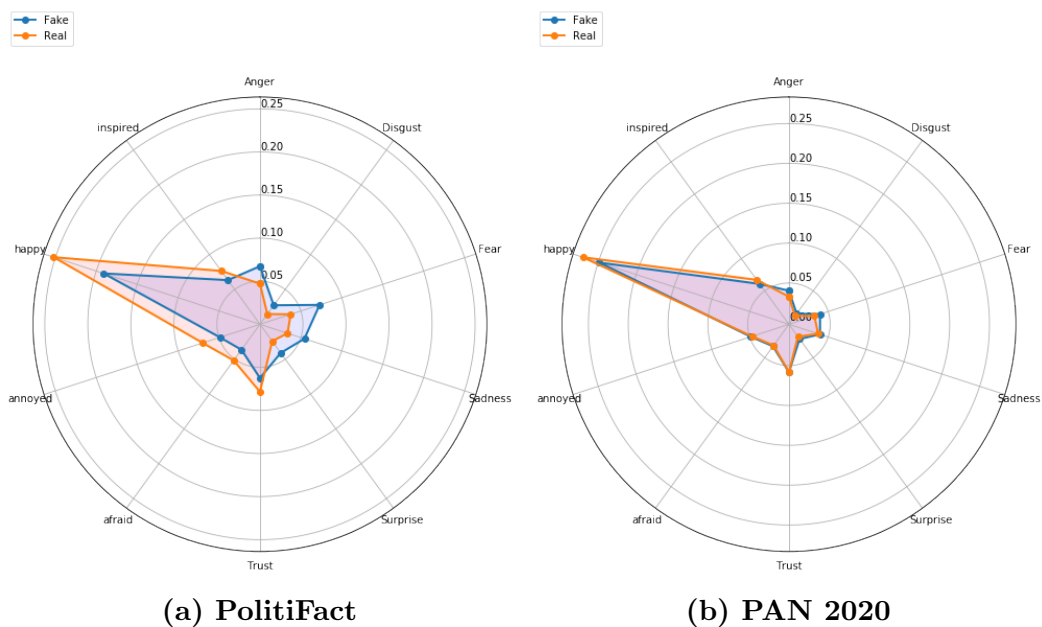


Figure 7.4: Radar charts of the emotion features: PolitiFact and PAN 2020 datasets.

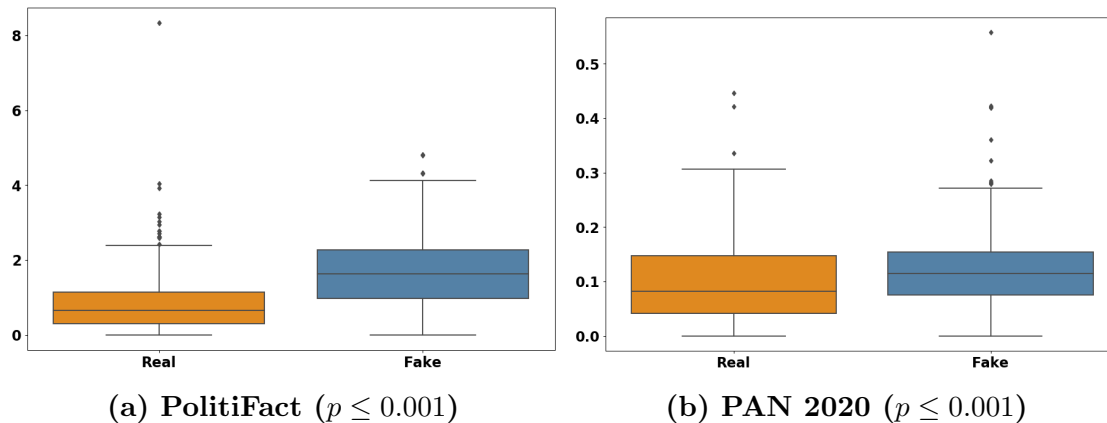


Figure 7.5: Box plots of user stress level on the PolitiFact and PAN 2020 datasets.

tend to frame their tweets with negative emotions targeting to make it catchier and circulate more among people. On the contrary, non-fake news spreaders are general individuals whose motive of using social media platforms is to connect with other

people or family, share their achievements, advice, and support [209] and are more skeptical about sharing fake news.

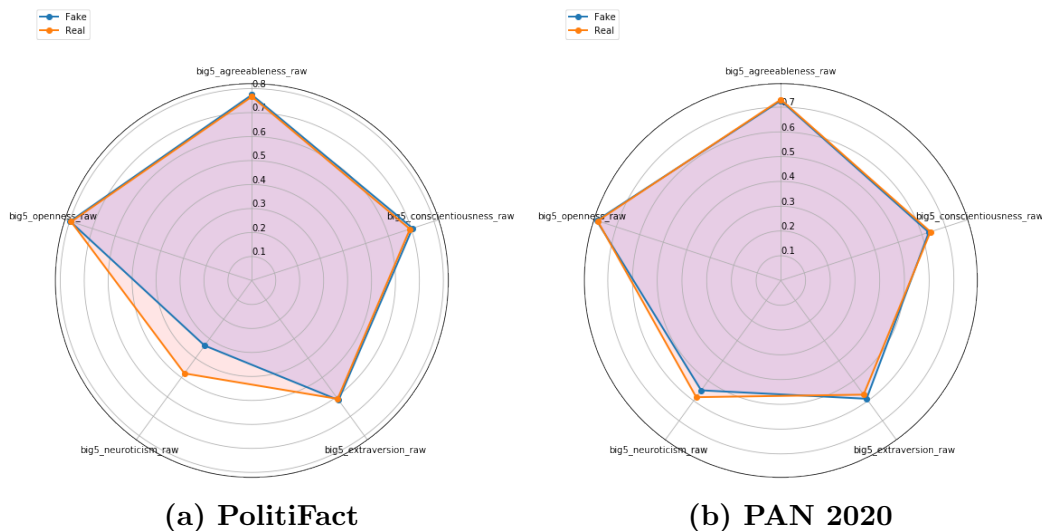


Figure 7.6: Radar charts of the Big-Five personality scores: PolitiFact and PAN 2020 datasets.

7.5.5 User Personality Traits

User Big Five personality traits are shown in Figure 7.6 for both types of users. Among the five traits, extroversion and neuroticism are statistically significant features in both datasets (all p -values are ≤ 0.001) and show the same trend, namely, fake news spreaders are estimated to be more extroverted and less neurotic than real news spreaders. Extroversion is related to the number of friends a user has, while neuroticism is related to frequency of posting [210]. Thus, fake news spreaders are estimated to be people who may share fake news to capture the interest of and make fun with their friends and/or possibly connect with more people. On the other end, sharing fake news is a rarer phenomenon as compared to real news sharing [67]; hence fake news spreaders are estimated to be less neurotic because they share less than

real news spreaders.

The other three personality traits are statistically significant features only in the PolitiFact dataset (all p -values are ≤ 0.001), where we found that fake news spreaders are estimated to be more agreeable, conscious, and open than real news spreaders. Agreeableness is related to the type of feelings (positive or negative) expressed via social media updates, conscientiousness to posting about political news, and openness to the sharing of various forms of media [210]. Thus, fake news spreaders are estimated to be people whose posting behavior is driven by emotions (either positive or negative) and have more interest in political events.

7.5.6 User Readability Level and Writing Style

Different from emotional and personality features, readability features do not generalize across the considered datasets. In general, fake news spreaders in the PolitiFact datasets have a lower readability level than non-fake news spreaders, while the trend is the opposite in the PAN 2020 dataset. Figures 7.7 and 7.8 show the box plots of two of the readability measures we considered on the PolitiFact and PAN 2020 dataset, respectively.

Similarly, Table 7.2 highlights the pattern of writing style among fake news spreaders and real news spreaders. If the value of a feature was higher (on average) for real news spreaders as compared to fake news spreaders, it is denoted as $R > F$ (and $R < F$ otherwise) in the table. Fake news spreaders tend to use more uppercase characters and fewer hashtags in their tweets but share more breaking news than real news spreaders, and this trend generalizes for both datasets. Moreover, fake news spreaders in PolitiFact incorporates more uppercase words and URLs but fewer words, lowercased characters, punctuation, trailing periods ('...'), stop words, and

Features	PolitiFact	PAN 2020
Hashtags	$R > F$	$R > F$
Retweets		$R > F$
Char	$R > F$	$R < F$
Uppercase char	$R < F$	$R < F$
Lowercase char	$R > F$	$R < F$
Lowercase word	$R > F$	
Uppercase word	$R < F$	
Breaking	$R < F$	$R < F$
Emoji		$R > F$
Trailing Period	$R > F$	
Punctuation	$R > F$	
Word Count	$R > F$	
Stop words	$R > F$	
URLs	$R < F$	
Mentions	$R > F$	

Table 7.2: Writing style features that differ in user feed. All differences are statistically significant ($p \leq 0.002$ for PolitiFact and $p \leq 0.04$ for PAN 2020). Shaded cells indicate the same pattern in both datasets.

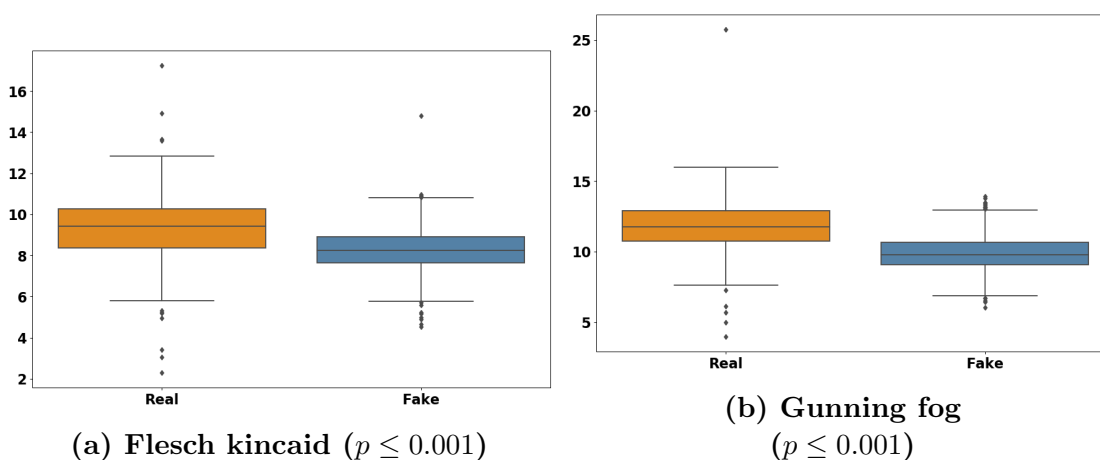


Figure 7.7: Readability index of tweets written by fake news spreaders vs real news spreaders in PolitiFact.

mentions than real news spreaders in their tweets. In the PAN 2020 dataset, fake news spreaders use more lowercase characters, fewer emojis, and retweet less than

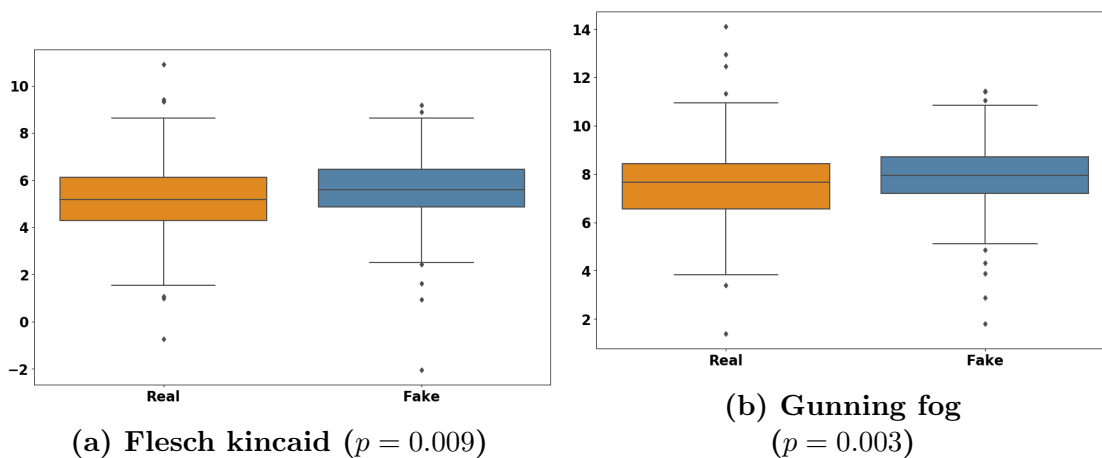


Figure 7.8: Readability index of tweets written by fake news spreaders vs real news spreaders in PAN 2020.

real news spreaders. This indicates that fake news spreaders aim to gain people’s attention by sharing breaking news and using more uppercase words and URLs in their tweets.

7.6 Experiments

This section reports on our experimental results of using the features described in Section 7.4 to automatically identifying fake news spreaders.

7.6.1 Experimental Setting

We addressed the problem of automatically identifying fake news spreaders as a binary classification task. In particular, we used the combination of all the groups of features for the prediction. Once the features are computed, the classification is performed by using the best classifier among linear support vector machine (SVM), logistic regression, and random forest. We used class weighting to deal with the class imbalance and performed 5-fold cross-validation. Additionally, we also used each group of features as input to the best classifier to examine the contribution of these

features in identifying a user likely to spread fake news. As an evaluation metric, we used Average Precision⁹ (AvgP) which is a metric commonly used when dealing with unbalanced binary datasets [211], as in the case of the Politifact dataset. The average precision is the area under the precision curve, computed by plotting precision against the true positive rate. The average precision score gives the probability that a classifier will correctly identify a randomly selected positive sample (e.g., a fake news spreader in our case) as being positive. In our problem, we are interested in identifying fake news spreaders with high precision. These are the users to target with correction strategies to mitigate the further spread of fake news. In the tables reported in this section, the best average precision values are highlighted in bold.

Baselines for comparison. We compared our proposed approach with the two best performing approaches used by the participants to the PAN CLEF 2020 shared task, namely the approaches proposed by Buda and Bolonyai [212] and Pizarro [213]. These baselines are described here below:

Buda and Bolonyai [212] utilized n-grams based approach and combined them with statistical features from the tweets, such as their average length or their lexical diversity. Specifically, they used an ensemble model of Logistic Regression with five sub-models, namely, logistic regression, linear SVM, random forest, and XGBoost with n-grams and XGBoost with statistical features.

Pizarro [213] utilized a character and word n-grams-based approach with a linear support vector machine as the classifier.

⁹We used the average precision implementation provided by the Python Scikit-learn library: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

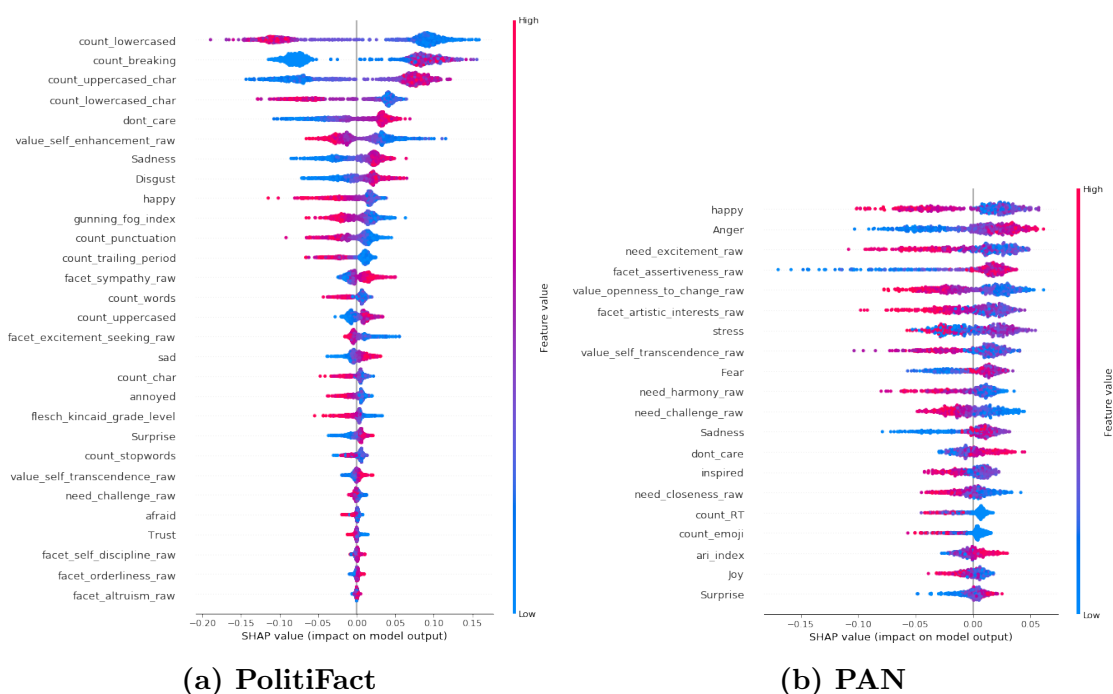


Figure 7.9: SHAP summaries of the important features: PolitiFact and PAN datasets. Y-axis represents the features in order of importance. X-axis represents the shap values, positive values (greater than zero) represents a higher chance of classifying a user as a fake news spreader and negative values represent a higher chance of classifying a user as a real news spreader.

Approach	PolitiFact	PAN 2020
Buda and Bolonyai [212]	0.737	0.783
Pizarro [213]	0.966	0.714
Our Features (Random Forest)	0.995	0.795
Our Features (Linear SVM)	0.595	0.687
Our Features (Logistic Regression)	0.672	0.717

Table 7.3: Average precision of our proposed features (in input to a Random Forest classifier) on PolitiFact and PAN 2020 datasets and comparison with baselines. Best values are in bold.

7.7 Classification Results

Classification results are reported in Table 7.3 to allow comparison between the performances of baselines and our method on both PolitiFact and PAN 2020 datasets. As we can see, our proposed features consistently outperform both baseline approaches. Specifically, we got an average precision of 0.995 vs. the best baseline results of 0.966 achieved by Pizarro [213] on the PolitiFact dataset and an average precision of 0.795 vs. the best baseline results of 0.783 achieved by Buda and Bolonyai [212] on the PAN 2020 dataset. Among the considered classifiers, random forest achieved the best performance. Furthermore, the baseline methods are mainly n-grams-based and, consequently, they are not easy to interpret. On the contrary, the features we consider in Section 7.4 achieve better performances and can also be analyzed to provide significant patterns to characterize fake news spreaders as we have shown in Section 7.5.

In addition, we investigated the performance of each considered group of features individually (demographics, emotions, behavior, network, readability, personality, and writing style) when the best classifier (i.e., random forest) is used. Results are re-

Features	PolitiFact	PAN 2020
Demographics	0.777	-
Emotions	0.976	0.787
Behavior	0.866	-
Network	0.776	-
Readability	0.897	0.635
Personality	0.979	0.786
Writing Style	0.990	0.713

Table 7.4: Average precision per feature group on PolitiFact and PAN 2020 datasets.

ported in Table 7.4. We observe that emotions and personality features are the most important groups of features for the PAN 2020 dataset. In the PolitiFact dataset, the writing style is the most important group of features, while emotions and personality are the second most important groups of features. Hence, our results reveal that emotions and personality are strong predictors of fake news spreaders in both datasets. Since the Twitter IDs of the users in the PAN 2020 dataset are concealed, it was not possible for us to collect the additional user data required to generate some features like demographics, behavior, and network features. However, the features extracted from the text show, in general, better performances than demographics, behavior, and network features in both datasets, as shown in Table 7.4. Combining all the groups of features together further improves the average precision of the classification task (cf. Table 7.3).

7.8 Feature Importance and Shapley Additive Explanations

Considering all the features from each group, we have a total of 91 and 99 features for the PAN and PolitiFact datasets, respectively, which can still be too many for

Features	PolitiFact	PAN 2020
Important features	0.994	0.776
All features	0.995	0.795

Table 7.5: Average precision of important features from Figure 7.9 vs. all features on PolitiFact and PAN 2020 datasets.

the size of the considered datasets (PolitiFact and PAN) to perform real vs. fake news spreader classification. Therefore, we used the statistical tests (Analysis of Variance (ANOVA) and Wilcoxon rank-sum depending on the data distribution) as in [194, 1] to perform feature selection. For each dataset, only features where the two averages (real vs. fake news spreader) were significantly different according to the statistical test (p -value < 0.05) were considered. Also, features are sorted by F-value in descending order to determine the importance. Among these features, we selected the top- k most important features to feed the classification algorithm, where k is the square root of the training set size (rule of thumb). For each dataset, the selected important features are shown in Figure 7.9.

Table 7.5 shows our classification results with important features using the best classifier, i.e., random forest. We observe that using only important features lowers the performance by a very small margin 0.19% and 0.1% in the PAN 2020 dataset and PolitiFact datasets, respectively. However, it still outperforms the scores of both baselines shown in Table 7.3.

Further, to explain why users are classified as fake news spreaders or real news spreaders, we used the SHAP values (SHapley Additive exPlanations) of the selected features, a widely used approach inspired by cooperative game theory [214]. We leveraged a tree explainer which is basically used to compute SHAP values for tree-based models. Since we want to learn about how each feature is influencing the decision of

the model, we used the global importance, i.e., the sum of all the absolute Shapley values per feature across the dataset. Figure 7.9 shows the SHAP summary plot that demonstrates the contribution of each feature in predicting users likely to spread fake news. The higher the SHAP value (i.e., closer to 1.0), the higher the probability of being a fake news spreader. As shown in the figure, writing style features (like frequency of lowercase words, uppercased characters) appear as the most important features in the model for the Politifact dataset. We observe that users writing tweets with fewer lowercased words, more uppercased characters, more breaking, less punctuation, shorter text, and fewer stopwords are more likely to be fake news spreaders according to the PolitiFact dataset. On the other end, features indicating emotions like happiness and anger and personality facets such as excitement, assertiveness, openness to change, artistic interests appear as the most important features in the model for the PAN 2020 dataset. We see that the users with less concern about others' welfare and interests (self-transcendence), less concordance (harmony), and having the willingness to change (openness to change) are more likely to be fake news spreaders, according to the PAN 2020 dataset.

Additionally, we further confirm that negative emotions like anger, fear, disgust, stress, and sadness extracted from the tweets of a user are among the most important features and indicate that the users likely to spread fake news seem to embrace a language with more negative valiance than real news spreaders in both datasets. TO do: add parts from misinfo workshop paper

7.9 Conclusions

In this article, we performed a comprehensive analysis to understand the correlation between user characteristics based on different attributes such as user demograph-

ics, personality, emotion, writing style and readability, social media behavior, and the likelihood of a user being a fake news spreader. We considered two datasets to perform our analysis, namely the PolitiFact (FakeNewsNet) and PAN datasets, and investigated new features such as user tweeting behavior and stress level. Furthermore, we addressed the problem of identifying users likely to share fake news using the proposed groups of features in both datasets and compared the performance with baseline approaches from the PAN shared task. Specifically, we obtained an average precision of 0.99 on the PolitiFact dataset (vs. 0.96 achieved by the best baseline) and 0.80 on the PAN dataset (vs. 0.78 achieved by the best baseline).

Our results showed the potential of the proposed features in identifying fake news spreaders by outperforming baseline approaches in both considered datasets. Our findings showed that younger generation under 18 or users over 40 may be more vulnerable in case of fake news sharing, and females may be more likely to be fake news spreaders than male users. Similarly, fake news spreaders tend to express more negative emotion and stress in their tweets, and the political orientation of a fake news spreader is more likely to coincide with the bias of the source of the majority of circulating fake news items. Besides, the behavioral patterns show that fake news spreaders have newer accounts, spends less time but tweet more within a short time interval. Likewise, it shows the inferred user personality, writing styles, and readability of the user's tweets have the potential to identify whether the user is a fake news spreader effectively.

Using an automated tool to infer user demographics based on their screen name, description, and profile image could be a potential limitation of our study. Thus, inferred demographics of some of the users might not be entirely accurate. However,

it is impossible to test the tool's efficiency in the considered datasets as such metadata are not explicitly available to be used as ground truth. Labels in the PAN 2020 dataset are another limitation of the work presented in this article as a user is labeled as a fake news spreader if they have shared at least one fake news item in the past. We have proposed a way to compute more reliable labels for the Politifact dataset to overcome this limitation. Finally, we have considered only users keen to spread fake political news, and we leave as future work the study of fake news spreaders in other domains, e.g., gossip news.

CHAPTER 8:

JOINT CREDIBILITY ESTIMATION OF PUBLISHER, NEWS, AND USER VIA HETEROGENEOUS GRAPH REPRESENTATION LEARNING.

8.1 Introduction

In recent years, the way of news consumption by readers has gone through a massive transformation. The large global population relies on social media for daily news. In 2021, the Pew Research Center reported that more than half (69%) American adults consume news from Twitter, and among them, 70% say they rely on Twitter for breaking news [215]. Moreover, the easy access to the news and facility of social media that allows users to build connections with other users has made social media a diffusion ground for fake news eventually escalating its detrimental effects on public trust and society. Significant effort has been made to detect fake news. One of the most intuitive approaches that anyone looks for while validating news is fact-checking. However, fact-checking is expert-driven and labor-intensive leading to the high possibility of fake news items going viral over social media before fact-checking websites flag those news items. Several studies were also focused on capitalizing

linguistic and psychological attributes [32, 1, 31], writing styles [216, 1, 217]. Others leveraged the social context attributes [218] and hybrid attributes considering both linguistic and social context [30]. Alternatively, some used user-based features to characterize and identify users who tend to share fake news on social media [69].

Although significant effort has been made to develop methods to minimize the effects and spread of fake news, it is undeniable that the prevalence of fake news over social media continues to persist and has become a more serious problem than ever before. The majority of existing works focus on the detection of fake news or fake news spreaders separately undermining the fact that there exists a relation between news publishers, news items, and users in social media which can impose complementary information about the credibility of each entity of news ecosystem. Moreover, the existing works only address fake vs real news detection or news spreaders detection. Whereas in reality, users may contribute to the spread of fake news intentionally or unintentionally. In addition, the credibility of the news items may depend on the amount of false information present in the news contents and the credibility of the news publishers may depend on mostly which types of news items they publish.

For instance, let us consider a sample news ecosystem in social media as illustrated in Figure 8.1. Here, p_1 and p_2 are news publishers who publish news items n_1 , n_2 , n_3 and n_4 and u_1 , u_2 , u_3 and u_4 are users who share those news items in social media. There is a high possibility that hyper-partisan news publishers (i.e., extreme left bias or extreme right bias) can falsely modify the information of the events in their news items (fake news) to support their corresponding partisan. For instance, news publisher p_2 with extreme partisan bias (either left or right) publishes fake news item n_4 while least partisan bias publisher p_1 tend to publish real news n_1 and n_2 .

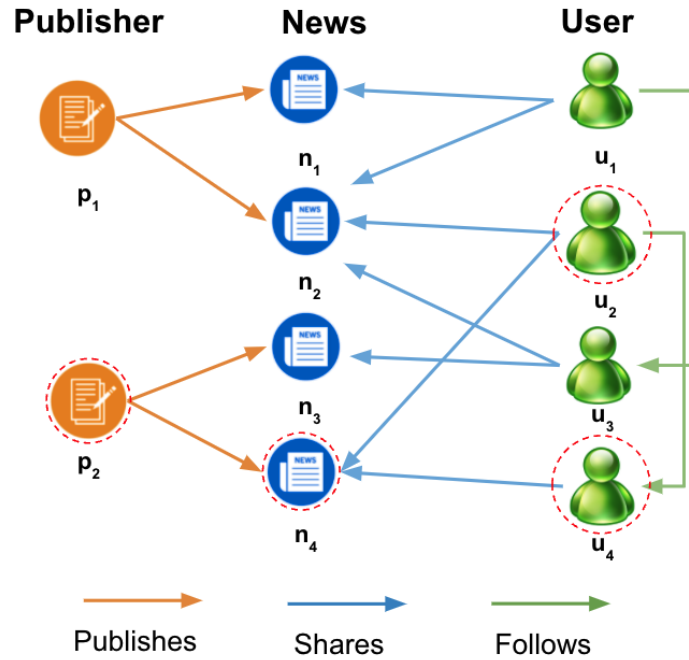


Figure 8.1: Sample news ecosystem in social media.

This assumption is based on the correlation between the political bias of a publisher and the veracity of its published news contents that have been theorized by studies in the field of journalism [114, 115] and also validated by our previous analysis [200]. However, publisher (e.g. p_2) can publish both real news (n_3) and fake news (n_4) and can have mixed credibility. Similarly, users in social media engage with news items and further share them among their connections. There is a high possibility that fake news spreaders like user u_2 and u_4 tend to share fake news item n_4 while users with no malicious intention like u_1 and u_3 tend to share real news items n_1 and n_2 . However, as people are poor at identifying fake news in social media, which is also supported by the study [189], people can share both real news items and fake news items. Therefore, users in the news ecosystem can be fake news spreaders, real news spreaders, or mixed news spreaders. Thus, it is evident that binary classification, i.e.

identifying publisher as credible or non-credible, user as fake news spreader or real news spreader is no longer sufficient.

Users in social media build connection where users with similar interest follows each other like user u_1 follow user u_3 and user u_2 follow user u_4 . Therefore, publisher-news relation, news-user relation, and user-user relation can provide complementary cues on the credibility of each entity in the news ecosystem.

In this chapter, we propose an approach that leverages the information about each entity of the graph (e.g. publishers, news items, and users in the news ecosystem) along with their relational information to detect their credibility jointly. Unlike existing research that includes only a classic binary condition, we model the problem of detecting the credibility degree of each entity as a multi-class classification problem. For instance, in the above-mentioned example, the credibility of the news publisher may range from very high to very low, news items may belong to mostly true, mostly false, etc and a user may frequently spread fake news items or both. For this, we used the characteristics of news items and users that we learned in our previous works [219, 220] as their corresponding information, and for incorporating relation we used the graph-based method.

In particular, our approach portrays the news ecosystem as a heterogeneous graph as shown in Figure 8.1 and models the interaction between entities to extract their representations using Relational Graph Convolution Network (RGCN) [221]. Precisely, we modified the RGCN model that can work with feature vectors of different dimensions and optimized on combined node-specific losses, as explained in Section 8.3.2. The generated vector representation feature takes into account the entire characteristics of the entity as well as the relationship between each entity of the news ecosystem

which can be used in downstream approaches like identifying the credibility degree of a news publisher, a news item, and a user jointly.

We report the results obtained when applying the proposed approach to the PoliFact dataset. Our results show a drastic improvement in the detection of fake news and fake news spreaders as compared to related approaches and considered baselines. Specifically, we outperformed baselines with a macro F1 score of 0.93 vs. the best baseline results of 0.87 for fake news detection and 0.79 vs. the best baseline result of 0.73 for fake news spreader detection. This indicates the importance of the relational attributes in determining the credibility degrees of entities in the news ecosystem.

8.2 Related Work

In this section, we briefly describe the related works on fake news detection and fake news spreader detection.

Several studies have focused on fake news detection methods by considering news contents features which are often extracted from news excerpts, headlines, and associated visual [31, 1, 32, 85, 100] and social context features including demographics, political orientation and network structure [222, 223]. For instance, Potthast et al. [31] considered the writing style of the news text, Horne and Adali [1] considered linguistic and complexity based features extracted from both news body and headline to determine the validity of news. While other studies [200, 70] utilized images as additional cues incorporated in fake news. In addition, several studies have been conducted to understand and characterize the users that are likely to spread fake news on social networks. Vosoughi et al. [64] revealed that the fake news spreaders had, on average, significantly fewer followers, followed significantly fewer people, and were significantly less active on Twitter. Shrestha and Spezzano showed that

social network properties help in identifying active fake news spreaders [65]. Shu et al. [66] analyzed user profiles to understand the characteristics of users that are likely to trust/distrust fake news. Guess et al. [67] analyzed the user demographics as predictors of fake news sharing on Facebook and found out political-orientation, age, and social media usage to be the most relevant. Shrestha et al. [68] analyzed the linguistic patterns used by a user in their tweets and personality traits as a predictor for identifying users who tend to share fake news on Twitter data [69, 68].

In addition to the approaches using feature engineering, researchers have also been looking into deep learning approaches in order to encode information from news and social context to further harness the detection of fake news and fake news spreaders. SAFE [70] used TextCNN [71] and FakeBERT [72] used BERT to encode textual information of news content and visual. Similarly, graph-based approaches using popular GCN have also been utilized to encode the propagation of news on social media for the detection of fake news[73, 74]. Likewise, Giachanou et al. [75] leveraged GCN to process the user Twitter feed in combination with features representing user personality traits and linguistic patterns used in their tweets to address the problem of discriminating between fake news spreaders and fact-checkers. Shu et. al developed a model dFEND [76] that utilized bidirectional RNN with GRU to capture word and sentence-level representation from news articles and user comments (tweets) to detect fake news. Further, they used BERT to encode textual information from news items and implemented a two-layered multi-layered perceptron (MLP) to predict fake news [77]. However, these methods focus on modeling fake news detection or fake news spreaders detection separately.

In this work, we attempt to jointly predict the credibility of news items, users, and

news publishers in the news ecosystem by leveraging relational graph convolutional network (RGCN) [221].

8.3 Methodology

In this section, we describe our proposed approach, a graph representation learning-based approach to generate a representation for each entity from their node and relational information in a supervised way.

Definition 2 (Heterogeneous Graph). *A heterogeneous graph, graph with multiple node types connected with multiple relation types, is defined as a directed graph $G = (V, E, T_n, T_r)$ where V , E , T_n and T_r represents the set of nodes, set of edges, set of node types and set of edge (relation) types, respectively. Each node $v \in V$ and edge $e \in E$ belong to specific node type in T_n and edge type in T_r mapped by function $\Phi : V \rightarrow T_n$ and $\psi : E \rightarrow T_r$, respectively where the graph contains multiple node type and edge type i.e. $|T_n| + |T_r| > 2$.*

Example 2. *News ecosystem in social media can be modeled as an example of heterogeneous graph. As shown in Figure 8.1, a heterogeneous news ecosystem graph can be defined as $G = (V, E, T_n, T_r)$ where nodes $v_i \in V$ consists of three different node types T_n : **publishers**, **news items** and **users** and $r \in T_r$ represents three different relation types: ‘publisher–**publishes**–news item’, ‘user–**shares**–news item’, ‘user–**follows**–user’, among the nodes connected by the edge $(v_i, r, v_j) \in E$. Note that the inverse relations i.e. ‘news item–**published by**–publisher’, ‘news item–**shared by**–user’, and ‘user–**followed by**–user’, are also included in graph G such that there is bidirectional information flow between the connected nodes.*

8.3.1 Relational Graph Convolutional Networks

The main difference between using a graph-based approach over the traditional machine learning model is that the model learns not only the information about the nodes but also the connection information encoded by the edges between the nodes. One of the most popular graph-based approaches for node classification is graph convolutional networks (GCNs) [224] which were proposed to automatically extract features from graphs.

The key idea behind the GCNs is message passing framework where vector representation of the neighboring nodes N_i in l^{th} hidden layer of the neural network are passed as an incoming message to the node v_i which are aggregated using some aggregation functions like sum, mean, etc. The aggregated message are then linearly transformed using multiplication by a weight matrix W^l which is then passed through an activation function $\sigma(\cdot)$ to produce the new vector representation of the node at $l + 1^{th}$ layer. The generated vector is called node representation as shown in Eq 8.1.

$$h_i^{l+1} = \sigma \left(\sum_{j \in N_i} \frac{1}{c_{ij}} h_j^l W^l \right) \quad (8.1)$$

where $j \in N_i$ indicates the neighboring node of node v_i whose vector representation in l^{th} hidden layer of the neural network is denoted as h_j^l and c_{ij} is a normalization constant for the edge (v_i, v_j) between node v_i and node v_j .

GCN works only with homogeneous graph with only one edge type. Therefore, Relational Graph Convolutional Network (RGCN) [221] was proposed as an extension of GCNs for labeled multi-graphs. Unlike regular GCNs, RGCN encode node based on the aggregation of its neighbors of particular relation type and transforms using

different trainable weight matrices based on relation type (relation-specific transformations) and is defined as,

$$h_i^{l+1} = \sigma \left(\sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} h_j^l W_r^l + h_i^l W_0^l \right) \quad (8.2)$$

where N_i^r is a set of neighbors of node v_i connected by specific edge (relation) type $r \in R$. W_r^l and $c_{i,r}$ are relation specific weight matrix and normalization constant in l^{th} hidden layer respectively. Instead of only aggregating representations of neighboring nodes, RCGN includes representation of node v_i itself from layer l to ensure that the generated representation in layer $l + 1$ includes self information from previous layer.

One limitation of RCGN is that it requires the dimension of feature vectors to be the same for all node types. Intuitively, several real-world scenarios can be represented as heterogeneous graphs and have node-specific feature vectors with completely different dimensions. Thus, we propose an approach, Role-Relational Graph Convolutional Network (Role-RGCN), a modified RCGN that can work with feature vectors of different dimensions and optimized on combined node-specific losses as explained in Section 8.3.2.

8.3.2 Role-Relational Graph Convolutional Network (Role-RGCN)

In this section, we describe our proposed model, Role-RGCN based node representation learning to extract node representation from the heterogeneous graph that can be used in downstream approaches like node classification.

We defined a heterogeneous graph where each node is characterized by a specific role and nodes with different roles have a different kind of features. Figure 8.2 il-

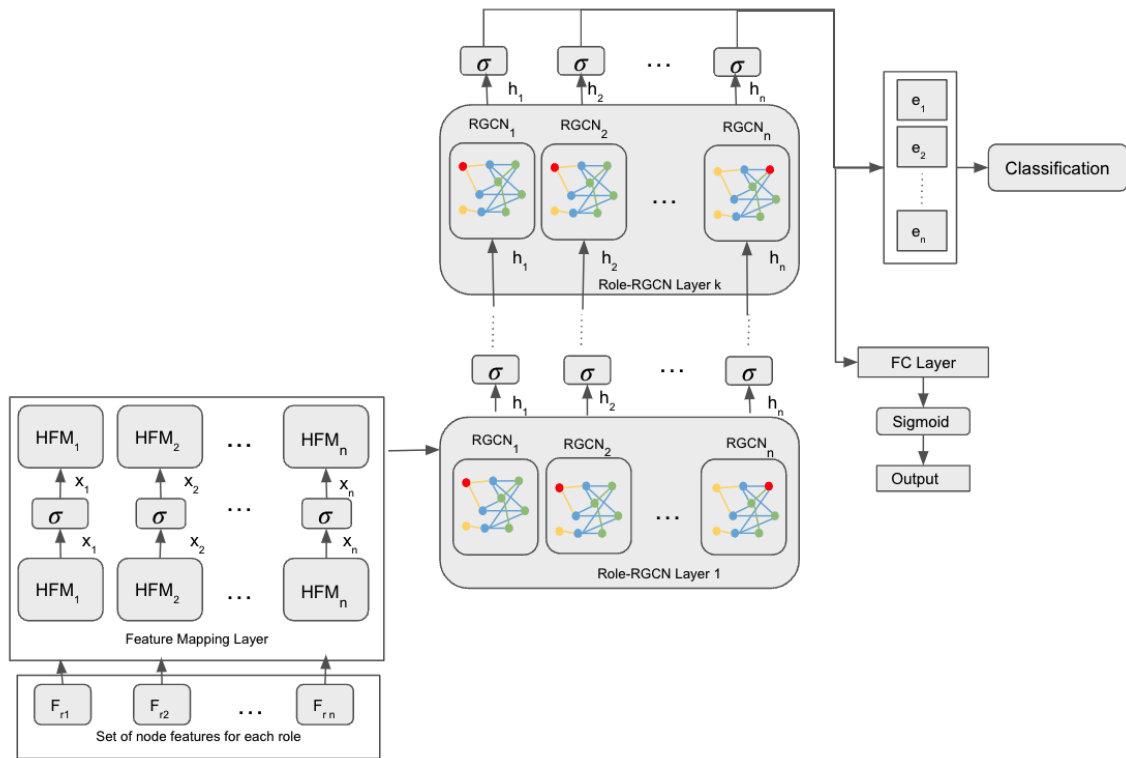


Figure 8.2: Role-RGCN Architecture

illustrates the architecture of our proposed model Role-RGCN. The first layer of the model comprises of “Feature Mapping Layer” where each heterogeneous feature mapping layer (HFM_i) transforms the feature of a node of a specific role in a common k -dimensional vectorial representation space.

Then the generated k -dimensional vector representation is passed through the “Role-RGCN Layer” where each node of a specific role has a dedicated RGCN model. The operation in the Role-RGCN layer is inspired by the fact that nodes with different roles process the information differently. Thus, the representation associated with each role is created through the aggregation process and when a node aggregates the information from its neighboring nodes it should distinguish from which kind

of node the information is coming from. To include the relational attributes, the generated vectorial representation associated with each role is consequently passed through Role-RGCN layers. A single Role-RGCN layer aggregates the information of neighbors from 1 hop and with the increase in the number of Role-RGCN layers, our model aggregates information of neighbors from multiple hops. Note that the representation associated with each role will be determined during the training phase, i.e., they will be trained with the entire network. Moreover, to improve the training convergence time of our model, we scaled all the input feature vectors with a standard scaler. We optimized the learning process using the Adam optimizer and used the global loss function, i.e. combining the cross-entropy loss for each role. The model outputs the representation of node that encapsulates the information about the node and its relation with other nodes.

Finally, the generated representations associated with each role are passed through a classifier for identifying the label of each role. During the classification process, our proposed model identifies which kind of node to consider and selects their corresponding representation as input to the classifier.

8.4 Datasets

This section describes the datasets we used to evaluate our proposed model. The FakeNewsNet dataset [51] consists of two datasets, PolitiFact and GossipCop, from two different domains, i.e., politics and entertainment gossip, respectively. Each of these datasets contains details about news content, publisher, social engagement information, and user in social networks. GossipCop focuses on gossip, which is related to a different form of misinformation. For this reason, this chapter only uses the PolitiFact dataset, which consists of news with known ground truth labels

Entity	# Fake	# Real	# Mixed
News	300	206	
Users	648	398	89

Entity	#Very Low	#Low	#Mixed	#Mostly Factual	#High	#Very High
Publisher	5	16	10	9	26	5

Table 8.1: Datasets and statistics.

collected from the fact-checking website PolitiFact¹ where journalists and domain experts fact-checked the news items as fake or real. Overall, the dataset contains 292 news sources publishing 990 news items with ground truth after removing two articles, `politifact14920` and `politifact14940`, that were labeled both real and fake. Among 990 news items, 701 news items were shared by 295,469 users. Note that the baselines with which we compared our approach used news items with at least 3 comments (tweets) [76]. Thus, we selected only those news items that have at least 3 tweets resulting in 506 news items with ground truth published by 197 publishers. We leveraged trustworthiness score provided by MediaBias/FactCheck (MBFC) [225] website as the ground truth label for publisher which ranges from ‘very high’, ‘high’, ‘mixed’, ‘mostly factual’, ‘low’ and ‘very low’. Among 197 publishers only 71 news publishers’ information was available on the MBFC website.

As this dataset only provides ground truth for news, we computed the labels for the users (fake news spreader, real news spreader, or mixed) as explained here below. First, we filtered out those users who had shared the same news item multiple times, and then we selected only those users who had shared at least eight unique news. Next, the resulting group of 1,135 users is labeled as fake news spreaders, real news

¹<https://www.politifact.com/>

spreaders, or mixed news spreaders as follows: (1) a user is a fake news spreader if at least 60% of the news items they shared are fake, or (2) a user is a real news spreader if at least 60% of the news items they shared are real (3) mixed news spreader otherwise. We labeled 648 users as fake news spreaders, 398 as real news spreaders, and 89 as mixed news spreaders. The size of the dataset is shown in Table 8.1.

8.5 Experiments

This section reports on our experimental evaluation of our proposed method to generate node representation and compare its performance against several approaches that have been used to automatically identify fake news and fake news spreaders.

8.5.1 Baselines

We compared the performance of our proposed model with various state-of-the-art (SOTA) approaches in the field of detecting entities in misinformation on the same PolitiFact dataset. We used the features described in Chapter 6 and Chapter 7 as input features of each entity in our proposed model and all considered baseline models. We also implemented a machine learning model that takes those features directly as input to the Random Forest as a baseline model. For comparison with Fake news detection models, we selected recent state-of-the-art approach dFEND [76], BERT+MLP [77] and SAFE [70]. Similarly, for fake news spreaders detection, we compared our proposed approach with the two best performing approaches used by the participants to the PAN CLEF² shared task, namely the approaches proposed by Buda and Bolonyai [212] and Pizarro [213]. These baselines are described here below:

- dFEND [76]: detects fake news by using bidirectional RNN with GRU to capture

²<https://pan.webis.de/>

word and sentence-level representation from news articles and word sequence representation from user comments (tweets) and concatenated of these representations to detect fake news. For the fair comparison on fake news detection, we look at the performance of the variant of dFEND where the information from news articles is only included.

- BERT+MLP [77]: employs cased BERT-Large model to generate embedding from news content. The generated news embedding is then used to predict fake and real news using a two-layer multi-layered perceptron (MLP) to predict news as fake or real.
- SAFE [70]: used Text-CNN [71] to extract features from news content to detect a news item as real or fake.
- Buda and Bolonyai [212]: detects fake news spreaders by using n-grams based approach and combined them with statistical features from the tweets, such as their average length or their lexical diversity. Specifically, they used an ensemble model of Logistic Regression with five sub-models, namely, logistic regression, linear SVM, random forest, and XGBoost with n-grams and XGBoost with statistical features.
- Pizarro [213]: detects fake news spreaders by using a character and word n-grams-based approach with a linear support vector machine as the classifier.
- Features+Random Forest: we used the features generated from news content, user tweets, and source bias from MBFC as input to Random Forest.
- Markov Random Field-based model (MRF's) [79]: we implemented a pairwise

Markov random field-based technique that can model the relationships among the nodes and jointly predict the credibility of each node type. The method is inspired by the work of Rayana and Akoglu [79]. The details on the method are explained in Appendix 8.7.1.

8.5.2 Experimental Setting

We implemented the GNN model using the PyTorch DGL package [226]. Specifically, we addressed the problem of automatically identifying fake news, fake news spreaders, and non-credible publishers as a multi-class classification task.

In the training phase, we adopt a two-layered Role-RGCN model such that the output of the previous layer (l) will then be the input vector to layer $l + 1$. We used ‘mean’ as an aggregation function over each relation $r \in R$ with 150 hidden units and leaky-relu activation function in each layer. We trained the model with 50 epochs, optimized using Adam optimizer, and a learning rate of 0.01. We used the cross-entropy loss function for each node type and minimized the loss over the combined losses of all roles. Finally, the output vector representation associated with each role is obtained that preserves the node features and relation information.

The extracted representation is fed into classifiers with default parameters for node classification. We report results for classification with a Random Forest model as it resulted in overall best performance among tested other classification algorithms, including Logistic Regression and Support Vector Machine.

We used class weighting to deal with class imbalance. For the fair comparison regarding train-test splits with considered baselines [76], we split 75% as train and 25% as a test. We executed the process 5 times and reported the average performance in Table 8.2. The distribution of each node type is retained using stratified k-fold

splitting and the folds are consistent among all experiments including baselines. To deal with potential news information leakage, as we computed ground truth for users depending on the ground truth of news they shared, we recomputed the ground truth of users only based on the news present while training the model. As an evaluation measure, we reported a macro F1-score. The best results are highlighted in bold in the tables.

8.5.3 Experimental Results

Table 8.2 shows the performance evaluation allowing comparison between the performances of baselines and our proposed model on the PolitiFact dataset consistently outperforms all baseline approaches by 6% at the minimum for both news and user. Specifically, we got a macro F1 score of 0.93 vs. the best baseline results of 0.87 achieved by a simple Random Forest classifier for fake news detection and 0.79 vs. the best baseline result of 0.73 achieved by Buda and Bolonyai [212] for fake news spreader detection on the PolitiFact dataset. Similarly, we got a comparable F1 score of 0.35 vs. 0.36 with a simple Random Forest classifier with a very minimal difference of 1%. This could be because very few news publishers in the PolitiFact dataset had ground truth and political bias i.e. (71 out of 197) that were collected from the MBFC website. Using a large number of publishers without ground truth and features might have added unwanted noise in publisher representation. We believe that an improvement can be seen in the prediction for publisher’s credibility degree if we could add a few more information (other features) about publishers. Furthermore, the features we consider for each type of node as explained in Section 8.3.2 achieve better performances than SOTA models for fake news detection.

The performance reported in Table 8.3 shows that our model also outperformed

Approach	News	User	Publisher
dEFEND [76]	0.85		
BERT+MLP [77]	0.71		
SAFE [70]	0.73		
Buda and Bolonyai [212]		0.73	
Pizarro [213]		0.61	
Features + Random Forest	0.87	0.61	0.36
MRF's [79]	0.54	0.39	0.12
Role-RGCN	0.93	0.79	0.35

Table 8.2: Macro F1 score of our proposed model on PolitiFact and comparison with baselines. Best values are in bold.

Approach	News	User	Publisher
RGCN	0.90	0.75	0.19
Role-RGCN	0.93	0.79	0.35

Table 8.3: Macro F1 score of our proposed model on PolitiFact and comparison with classical RGCN. Best values are in bold.

a classical RGCN model.

8.6 Conclusion

In this chapter, we proposed an approach to jointly estimating the credibility degree of entities: publisher, news item, and user in a news ecosystem. We presented the problem of detecting the credibility of each entity as a multi-class classification problem. We address the problem utilizing a representation based on a proposed role-relational graph convolution network (Role-RGCN). In particular, we modified the RGCN model to deal with its limitation of feature size and optimized on combined node-specific losses. We automatically learn the node-level as well as the relational attributes between each entity of the news ecosystem in a supervised way and use these features for identifying the credibility degree of a news publisher, a news item, and a

user jointly. Our proposed approach outperforms existing state-of-the-art methods in the field of fake news detection with 93% f1-score and fake news spreader detection with 79% f1-score.

8.7 Appendices

8.7.1 Markov Random Field-based model

A Pairwise Markov Random Field (PMRF) is a probabilistic graphical model generally used in inference problems on networks [227] which models the relationships among random variables as a factor graph. A factor graph, $H = (A, F, E)$ is an undirected bipartite graph where a set of random variables (nodes) $A = a_1, \dots, a_n$ is on one side and a set of factors $F = \{f_1, \dots, f_n\}$ is on other side. A variable a_i can have any possible state (class) S_i . An edge between a random variable a_i and $f_i \in F$ exists if random variable a_i appears in factor f . A factor is a function $f : a_{(f_i, a_i)} S_i \rightarrow \mathbb{R}$ that evaluates the relationship of all random variables by assigning label to each variable as a real value. A PMRF graph is associated with a set of node (unary) potentials and a set of edge (pairwise) potentials. It is assumed that each variable has exactly one factor called node potential and is denoted as $\Phi(\cdot)$. For instance, node potential for variable a_i is denoted as $\Phi(a_i)$. Node potential represents the prior belief (probability) of a node being in each possible state and is determined based on prior knowledge. Similarly, a factor with two variables is called edge potential and is denoted as $\psi(\cdot, \cdot)$. An edge (compatibility) potentials $\psi(a_i, a_j)$ between node a_i and a_j represents the influence of state of variable a_i on the state of variable a_j . A possible world w is defined as a function $w : A \rightarrow S$ where each variable $a_i \in A$ is assigned a particular value of state $w(a_i) \in S_i$ in a factor graph. There can be multiple possible worlds

$\psi^{t='publish'}$	News	
Publisher	Real	Fake
Least-partisan	$1-\epsilon$	ϵ
Hyper-partisan	2ϵ	$1-2\epsilon$

$\psi^{t='share'}$	News	
User	Real	Fake
RNS	$1-2\epsilon$	2ϵ
FNS	2ϵ	$1-2\epsilon$

$\psi^{t='follow'}$	User	
User	RNS	FNS
RNS	$1-\epsilon$	ϵ
FNS	ϵ	$1-\epsilon$

Table 8.4: Edge potentials for each entity in news ecosystem. Here FNS represents fake news spreader and RNS represents real news spreader.

in a factor graph and each possible world can have a different probability. If the set of all possible worlds is denoted as W , then a factor graph represents a single joint probability of W . The probability of a single world w is given as,

$$P(w) = \frac{1}{Z} \prod_{f \in F} f(w(a_i) \forall a_i : (f, a_i) \in E) \quad (8.3)$$

where Z is the normalization factor. Then, we can compute a marginal probability on a factor graph that infers the probability of a random variable being in a particular state. In a possible world w , the marginal probability of a random variable $a_i \in A$ being in one of the possible states $s_j^i \in S_i$ is given by the the sum of probabilities of all possible worlds where a_i is assigned to state s_j^i and is given by,

$$P(a_i = s_j^i) = \sum_{w \in W, w(a_i) = s_j^i} P(w) \quad (8.4)$$

The loopy belief propagation (LBP) follows an iterative technique of message passing and is often used to approximate marginal probabilities of each node when factors have two variables and the graph contains loops [228].

In our setting, we assume a graph $G = (V, E, T_n, T_r)$ represents the news ecosystem as shown in Figure 8.1 where $P \subset V$ such that for each $p \in P$, $\Phi(p) = \text{“publisher”}$, $N \subset V$ such that for each $n \in N$, $\Phi(n) = \text{“news”}$ and $U \subset V$ such that for each $u \in U$, $\Phi(u) = \text{“user”}$.

We defined edge potentials ψ_{ij}^t where t refers to the type of relation between two nodes (variables); the publisher-news edges $(p_i, n_j, t = \text{publish}) \in E$, the user-news edges $(n_i, u_j, t = \text{share}) \in E$ and user-user edges $(u_i, u_j, t = \text{follow}) \in E$. For a small value of ϵ , we initialized edge potentials ψ_{ij}^t indicating homophily based on the following intuitions,

- There is a high probability that hyper-partisan news publishers tend to publish more fake news and it is less likely for least-partisan news publishers
- There is a high probability that fake news spreaders (real news spreaders) are more likely to share fake (real) news. However, as people are poor at identifying fake news in social media [189], with some probability real news spreaders can also share fake news unintentionally or for fun, whereas fake news spreaders can share real news with the intention to camouflage their malicious activities
- fake news spreader (real news spreader) often follows other fake news spreaders (real news spreaders) in social media with the intention to increase their number to perform malicious activities collectively

The assigned details about edge potential is shown in Table 8.4.

Prior potentials for each node type are computed as described below. We extracted the political bias of news publishers from mediabias/factcheck website. In particular, we assumed that hyper-partisan news publishers tend to be more non-credible as they

tend to publish more questionable news. Therefore, we set prior potential for news publishers as $\Phi(P_i) = \{0.2, 0.8\}$ for hyper-partisan and unbiased $\Phi(P_i) = \{0.5, 0.5\}$ otherwise. For news and user, we extracted features from news content and user's profile and tweets as described in our study [219, 220].

From our previous studies [219, 220], we have an understanding about high or low average value of which features contributes more towards the entity being more suspicious. Therefore, we combine that cues from all features to obtain a stronger estimation of suspiciousness score for each news node and user node. Given a set of F features x_{1i}, \dots, x_{Fi} for node i , we computed CDF of feature x_{li} where $1 \leq l \leq F$ to estimate the probability that the real-valued random variable X_l will have value less than or equal (greater than or equal) to x ,

$$f(x_{li}) = \begin{cases} 1 - p(X_l \leq x_{li}) & \text{if high average value for fake} \\ p(X_l \leq x_{li}) & \text{otherwise} \end{cases} \quad (8.5)$$

Then, we combine these f values to compute the suspiciousness score of a node i as follows.

$$S_i = 1 - \sqrt{\frac{\sum_{l=1}^F f(x_{li})^2}{F}} \quad (8.6)$$

where $S_i \in [0, 1]$ and the prior potentials of class is initialized as $\{1 - S_i, S_i\}$.

The generated posterior potentials are used as the credibility score for each node and are fed into a Random Forest machine learning model to predict the credibility degree of each entity.

CHAPTER 9:

MODELING THE DIFFUSION OF FAKE AND REAL NEWS THROUGH THE LENS OF THE DIFFUSION OF INNOVATIONS THEORY

9.1 Introduction

In recent years, social media has become a diffusion ground for information, including news on important events, breaking news, and emergencies. People encounter a huge amount of news in social media simply because of its ease of use and the nearly frictionless convenience of modification and further sharing [184, 30]. Social networking platforms like Facebook and Twitter facilitate easy access to any kind of news and allow users to build connections with other users and organizations from anywhere simply through a process known as “following.” Users then automatically receive postings made by those they are “following,” and spread those postings through their social-media activities like tweets, retweets, shares, and likes. Taken in aggregate across many users, a principal effect of this process is an effectively uncontrolled spread of information and loss of system-wide vetting of information. As will be described, this is especially the case for news that might be compromised in terms of quality since any independent individual – regardless of their knowledge and

ability to verify the accuracy of items – can act as news creator and distributor on such platforms. This eventually aggravates the possibility of misinformation and fake news dissemination – a possibility which has been identified as a very real societal problem. In order to better understand how this phenomenon may be addressed, it is critical to understand not only what are the characteristics of information consumed by users in social media but also who is likely to share information with friends and followers and what influences them to do so.

The Independent Cascade Model (ICM) and Linear Threshold Model (LTM) are two classical information diffusion models that assume that a user will share a given news item with some probability by only considering that some of their friends have previously shared the same news item [229]. However, recent social science literature on fake news sharing suggests that a user decision of sharing or not sharing a piece of given news does not only depend on the influence of their friends but also on specific characteristics of the users themselves (e.g., demographics, profile properties, behavior and activity, etc.), the news received (e.g., title and content characteristics, etc.), and the social context (e.g., number of followers and following, tie strength, etc.) [230]. All these aspects align with what is theorized by the Diffusion of Innovations Theory to explain how an innovation (which in our case is news) diffuses in a social network [231].

In this chapter, we propose an approach based on the Diffusion of Innovations Theory to model, characterize, and compare how real and fake news is shared in social media. Specifically, we address the following problem: *given that a user u is influenced on some given (real or fake) news item n by at least one of their followees v (i.e., u is following v and v has shared some news item n among their followers),*

predict whether the user u will also share the news item n among their followers.

We model the problem as a binary classification task and propose a set of features inspired by the Diffusion of Innovations Theory [230] that takes into account user, news, and social network characteristics to better predict real and fake news sharing in social media. The set of user-based features we consider include demographics, profile information, personality, emotions, user interest, and behavior. News-based features encode style, complexity, and psychological aspects of news headline and body. Network-based features consider, instead, the user following network to measure tie strength and quantify opinion leadership. Based on our review of related literature, all of these factors have never before been combined into a unique predictive model or tested on a large scale.

To test our proposed approach, we compiled a Twitter dataset of 1,572 users from the FakeNewsNet [51] state-of-the-art data, which contains 18,080 user-news item sharing and not sharing instances of 296 news items (127 real and 169 fake).

Our experimental results show that our proposed approach inspired by the Diffusion of Innovations Theory outperforms the results of classical information diffusion models, i.e., independent cascade and linear threshold models, which we used as baselines. More specifically, we show that the combination of our proposed user-, news-, and network-based features can predict real (resp. fake) news sharing with an AUROC of 97.39 (resp. 97.34) and an average precision of 95.23 (resp. 88.43) (vs. an AUROC of 64.48 (resp. 67.70) and an average precision of 71.83 (resp. 87.45) achieved by the best baseline). Among the proposed features, we observed that news-based features allow greater efficacy in predicting real and fake news sharing, followed by the user-based and network-based features.

Moreover, our analysis reveals some interesting patterns on similarities and differences between real and fake news sharing. We observed that shared real and fake news is typically shorter and contains more lexical diversity and less expression of negative emotions. Users who shared such news have higher tie strength, meaning they retweeted a higher percentage of tweets by their followees, even while news sharing is affected by the serendipitous qualities of social media. We also observed that shared real and fake news differs mostly based on syntactic features such as part-of-speech (POS) and punctuation, and that users who share fake news are ≥ 40 years of age, have fewer followers than followees, and are politically right-leaning.

9.2 Related Work

Studies have been conducted in computer science and social science to improve our understanding of the characteristics of users who are likely to contribute to spreading fake news on social networks.

Vosoughi et al. [64] found out that fake news spreaders had, on average, significantly fewer followers, followed significantly fewer people, and were significantly less active on Twitter. With respect to the social media platform Twitter, although bots contribute to spreading fake news, the dissemination of fake news on Twitter is mainly caused by human activity.

The characteristics of users who are likely to trust or distrust fake news have been analyzed by Shu et al. [66]. By looking at Twitter user profile data, Shu et al. found that, on average, users who share fake news are newer to the platform (shorter time since registration) than the users who share real news. Additionally, while bots were shown to be more likely to post a piece of fake news than real news, users who spread fake news are still more likely to be humans than bots. They also show that users

who are keen to spread real news are more likely to be more popular than users who are keen to spread fake news and that fake news is more likely spread by older people and females.

User demographics as predictors of fake news sharing have been analyzed by Guess et al. [67] on the Facebook platform where political orientation, age, and social media usage turned out to be the most relevant factors. In this study, the majority of fake news items included for analysis were dated from 2016 and were typified by pro-Trump sentiments. The researchers found that users who leaned to the political right were more likely to share those fake news items. Additionally, individuals identified as senior citizens tended to share more fake news (a fact the researchers theorized to be due to age-associated lower digital media literacy skills necessary to assess online news truthfulness). Finally, the researchers found that the more news people post in social media, the less likely they are likely to share fake news, an observation theorized to be the case because those users would be more familiar with the platform and what they share.

The author profiling shared task at the PAN at CLEF 2020 conference focused on determining whether or not the author of a Twitter feed was keen to spread fake news[69]. The teams who participated proposed different linguistic features to address the problem, including (a) n-grams, (b) style, (c) personality and emotions, and (d) embeddings. Among them, Shrestha et al. [68] showed that psycho-linguistic and personality features are significantly associated with user sharing behavior.

When dealing with news spread modeling in social media, researchers usually study the underlined network among its users. The considered models can be divided into two different categories according to network observability: (a) Independent

Cascade and Linear Threshold models assume the user connections to be explicit, while (b) Hawkes processes or epidemic models predict the number of infected people over time by working with an implicit network [232, 233, 234].

However, just considering the network to explain news diffusion may not be sufficient; hence some works tested hypotheses inspired by the Diffusion of Innovations Theory, which also considers news and user properties as important factors to explain news sharing behavior [235]. For instance, in a study conducted by Ma et al. [190], news preference, opinion leadership, and tie strength have been shown to be the most important factors at predicting news sharing, while homophily hampered news sharing in users' local networks. Also, Lee and Ma [191] have shown how people driven by gratifications of socializing, information seeking, and status-seeking were more likely to share news on social media platforms. This highlights a chicken-and-egg situation where it becomes difficult to identify a causal sequence of phenomena between actors and network effects.

9.2.1 Diffusion of Innovations Theory

Diffusion of innovations theory has origins in social science of the mid-20th Century. Starting with agriculture and the substantial impact of GMO technologies on the robustness of plants and yields, its focus has always been on modeling the rate and the spread of acceptance (i.e., diffusion) of technological innovations in settings where those innovations will challenge the stability of existing practices, and affect and perhaps challenge (a) established systems of authority, (b) stakeholders at multiple levels in existing social structures, and (c) the strength and influence of systems of knowledge, skills, and values that were previously external to the setting in which innovations are introduced [231].

Unlike many critical social science theories that focus on colonizing activity and perhaps aggression as a means of introducing innovations, the theory and resulting models describing the diffusion of innovations focus on how actors within the receiving system are essential to successful diffusion. With this in mind, technologies and innovations are not (or not only) seen as having “power over” incumbents in the receiving system, but as actants in a newly articulated network, that when adopted and diffused, afford “power to” actors who did not previously have such a position in the receiving system [236].

It is this orientation to networks through which innovations and their promoters come to have influence, that we use the diffusion of innovations theory in this chapter. We research the interaction of characteristics of (a) real and fake news items, (b) actors (users) who circulate those items in the imputably less-hierarchical environment of social media, and (c) the uptake and sharing of real and fake news items by other actors, to help us produce methods for predicting the existence of fake news.

9.3 Dataset

We used the PolitiFact dataset from FakeNewsNet to carry out our experiments. The size of this dataset is shown in Table 9.1. FakeNewsNet [51] is a well-known data repository that consists of two datasets, PolitiFact and GossipCop, from two different domains, i.e., politics and entertainment gossip, respectively. Each of these datasets contains details about news content, publisher, social engagement information, and social network. In this chapter, we only used the PolitiFact dataset (as gossip is different from fake news), which contains news with known ground truth labels collected from the fact-checking website PolitiFact¹ where journalists and domain

¹<https://www.politifact.com/>

Table 9.1: Size of the PolitiFact dataset.

# Users	# News	# Tweets	# Retweets	# Real News	# Fake News
281,596	992	438,504	619,239	560	432

experts fact-checked the news items as fake or real.

In order to test our proposed method to predict news sharing by influenced users, we computed the labels for user sharing or not sharing a given piece of news, as follows. First, we computed pairs of influencer and influenced users. An influencer is a user who tweeted a given piece of news, and an influenced user is a follower of that influencer. We considered (a) users (influencers) who have shared at least one news and have at least one follower (influenced user), and (b) followers who shared at least five instances of news (and ordered those instances in chronological order of shared time). Hence, we annotated news items as shared or not as follows:

- **Shared:** a piece of news that is shared/tweeted by a user (influencer) and then shared/retweeted by its follower (influenced user) as one of their two most recently shared news items.
- **Not Shared:** a piece of news that is shared/tweeted by a user (influencer) but never shared/retweeted by their follower.

For example, let us consider a sample news sharing network as shown in Figure 9.1. We have three users $U = \{u_1, u_2, u_3\}$ and three news $N = \{n_1, n_2, n_3\}$ whose veracity could be either real or fake. Each edge label specifies the interaction type between two entities. Here, user u_1 (influencer) is followed by users u_2 and u_3 (influenced users) and among the news items n_1 , n_2 and n_3 shared/tweeted by user u_1 , user u_2 shares/retweets news items n_1 and n_2 . Therefore, the instance $\langle u_1, u_2, n_1 \rangle$ and

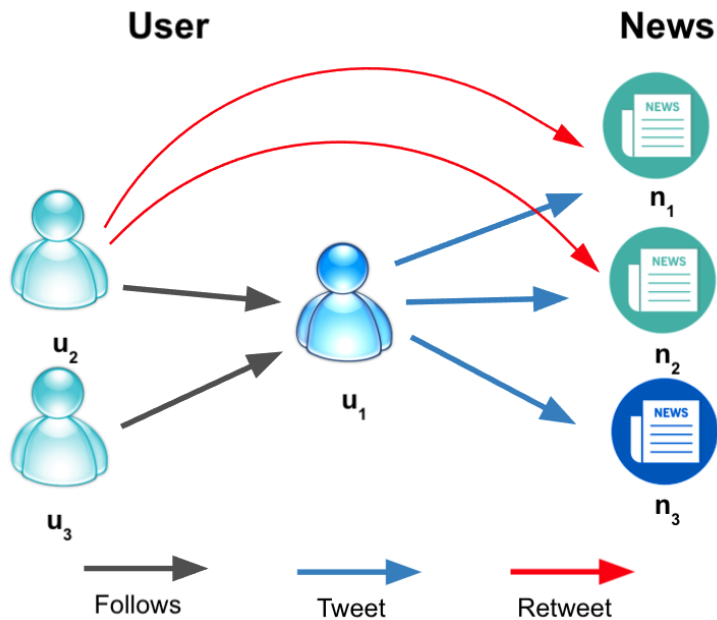


Figure 9.1: News sharing network between users in Twitter.

$\langle u_1, u_2, n_2 \rangle$ are labelled as *Shared* whereas the instance $\langle u_1, u_2, n_3 \rangle$ as *Not Shared*.

We used the remaining three or more news shared by followers to profile these users and compute the user's interest similarity with the given news item, as discussed in 9.4.1. In addition, for each follower, we crawled all tweets posted one month prior to the publish date of the shared news to compute remaining user-based features. We have a total of 1,572 influenced users in our dataset sharing real and fake news, as shown in Table 9.2.

Also, in order to compare real and fake news sharing dynamics, we consider two different datasets, one containing only sharing instances of real news items and the other containing only sharing instances of fake news items (cf. Table 9.2).

Table 9.2: Size of the datasets used in our experiments.

Dataset	# Users	# News	# Shared	# Not Shared
Fake News Sharing	1,557	169	527	7,134
Real News Sharing	1,572	127	2,617	7,802

9.4 Features

This section describes the three sets of features, namely user-based, network-based, and news-based features, that we considered to implement Diffusion of Innovations for modeling news sharing.

9.4.1 User-Based Features

Demographics

As the first group of user-based features, we consider user demographics which include age, gender, and political ideology. Previous research has highlighted how these features influence users' news-sharing behavior. For instance, according to the findings by Reis et al. [195], male and white users potentially share more news on Twitter, while the work by Shu et al. [66] has shown that older people and females are more likely to spread fake news on Twitter.

Typically, demographic features are not explicitly available on social media platforms. Hence, we used machine-learning-based methods to infer such attributes for the users, as detailed here below.

- Age and Gender: We utilized m3inference [196], a deep-learning-based system trained on Twitter data, to infer user age and gender. Based on the available metadata such as username, screen name, description, and profile image, m3inference predicts the *gender* of the user as male or female and the *age* of

the user grouped in four categories (≤ 18 , 19–29, 30–39 and ≥ 40).

- **Political Ideology:** The user political ideology can provide additional information about the user’s sharing behavior. In this chapter, we computed a user political leaning by using *#polar score*, a method defined by Hemphill et al. [197]. This method uses the hashtags used by users in their tweets to estimate their political ideology. In our implementation, a signed Chi-Squared score is first learned for each hashtag from the dataset of U.S. Congress members [198] with known political affiliation provided by Chamberlain et al. [198]. Next, for each user, the *#polar* scores of each hashtag used in their tweets are averaged to compute a global *#polar* score for the user. A positive user polar score indicates that the user inclines towards right-leaning political ideology. Vice versa, a negative user polar score indicates a left-leaning political ideology.

Explicit Features and Activity

We consider Twitter available user profile information as explicit features. This group of features includes:

- **Protected:** indicates whether a user has chosen to protect their tweets or not
- **Verified:** indicates whether a user is a verified user
- **Register Time:** the number of days passed since the registration of the account
- **Status count:** indicates the number of tweets (including retweets) by a user
- **Favor count:** indicates the number of tweets a user has liked

Shu et al.[237] found that these features exhibit significant patterns among the users that are likely to trust/distrust fake news and propagate them on Twitter. Thus, including these features can provide additional cues to determine the news sharing behavior of a user in social media.

In addition, we computed the user *insomnia index* as in [200] to analyze the user tweeting behavior within the day (24 hours). Specifically, we divided the time into day and night and considered the ‘night’ window as ‘9PM-6AM’ and the ‘day’ window as ‘6:01AM-8:59PM’ (we used the local time of the user), and analyzed the normalized difference between the number of tweets shared during these time windows (percentage of day posts and night posts) for each user.

Personality

To compute personality features, we leveraged IBM Watson Personality Insights service that uses linguistic analytics to infer individuals’ intrinsic personality characteristics, including Big Five personality traits, Needs, and Values, from digital communications such as social media posts. For this, we concatenated all the user tweets in a unique document to compute their personality characteristics as in [219]. We considered the following features (raw scores) provided by IBM Watson Personality Insights service:

- Big Five: The service computes the user Big Five personality traits described by the five-factor (FFM) or OCEAN model, a widely used taxonomy to describe people’s personality traits [203]. This taxonomy identifies five basic personality dimensions, which are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. For each personality dimension, the IBM

Watson Personality Insights service also provides a set of additional six facet features. For instance, extraversion' facets include excitement seeking, activity level, cheerfulness, assertiveness, friendliness, gregariousness.

- Needs: These features describe a user's needs as inferred by the text they wrote and include excitement, harmony, curiosity, ideal, closeness, self-expression, liberty, love, practicality, stability, challenge, and structure.
- Values: These features describe the motivating factors that influence a person's decision-making. They include self-transcendence, conservation, hedonism, self-enhancement, and openness to change.

Emotion

As people express their emotions, appraisals, and sentiments towards any news or article through the choice of words in their tweets, we tried to capture the emotion of users from their tweets. To compute these features, we concatenated all the tweets by each user to form a single document per user.

- Emotion Intensity: To determine the intensity of emotions such as anger, joy, sadness, fear, disgust, anticipation, surprise, and trust, we leveraged the Emotion Intensity Lexicon (NRC-EIL) [238]. Next, we computed feature vectors using the approaches proposed in [219]. Each lemmatized word in the text after removing all stopwords is looked up in the emotion intensity dictionary, and intensity scores of matching words are averaged element-wise to generate an emotion vector representation of the text.
- Stress: Along with the above-mentioned emotions, characteristics of stress such

as worries, feelings of frustrations, and irritations, can also be captured by the language used in the tweets. Thus, we also considered a feature capturing the user stress level, which we computed using the lexical dictionary created by Wang et al. [202]. As preprocessing, we removed words like 'RT,' 'Via,' and '&' from each tweet.

- **Sentiment Analysis:** The choice of words in a user's tweets can depict their emotional state. Hence, we performed sentiment analysis by using the Valence Aware Dictionary and sEntiment Reasoner (VADER) [239], a library specifically built for capturing sentiments expressed in social media texts. For each user, we measured the average sentiment (positive, negative, and neutral) across all their tweets.

User Interest in News

We compute the user interest in a given piece of news by making use of Latent Dirichlet Allocation (LDA), a popular algorithm for topic modeling. We used the implementation from the Gensim package [240] and trained the model with 100 topics on Wikipedia data to infer topics of the text in our dataset. Specifically, we considered the following two approaches to compute the topical similarity between user's interest and shared news item:

- **Similarity 1:** cosine similarity between the topics extracted from the news item to be shared and the topics extracted by the document containing the concatenation of all the user's previously shared news (three or more news items shared by the influenced user, as discussed in Section 9.3).

- Similarity 2: cosine similarity between the topics extracted from the news item to share and the concatenation of all the tweets from the timeline of the influenced user.

9.4.2 Network-Based Features

Twitter Follower-Following (TFF) ratio

Vosoughi et al. [64] have shown that fake news spreaders had fewer followers and followed fewer people than real news spreaders. Thus, in this chapter, we computed the Twitter follower to following (TFF) ratio as in [66] to measure user connectivity in the Twitter social network. TFF is computed as $TFF = \frac{\#Follower+1}{\#Followee+1}$ which indicates the ratio of the number of followers to the number of followees of the user. The greater the ratio, the higher the popularity of the user.

Weak and Strong ties

According to Ma et al. [190], perceived tie strength in online social networks is positively associated with news sharing intention in social media. Granovetter [241] showed that strong ties are the friends and weak ties represent acquaintances and the information diffuses faster among people with strong ties. While strong ties indicate a large number of shares between two people in a network, weak ties depict fewer shares between influencer and follower.

To compute the tie strength between the influencer and influenced user pairs, we considered two different approaches, described here below:

- Receiver's perspective tie strength: For a given influencer-influenced user pair in the dataset, we computed, out of all the retweets by the influenced user, what

percentage of them were also previously tweeted by the given influencer².

- Time-based tie strength: the average time taken by the influenced user to retweet/share news items previously tweeted by the given influencer.

Degree Centrality and PageRank

A user in a social network can be characterized based on their connectivity in the network. Measuring how central and popular users are in the network uncovers the influential users.

- Degree centrality is the simplest centrality measure that counts the number of neighbors a node has. Since Twitter has directed networks, we computed two measures of degree: in-degree, which indicates the number of followers a user has, and out-degree indicates the number of followings of a user. In- and out-degree centrality have an important connection in evaluating the influence of a node in a network. The higher the in-degree of a node, the more influential the node.
- PageRank measures the popularity of nodes in a network [235]. Unlike degree centrality, PageRank tries to account for both quantity and quality to evaluate how important a node is in a network. Meaning, Page Rank computes how influential a node is based not only on the count of followers but also on the quality of followers (i.e., how influential a follower is).

²We did not consider influencer's perspective, i.e., which percentage of influencer's tweets have been retweeted by the follower since it is the same as the Bernoulli distribution we have used as a baseline (cf. Section 9.5.2).

9.4.3 News-Based Features

As news-based features, we considered the ones proposed in our previous work [219] to detect fake news on the same PolitiFact dataset used in this chapter. These features include stylistic, psychological, and complexity features that are computed for both title and body text of news items.

Stylistic Features

We considered a subset of the features computable by the 2015 Linguistic Inquiry and Word Count (LIWC) tool [174] that represent the functionality of text to capture the user writing style from the sets of their authored tweets. This set of features includes:

Regarding the part of speech features, we used the Python Natural Language Toolkit part of speech (POS) tagger to compute the number of nouns (NN), proper nouns (NNP), personal pronouns (PRP), possessive pronouns (PRP\$), Wh-pronoun (WP), determinants (DT), Wh-determinants (WDT), cardinal numbers (CD), adverbs (RB), interjections (UH), verbs (VB), Adjective (JJ), past tense verbs (VBD), gerund or present participle verbs (VBG), past participle verbs (VBN), non-3rd person singular present verbs (VBP), and third-person singular present verbs (VBZ).

Psychology Features

Social psychology is the study of the dynamic interaction between individuals and the people around them. Psychology plays an important role in the field of social media. Thus, we computed the positive (pos) and negative (neg) sentiment metrics using the LIWC tool and emotion features, such as anger, joy, sadness, fear, disgust, anticipation, surprise, and trust by using the Emotion Intensity Lexicon (NRC-EIL)

[238] and the approach proposed in [219]. The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). We computed these scores for both text and title of different news.

Complexity Features

The complexity of text in natural language processing depends on how easily the reader can read and understand a text. We used the Simple Measure of Gobbledygook Index (SMOG) readability measure as a complexity feature in our analysis [242]. Higher scores of readability indicate that the text is easier to read. This group of features also includes lexical diversity or Type-Token Ratio (TTR) [219] and the average length of each word (avg wlen).

9.5 Experiments

This section reports on our experimental results of using the features described in Section 9.4 to model real and fake news sharing.

9.5.1 Experimental Settings

We used a binary classification task to automatically identify whether a user will share a news item or not. Specifically, we used the features described in Section 9.4 as input to various machine learning algorithms, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Extra Trees Classifier, and selected the best performing classifier as our proposed model. As our data is unbalanced, we used class weighting to deal with it and performed 10-fold cross-validation. We considered the Area Under the ROC Curve (AUROC) and Average Precision (AvgP)

as evaluation metrics, which are well-suited for unbalanced data.

9.5.2 Baselines for Comparison

As our model is predicting which users will become infected (as opposed to the number of infected users), we compared with two well-known information diffusion models, namely the Independent Cascade Model (ICM) and the Linear Threshold Model (LTM) [232], as they work with the explicit network.

The Independent Cascade Model is a stochastic information diffusion model where nodes can be in two states: *active*, meaning that the node is already influenced by the information in diffusion, and *inactive*, meaning that the node is unaware of the information or not influenced by the information in diffusion. At each step, a newly active node u has the chance to influence an inactive neighbor v according to an influence probability p_{uv} . Each probability p_{uv} is independent of the others. According to Li et al., several ways exist to determine influence probabilities [243] from propagation data. Among them, we considered the following two heuristics proposed by Goyal et al. as they are scalable for large data [244]:

- *Bernoulli distribution.* Under this heuristic, the influence probability p_{uv} is computed as $p_{uv} = \frac{A_{u2v}}{A_u}$, where, in our case, A_u is u 's total number of tweets and A_{u2v} is the number of u 's tweets retweeted by v .
- *Jaccard Index.* Under this heuristic, the influence probability p_{uv} is computed as $p_{uv} = \frac{A_{u2v}}{A_{v|u}}$, where, in our case, $A_{v|u}$ is the number of tweets either u or v tweeted and A_{u2v} is the same as before.

In the Linear Threshold Model, each edge (u, v) is associated with a weight b_{uv} and each node u has a threshold value θ_u . Threshold values are generally assigned

uniformly at random to nodes from the interval $[0,1]$. At each step i , a node v will become active if $\sum_{u \in N_{in}(v), u \in A_{i-1}} b_{uv} \geq \theta_v$, where E is the set of edges in the network, $N_{in}(v) = \{w | (w, v) \in E\}$ is the set of v 's incoming neighbors, and A_{i-1} is the set of nodes that are active in the previous step. As reported Li et al. [243], the two most adopted heuristics to assign probabilities b_{uv} are to (1) uniformly assigning the edge (u, v) with a probability from the set $\{0.1, 0.01, 0.001\}$ at random, or (2) setting b_{uv} equals to the inverse of the in-degree of node v .

In order to compute AUROC and average precision for baselines, we scored each instance $\langle u, v, n \rangle$ in our dataset, where u is the node tweeting news item n and v is a follower of u , as follows. For the independent cascade model, the score is set to the independent probability p_{uv} . For the linear threshold model, the score is set to u 's Shapley value $\varphi(u)$ of the following coalition game. The set N of players is given by the set of all nodes tweeting news n , and the characteristic function $\nu : 2^N \rightarrow \mathbb{R}$ is defined as follows: $\nu(H) = 1$ if node u is in the set of influenced nodes according to a linear threshold model where H the set of seed nodes, and $\nu(H) = 0$ otherwise. Thus, $\varphi(u)$ quantifies the contribution of node u in influencing v . We normalized $\varphi(u)$ by dividing it by $\max_{w \in N} \varphi(w)$ under the linear threshold model.

9.5.3 Results and Analysis

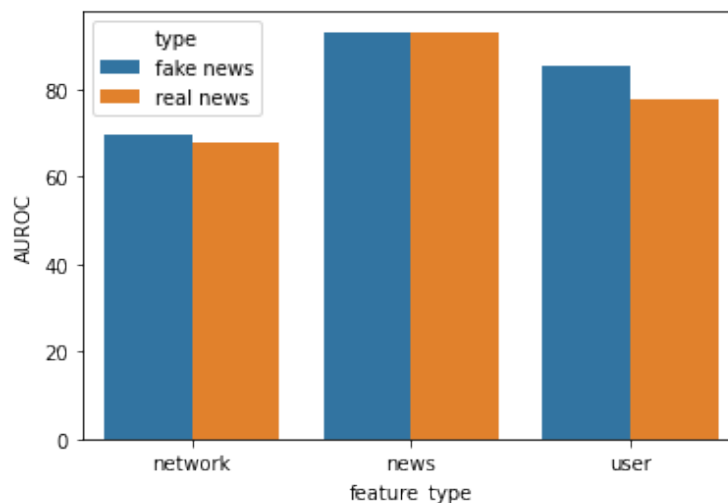
Classification results are reported in Table 9.3 according to AUROC and average precision. As the table shows, among all the considered classifiers, XGBoost achieved the best results with an AUROC of 97.34 and average precision of 88.43 for fake news sharing, an AUROC of 97.39 and an average precision of 95.23 for real news sharing. Further, our model with proposed features consistently outperformed both baseline models with a margin of 30% in AUROC, approximately. Specifically, the

Table 9.3: Performance of our proposed features according to different classifiers on both real and fake news sharing and comparison with baselines. Best values are in bold.

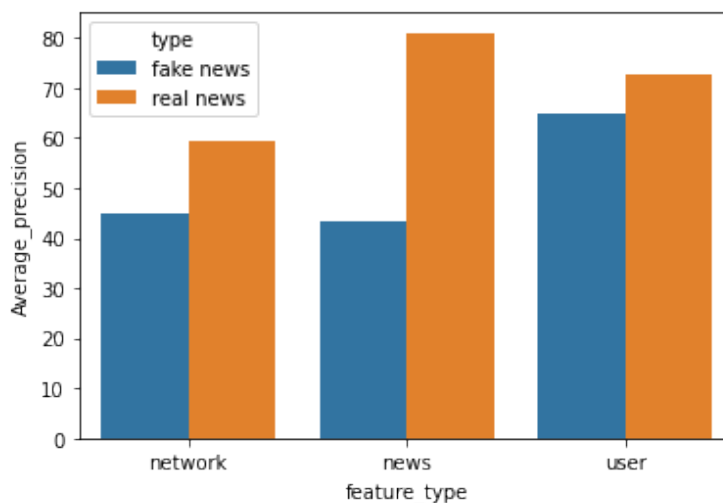
	Fake News		Real News	
Classifier	AUROC	AvgP.	AUROC	AvgP.
Logistic Regression	93.78	65.22	92.39	84.15
SVM	89.08	36.58	87.05	66.91
Random Forest	97.86	85.11	97.38	94.61
Extra Trees	96.82	76.75	95.26	85.94
XGBoost	97.34	88.43	97.39	95.23
ICM (Bernoulli)	67.67	56.72	64.41	48.54
ICM (Jaccard)	67.70	53.03	64.48	46.05
LTM (Random)	64.20	78.57	59.35	63.77
LTM (Inverse Degree)	63.78	87.45	56.45	71.83

best baselines achieved an AUROC of 67.70 (ICM with Jaccard index) and an average precision of 87.45 (LTM with inverse degree) for fake news sharing, and an AUROC of 64.48 (ICM with Jaccard index) and an average precision of 71.83 (LTM with inverse degree) for real news sharing.

In addition, we considered each group of features (network-based, news-based, and user-based) in input to the best classifier (i.e., XGBoost) to measure their contribution to the implemented diffusion of innovations to model real and fake news sharing. As Figure 9.2 shows, a significant amount of contribution in decision-making (for both real and fake news sharing) is given by news-based features, followed by user-based features and network-based features according to AUROC. Although the same trend of contribution can be seen for real news sharing (i.e., news-based > user-based > network-based), user-based features are the most important group of features followed by network-based features and news-based features for fake news sharing according to



(a) AUROC



(b) Average Precision

Figure 9.2: Feature group analysis: AUROC and average precision per feature group for shared real and fake news.

average precision as shown in Figure 9.2b. Thus, our findings reveal that, in general, network-based features are not enough for predicting news sharing, and better results can be obtained by coupling these features with user- and/or news-based features.

Thus, to further investigate this trend, we performed statistical feature analysis as

Table 9.4: Features that differ in fake and real news sharing. S means shared and NS means not shared. All differences are statistically significant ($p < 0.05$). Same feature with same trend for both title and text is denoted as ‘t&t’.

	Features	Fake News	Real News	Features	Fake News	Real News	Features	Fake News	Real News	
News	neg t&t	$NS > S$	$NS > S$	NNP t&t	$NS > S$	$NS > S$	focuspast t&t	$NS > S$	$NS > S$	
	negemo text	$NS > S$	$NS > S$	AllPunc title	$NS > S$	$S > NS$	WPS title	$NS > S$	$NS > S$	
	WC t&t	$NS > S$	$NS > S$	Quote text	$NS > S$	$S > NS$	shehe text	$NS > S$	$NS > S$	
	VB t&t	$NS > S$	$S > NS$	Anger title	$NS > S$	$NS > S$	Sadness title	$NS > S$	$NS > S$	
	TTR text	$S > NS$	$S > NS$	VBG t&t	$NS > S$	$S > NS$	VP t&t	$NS > S$	$S > NS$	
	Trust title	$NS > S$	$S > NS$	RB t&t	$NS > S$	$S > NS$	PRP text	$NS > S$	$S > NS$	
	Fear title	$NS > S$	$NS > S$	Disgust text	$NS > S$	$NS > S$	smog index title	$NS > S$	$NS > S$	
	Surprise t&t	$NS > S$	$NS > S$	JJ text	$NS > S$	$NS > S$	WP text	$NS > S$	$S > NS$	
	shehe title	$NS > S$	$S > NS$	DT text	$NS > S$	$S > NS$	work text	$S > NS$	$NS > S$	
	WDT text	$NS > S$	$S > NS$	VDN t&t	$NS > S$	$NS > S$	i title	$NS > S$	$S > NS$	
	PRP title	$NS > S$	$S > NS$	VBD title	$NS > S$	$NS > S$	work title	$NS > S$	$S > NS$	
	Exclam title	$NS > S$	$S > NS$	PRP\$ title	$NS > S$	$S > NS$	PRP\$ text	$NS > S$	$S > NS$	
	Tone text		$S > NS$							
	User	≤ 18	$NS > S$	$S > NS$	19 – 29	$NS > S$	$S > NS$	30 – 39	$NS > S$	$S > NS$
		≥ 40	$S > NS$	$NS > S$	polar score	$S > NS$	$NS > S$	day posts (%)	$NS > S$	$S > NS$
		friends	$S > NS$	$NS > S$	favourites	$NS > S$	$S > NS$	statuses	$NS > S$	$S > NS$
register time		$S > NS$		weekday posts	$S > NS$	$NS > S$	weekend posts	$NS > S$	$S > NS$	
big5 neuroticism		$NS > S$	$S > NS$	need stability	$NS > S$	$S > NS$	need love	$NS > S$	$S > NS$	
value self enhancement		$NS > S$	$S > NS$	value conservation	$NS > S$		need curiosity	$S > NS$		
need harmony		$NS > S$		big5 extraversion	$NS > S$		need closeness	$NS > S$		
big5 openness		$S > NS$		need liberty	$S > NS$		value hedonism	$NS > S$		
need self expression		$S > NS$		Surprise	$NS > S$		Anticipation	$NS > S$	$S > NS$	
Trust		$NS > S$	$S > NS$	Anger	$NS > S$		Joy	$S > NS$		
neg		$NS > S$		pos	$S > NS$		similarity 1	$NS > S$	$NS > S$	
similarity 2		$NS > S$	$NS > S$							
Network	deg in	$S > NS$	$NS > S$	deg out	$NS > S$	$S > NS$	pagerank	$S > NS$	$NS > S$	
	TFF	$NS > S$	$S > NS$	tie strength (RP)	$S > NS$	$S > NS$				

in [219]. We would expect that the uncovered statistically significant feature values that differ among different labels, i.e., shared and not shared for real and fake news, would help machine learning classifiers to clearly separate data and classify precisely. Table 9.4 highlights the pattern difference among news shared and not shared for both real and fake news. If the value of a feature was higher (on average) for a piece of news shared as compared to not shared, it is denoted as $S > NS$ (and $NS > S$ vice versa) in the table.

Similarities We start by discussing similarities between real and fake news sharing behavior. Regarding news-based features, we observe that, independently of news veracity, shared news items express a less negative sentiment (neg) and fewer emotions (anger, fear, surprise, sadness, disgust, and negative emotion) as compared to non-

shared news items. From a stylistic point of view, shared news items have a shorter content (WC) and use fewer words per sentence (WPS) in the title, fewer proper nouns (NNP) and adjectives (JJ), and fewer third-person pronouns (shehe) in the text, and focus less on past (focuspast, past tense verbs (VBD) and past participle verbs (VBN)) than non-shared news items. Also, shared news (both real and fake) has more lexical diversity (TTR) in the body, but news titles require a lower educational level to be read (SMOG). Dealing with user-based features, we observe a similar behavior regarding user interest in shared news (both real and fake), i.e., there is a lower user-news similarity for shared news as compared to non-shared news, both in terms of previously shared news (similarity 1) and timeline tweets (similarity 2). One motivation could be the serendipitous attributes of social media, where users can easily access a vast amount of diverse news topics. Moreover, social media users can easily connect to other users by following them, which ideally creates an environment where information flows from one user to another leading to the possibility that a user might encounter a news item that neither matches with its interest profile nor with previously shared news by its followers. Similarly, both real and fake news is shared mostly by users who have higher tie strength based on the receiver's perspective indicating a higher percentage of tweets were tweeted after being persuaded by their followee.

Differences When we look at news-based differences between real and fake news sharing, we see that people tend to share fake news that is characterized by less use of parts of speech such as PRP, PRP\$, VBG, VBP, RB, VB, WP, WDT, DT, VBP, has fewer quotes (Quote) in news body text, uses less punctuation (allpunc), e.g., exclamation marks, and fewer work related words (work) in the title (but more in

news body text) and tends to express less positive emotion such as Trust. Whereas, we notice the above trends to be the opposite for shared real news. Moreover, we observe that users who shared fake news are 40 or more years old (while users under 40 shared less fake news) and have a right-leaning political orientation (polar score), and they tend to post fewer tweets during day time (post day perc) and weekend (but more in weekdays), have fewer favorites and statuses counts and tends to express less positive emotions (Trust, Anticipation) in their tweets. They also portray a less neurotic personality, are less motivated by self-enhancement, and care less about love and stability. Similarly, users sharing fake news have fewer followers than followee (TFF, friends count, deg out, deg in). However, sometimes a user may have few followers (low in-degree), but those followers could be highly influential, leading to a higher page rank score. Thus, users who shared fake news are highly influenced by their followee. The opposite holds for users sharing real news.

However, we can say that not all features are statistically significant for both real and fake news sharing. Among these features, we observed that the shared fake news body text has fewer possessive pronouns (PRP\$), whereas shared real news body text tends to use more emotional tone words. Regarding users who shared fake news, we observed that they are older according to register time, tend to express fewer negative emotions (for example, Anticipation, Anger, neg), lower trust and surprise. Posts by members of these same groups include more positive emotions via their tweets (for example, pos, Joy). Moreover, personality traits associated with negative emotions show that they exhibit ambivert characteristics in their Tweets by reflecting less appreciation to other's feelings (harmony), fewer connections with family/friends (closeness), a lower willingness to interact with others (extraversion), less hedonism

and conservation, and, at the same time, more liberal, self-expressive, curious and open to new experiences. This orientation simply to “new experiences” and “self expression” has been associated with an orientation to both verified news items and fake news items in the interest of “intellectual diversity” [245], a term commonly associated with interests of individuals expressing agreement with the Conservative political position in the United States.

9.6 Conclusion

In this chapter, we addressed the problem of modeling news sharing in social media. Specifically, we proposed and implemented an approach based on the Diffusion of Innovations Theory using data extracted from Twitter to model, characterize, and compare how real and fake news is shared in social media.

Consistent with the Diffusion of Innovations theory, we have found that there is no single point of authority that can be said to cause successful diffusion of fake news in networks included in our analysis. Similarly, there is no single point of failure. We have shown that it is the dynamic interaction of (a) news features, (b) user features, and (c) emergent and gradually stabilizing network features that produce successful diffusion. Consequently, identifying means of predicting fake news sharing requires multiple perspectives on this dynamic interaction.

More specifically, we considered the problem of predicting whether a user will share a news item with followers, given that the news item was shared by one of their followees (that is, prediction a user was influenced by factors associated with that given news item and/or factors associated with the followee’s sharing of the item). To study this question, we proposed using three main sets of features, (a) news-based features computed from news headline and body, (b) user-based features computed

from user Twitter feed and profile, and (c) network-based features computed from the user following network, and performed a comprehensive analysis on a large Twitter dataset with ground truth defined as news *Shared* and *Not Shared*.

Our experiments showed that the proposed features permitted prediction of real and fake news sharing that outperformed considered baseline approaches. Specifically, we obtained an AUROC of 97.39 (resp. 97.34) and average precision of 95.23 (resp. 88.43) for predicting real (resp. fake) news sharing. Further, our analysis revealed that other features beyond classical network-related features must be considered in order to produce high-confidence modeling of real and fake news sharing, and effectively highlighted distinctive patterns of similarities and differences between real and fake news sharing.

Part V

Concluding Remarks

CHAPTER 10:

CONCLUSION AND FUTURE DIRECTIONS

In this chapter, we summarize the contributions of this dissertation and highlight the potential research directions for the future.

10.1 Conclusion

In this dissertation, we present our in-depth analysis on false information prevalent on the web and various types of actors interacting with it such as trustworthy, unreliable, and fraudulent reviewers, and fake and real news spreaders. Specifically, our analysis focus on interactive platforms like e-commerce and rating platforms Such as Amazon and Yelp!, and social media platforms such as Twitter in order to understand and identify false information and actors responsible for its creation and spread.

In Chapter 3, we investigated why false information is succeeding in their motive and shows that humans are less accurate in identifying fake news when they have only the text of the article compared to when they rely on meta-data, especially image. We also revealed that the professionalism of the image was a helpful heuristic that enabled more accurate judgments among the participants.

Next in Chapter 4, we measured the impacts of fraudulent reviews in the recommender system where we observed that users are exposed to better recommendations when spammers are excluded and Amazon non-mainstream users are the ones who

are most affected by shilling attacks.

Further in Chapter 5, we built an automatic framework, DeepTrust to learn relevant features from the user’s temporal review sequence in an unsupervised way and use these features for classifying users into trustworthy, unreliable, or fraudulent reviewers. Our DeepTrust framework achieves an F1-score of 93% at classifying those types of users. In Chapter 6, we learned the characteristics of fake news on two different news domains such as political and gossip using psycho-linguistic, complexity of the text, emotion, and stylistic attributes extracted from news contents. We found that, although majority of the findings by Horne and Adali [86] were confirmed from our analysis, some of the observations were not generalizable. For instance, their findings that fake news articles use smaller words and fewer quotes and fake titles contain fewer stop words were not generalized in our analysis. Furthermore, our results highlighted some new patterns such as real news articles use more positive tone and are more descriptive than fake news articles, and fake news titles and bodies express more negative emotions than real news. We also found that the political news domain is different than gossip news domain with more religion-related words in fake political news articles. Similarly, in Chapter 7, we characterized fake news spreaders based on their demographics, network-based and psycho-linguistic, emotion, and personality extracted from their tweets. We found some prominent patterns of fake news spreaders in terms of demographics, their behavior, and writing styles. For instance, we found that the users under 18 or over 40 and predicted females may be more vulnerable in case of fake news sharing and fake news spreaders tend to express more negative emotion and stress in their tweets. We also show the predictive performances of the learned characteristics in accurately identifying fake news spreaders with an

average precision of 99% on the PolitiFact dataset and 80% on the PAN dataset.

Likewise, in Chapter 8, we built Role-RGCN, a model based on heterogeneous graph representation learning for the joint estimation of credibility degree of entities involved in the news ecosystem utilizing the understanding gained from Chapter 6 and Chapter 8. We modified the classical RGCN model and improved over its limitation that it requires the node features to be of the same size for all roles. We evaluated the proposed model using the news ecosystem and addressed the problem of estimating the credibility degree of each node with a specific role as a multi-class problem where the performance shows the high accuracy of model in identifying credibility degree of news items with 93% f1-score and users with 79% f1-score. We also used the knowledge on fake news and fake news spreaders to understand information diffusion in social media in Chapter 9 and show that the proposed features obtained 97% AUROC score in predicting real and fake news sharing in social media. We also found that there is no single point of success or failure regarding the features in the diffusion of news in social media rather we observed that it is the dynamic interaction of news features, user features, and network features that produce successful diffusion.

Overall, this dissertation seeks a comprehensive understanding of false information and actors responsible for the creation and spread of false information on the web. We characterized and developed methods to accurately identify two types of false information and trustworthiness of users on various platforms. The analysis shows that each of our proposed methods outperforms the existing state-of-the-art methods in the detection of false information and actors in real-world opinion-based systems and social media.

10.2 Future Directions

There are several future research directions that will further help to address the challenging issues in the false information domain such as understanding and detection of other forms of false information including but not limited to rumors, satires, adversarial fake news, and adversarial fraudulent reviews. Further, while we posed that most of the existing research utilizing datasets labeled as fake or real is not sufficient and requires methods that address a problem of identifying the credibility degree of news items based on the amount of truthfulness in news content, a common and most prominent issue is a dearth of labeled data of news with various degrees of truthfulness. We will further describe each of these directions below.

- Adversarial attacks and detection methods.

In recent years, natural language generation and image processing fields have shown continuous signs of progress in the field of applied artificial intelligence. The researchers are developing general-purpose algorithms to automatically generate text and images using neural networks. For instance, the team from OpenAI developed GPT-2 [246] trained on the vast amount of text from the web to generate realistic content, text translations, question answering, and so on. However, one of the inevitable drawbacks includes malicious actors increasingly misusing such algorithms to generate controlled and realistic false information at speed and scale. Thus, it is crucial to take care of such AI-generated false information and develop methods to understand and detect such adversarial-generated false information.

- Creation of standard datasets and benchmarks.

Although several researches have been done to identify false information, most of them have pointed out the lack of benchmark datasets for proper analysis. Having benchmark datasets is as important as developing algorithms for false information detection. Especially, the research community lacks the dataset with varying degrees of the truthfulness of news content which poses challenges for researchers for a fair evaluation of their model performance. Therefore, future works in the false information domain should consider including benchmark datasets with different degrees of truthfulness rather than limiting them to real or fake. It would also be interesting to look at the performance of the existing state-of-the-art methods in identifying such multi-labeled false information.

REFERENCES

- [1] Benjamin D Horne and Sibel Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.
- [2] Lstm description. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- [3] Clpsych dataset.
- [4] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.
- [5] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.
- [6] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.
- [7] Ben Popken. Russian trolls went on attack during key election moments., 2018.
- [8] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profile for fake news detection. *arXiv preprint arXiv:1904.13355*, 2019.

- [9] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–18, 2020.
- [10] Atefeh Heydari, Mohammad ali Tavakoli, Naomie Salim, and Zahra Heydari. Detection of review spam: A survey. *Expert Systems with Applications*, 42(7):3634–3642, 2015.
- [11] Sungwoo Choi, Anna S Mattila, Hubert B Van Hoof, and Donna Quadri-Felitti. The role of power and incentives in inducing fake reviews in the tourism industry. *Journal of Travel Research*, 56(8):975–987, 2017.
- [12] Michael Luca. Reviews, reputation, and revenue: The case of yelp.com, 2013.
- [13] Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016.
- [14] Markus Schuckert, Xianwei Liu, and Rob Law. Insights into suspicious online ratings: direct evidence from tripadvisor. *Asia Pacific Journal of Tourism Research*, 21(3):259–272, 2016.
- [15] Andreas Munzel. Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus. *Journal of Retailing and Consumer Services*, 32:96–108, 2016.
- [16] Saba Salehi-Esfahani and Ahmet Bulent Ozturk. Negative reviews: Formation, spread, and halt of opportunistic behavior. *International Journal of Hospitality Management*, 74:138–146, 2018.

- [17] Elizabeth Dwoskin Craig Timberg. Washington post finds fake reviews on amazon, 2018.
- [18] Michael Barthel, Amy Mitchell, and Jesse Holcomb. Many americans believe fake news is sowing confusion. washington, dc: Pew research center, 2016.
- [19] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? *ACM SIGMETRICS Performance Evaluation Review*, 44(1):179–192, 2016.
- [20] Theodoros Lappas, Gaurav Sabnis, and Georgios Valkanas. The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research*, 27(4):940–961, 2016.
- [21] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.
- [22] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings*, 2014.
- [23] Marc Fisher, John Woodrow Cox, and Peter Hermann. Pizzagate: From rumor, to hashtag, to gunfire in dc. *Washington Post*, 6, 2016.
- [24] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736, 2013.

- [25] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics, 2011.
- [26] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.
- [27] Elizabeth Dwoskin and Craig Timberg. How merchants use facebook to flood amazon with fake reviews. *Washington Post*, 2018.
- [28] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *CoRR*, abs/1804.08559, 2018.
- [29] Xinyi Zhou and Reza Zafarani. Fake news: A survey of research, detection methods, and opportunities. *CoRR*, abs/1812.00315, 2018.
- [30] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsletter*, 19(1):22–36, 2017.
- [31] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *ACL*, pages 231–240, 2018.
- [32] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *ACL*, pages 3391–3401, 2018.

- [33] Dong ping Tian et al. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8(4):385–396, 2013.
- [34] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.
- [35] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433. IEEE, 2015.
- [36] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACMMM*, pages 795–816, 2017.
- [37] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *SIGKDD*, pages 849–857. ACM, 2018.
- [38] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230. ACM, 2008.
- [39] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. What yelp fake review filter might be doing? In *Seventh international AAAI conference on weblogs and social media*, 2013.
- [40] Karen Weise. A lie detector test for online reviewers. *Bloomberg Business Week*, 2011.

- [41] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 632–640. ACM, 2013.
- [42] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Exploiting burstiness in reviews for review spammer detection. In *Seventh international AAAI conference on weblogs and social media*, 2013.
- [43] Parisa Kaghazgaran, Majid Alfifi, and James Caverlee. Tomcat: Target-oriented crowd review attacks and countermeasures. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 302–312, 2019.
- [44] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM, 2010.
- [45] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 495–503. SIAM, 2016.
- [46] Craig Silverman. This analysis shows how viral fake election news stories outperformed real news on facebook. *BuzzFeed News*, 16, 2016.
- [47] Robin Burke, Michael O’Mahony, and Neil Hurley. Robust collaborative rec-

- ommendation. In *Recommender systems handbook*, pages 961–995. Springer, 2015.
- [48] Carlos E Seminario. Accuracy and robustness impacts of power user attacks on collaborative recommender systems. In *RecSys*, pages 447–450. ACM, 2013.
- [49] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S. Glance. What yelp fake review filter might be doing? In *ICWSM*, pages 409–418, 2013.
- [50] Julian McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *WWW*, pages 897–908, 2013.
- [51] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 8, 2018.
- [52] Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névél, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR Workshop Proceedings, September 2020.
- [53] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020.

- [54] Daria Plotkina, Andreas Munzel, and Jessie Pallud. Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research*, 109:511–523, 2020.
- [55] Huan Sun, Alex Morales, and Xifeng Yan. Synthetic review spamming and defense. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1088–1096, 2013.
- [56] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1143–1158, 2017.
- [57] Amy Mitchell, Jeffrey Gottfriedd, Michael Barthel, and Nami Sumida. Distinguishing between factual and opinion statements in the news. *Pew Research Center*, 2018.
- [58] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *WSDM*, pages 312–320, 2019.
- [59] Miriam J Metzger and Andrew J Flanagin. Psychological approaches to credibility assessment online. *The handbook of the psychology of communication technology*, 32:445–466, 2015.
- [60] Miriam J Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

- [61] James Cheo. Fake news can make - or break - stock prices., 2018.
- [62] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.
- [63] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.
- [64] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
- [65] Anu Shrestha and Francesca Spezzano. Online misinformation: from the deceiver to the victim. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining*, pages 847–850. ACM, 2019.
- [66] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439. ACM, 2019.
- [67] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019.
- [68] Anu Shrestha, Francesca Spezzano, and Abishai Joy. Detecting fake news spreaders in social networks via linguistic and personality features. In *CLEF*, 2020.

- [69] Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CLEF*, 2020.
- [70] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367. Springer, 2020.
- [71] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.
- [72] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.
- [73] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.
- [74] Yi Han, Shanika Karunasekera, and Christopher Leckie. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316*, 2020.
- [75] Anastasia Giachanou, Esteban A. Rissola, Bilal Ghanem, Fabio Crestani, and Paolo Rosso. The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In *Natural Language Processing and Information Systems: 25th International Conference on Applications of*

Natural Language to Information Systems, NLDB 2020, page 181. Springer Nature.

- [76] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.
- [77] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055, 2021.
- [78] Parisa Kaghazgaran, James Caverlee, and Anna Squicciarini. Combating crowd-sourced review manipulators: A neighborhood-based approach. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 306–314, 2018.
- [79] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994. ACM, 2015.
- [80] Guan Wang, Sihong Xie, Bing Liu, and S Yu Philip. Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining*, pages 1242–1247. IEEE, 2011.
- [81] Abhinav Mishra and Arnab Bhattacharya. Finding the bias and prestige of

- nodes in networks based on trust scores. In *Proceedings of the 20th international conference on World wide web*, pages 567–576. ACM, 2011.
- [82] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 221–230. IEEE, 2016.
- [83] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 119–130, 2013.
- [84] Daniel A. Effron and Medha Raj. Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share. *Psychological Science*, November 2019.
- [85] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3):21:1–21:42, April 2019.
- [86] Benjamin D. Horne, Dorit Nevo, John O’Donovan, Jin-Hee Cho, and Sibel Adali. Rating reliability and bias in news articles: Does AI assistance help everyone? In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019*, pages 247–256, 2019.
- [87] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, Jun 2010.

- [88] Sudipta Basu. The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics*, 24(1):3–37, December 1997.
- [89] Lawrence E. Boehm. The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, 20(3):285–293, 1994.
- [90] Dustin Carnahan and R Kelly Garrett. Processing style and responsiveness to corrective information. *International Journal of Public Opinion Research*, 32(3):530–546, 2020.
- [91] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, July 2019.
- [92] Maria Glenski, Corey Pennycuff, and Tim Weninger. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems*, 4(4):196–206, December 2017. Conference Name: IEEE Transactions on Computational Social Systems.
- [93] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What’s in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *Seventh International AAAI Conference on Weblogs and Social Media*, June 2013.
- [94] William J. Brady, Molly Crockett, and Jay Joseph Van Bavel. The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online. preprint, PsyArXiv, March 2019.

- [95] Bence Bago, David G. Rand, and Gordon Pennycook. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, January 2020.
- [96] Leticia Bode and Emily K Vraga. See something, say something: Correction of global health misinformation on social media. *Health communication*, 33(9):1131–1140, 2018.
- [97] Emily K Vraga and Leticia Bode. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645, 2017.
- [98] Bartosz W Wojdyski, Matthew T Binford, and Brittany N Jefferson. Looks real, or really fake? warnings, visual attention and detection of false news articles. *Open Information Science*, 3(1):166–180, 2019.
- [99] R Kelly Garrett and Shannon Poulsen. Flagging facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, 24(5):240–258, 2019.
- [100] Xunru Che, Danaë Metaxa-Kakavouli, and Jeffrey T. Hancock. Fake News in the News: An Analysis of Partisan Coverage of the Fake News Phenomenon. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '18*, pages 289–292, Jersey City, NJ, USA, October 2018. Association for Computing Machinery.
- [101] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining*, pages 354–367, 2020.

- [102] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, 2020.
- [103] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. “fake news” is not simply false information: a concept explication and taxonomy of online content. *American Behavioral Scientist*, page 0002764219878224, 2019.
- [104] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450. ACM, 2012.
- [105] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 603–612, Republic and Canton of Geneva, CHE, April 2018. International World Wide Web Conferences Steering Committee.
- [106] Ronald Robertson. Partisan Bias Scores for Web Domains, 2018.
- [107] Sam Wineburg and Sarah McGrew. Lateral reading: Reading less and learning more when evaluating digital information. 2017.

- [108] Juliet DNSc and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc, Los Angeles, fourth edition edition, December 2014.
- [109] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.
- [110] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [111] José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martinez, and Joaquín Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *ICPR*, volume 3, pages 314–317. IEEE, 2000.
- [112] Samara Klar. A multidimensional study of ideological preferences and priorities among the american public. *Public Opinion Quarterly*, 78(S1):344–359, 2014.
- [113] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545, July 2020. ISBN: 9781920498115 Publisher: National Academy of Sciences Section: Social Sciences.
- [114] Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier, 2015.

- [115] Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173, 2007.
- [116] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Understanding and reducing the spread of misinformation online, Nov 2019.
- [117] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. Combating Fake News: An Agenda for Research and Action. Technical report, Harvard & Northeastern University, May 2017.
- [118] Luciana Carraro, Luigi Castelli, and Claudia Macchiella. The Automatic Conservative: Ideology-Based Attentional Asymmetries in the Processing of Valenced Information. *PLOS ONE*, 6(11):e26456, November 2011.
- [119] Luigi Castelli and Luciana Carraro. Ideology is related to basic cognitive processes involved in attitude formation. *Journal of Experimental Social Psychology*, 47(5):1013–1016, September 2011.
- [120] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [121] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1):eaau4586, January 2019.

- [122] Pew Research Center. Demographics of Social Media Users and Adoption in the United States, June 2019.
- [123] Paul-Alexandru Chirita, Wolfgang Nejdl, and Cristian Zamfir. Preventing shilling attacks in online recommender systems. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 67–74, 2005.
- [124] Mingdan Si and Qingshan Li. Shilling attacks against collaborative recommender systems: a review. *Artificial Intelligence Review*, 53(1):291–319, 2020.
- [125] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.
- [126] Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2008.
- [127] Anu Shrestha, Francesca Spezzano, and Maria Soledad Pera. Who is really affected by fraudulent reviews? an analysis of shilling attacks on recommender systems in real-world scenarios. *arXiv preprint arXiv:1808.07025*, 2018.
- [128] Agnideven Palanisamy Sundar, Feng Li, Xukai Zou, Tianchong Gao, and Evan D Russomanno. Understanding shilling attacks and their detection traits: a comprehensive survey. *IEEE Access*, 8:171703–171715, 2020.
- [129] Michael P O’Mahony, Neil J Hurley, and Guérolé CM Silvestre. Recommender systems: Attack types and strategies. In *AAAI*, pages 334–339, 2005.

- [130] Minghong Fang, Neil Zhenqiang Gong, and Jia Liu. Influence function based data poisoning attacks to top-n recommender systems. In *Proceedings of The Web Conference 2020*, pages 3019–3025, 2020.
- [131] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and V. S. Subrahmanian. REV2: fraudulent user prediction in rating platforms. In *WSDM*, pages 333–341. ACM, 2018.
- [132] Edoardo Serra, Anu Shrestha, Francesca Spezzano, and Anna Cinzia Squicciarini. Deeptrust: An automatic framework to detect trustworthy users in opinion-based systems. In *CODASPY*, pages 29–38. ACM, 2020.
- [133] Shyong K Lam and John Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, pages 393–402. ACM, 2004.
- [134] Carlos E Seminario and David C Wilson. Attacking item-based recommender systems with power items. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 57–64, 2014.
- [135] Carlos E Seminario and David C Wilson. Nuke’em till they go: Investigating power user attacks to disparage items in collaborative recommenders. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 293–296, 2015.
- [136] Soumya Wadhwa, Saurabh Agrawal, Harsh Chaudhari, Deepthi Sharma, and Kannan Achan. Data poisoning attacks against differentially private recommender systems. In *Proceedings of the 43rd International ACM SIGIR Confer-*

- ence on Research and Development in Information Retrieval*, pages 1617–1620, 2020.
- [137] Chen Lin, Si Chen, Hui Li, Yanghua Xiao, Lianyun Li, and Qian Yang. Attacking recommender systems with augmented user profiles. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 855–864, 2020.
- [138] Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. How dataset characteristics affect the robustness of collaborative recommendation models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 951–960, 2020.
- [139] Karen Weise. A lie detector test for online reviewers. In *Bloomberg BusinessWeek*. 2011.
- [140] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 985–994, 2015.
- [141] Santosh K. C. and Arjun Mukherjee. On the temporal dynamics of opinion spamming: Case studies on yelp. In *WWW*, pages 369–379, 2016.
- [142] Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Cinzia Squicciarini. Uncovering crowdsourced manipulation of online reviews. In *Proceedings of the*

- 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 233–242, 2015.
- [143] Michael D Ekstrand. Lenskit for python: Next-generation software for recommender systems experiments. In *CIKM*, pages 2999–3006, 2020.
- [144] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.
- [145] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *IEEE ICDM*, pages 263–272, 2008.
- [146] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- [147] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *AAIM*, pages 337–348. Springer, 2008.
- [148] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.
- [149] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1235–1244, 2015.

- [150] Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop*, volume 2007, pages 5–8, 2007.
- [151] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. Leave no user behind: Towards improving the utility of recommender systems for non-mainstream users. In *WSDM*, pages 103–111. ACM, 2021.
- [152] Srijan Kumar, Francesca Spezzano, V. S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *IEEE ICDM*, pages 221–230, 2016.
- [153] Shyong K. Lam and John Riedl. Shilling recommender systems for fun and profit. In *WWW*, pages 393–402, 2004.
- [154] Jennifer Golbeck, James Hendler, et al. Filmtrust: Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer communications and networking conference*, volume 96, pages 282–286, 2006.
- [155] Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. The eigen-trust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651, 2003.
- [156] Abhinav Mishra and Arnab Bhattacharya. Finding the bias and prestige of nodes in networks based on trust scores. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 567–576, 2011.

- [157] Srijan Kumar, Francesca Spezzano, V. S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 221–230, 2016.
- [158] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and V. S. Subrahmanian. REV2: fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 333–341, 2018.
- [159] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. BIRDNEST: bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pages 495–503, 2016.
- [160] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. Identify online store review spammers via social review graph. *ACM TIST*, 3(4):61:1–61:21, 2012.
- [161] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*, 2013.
- [162] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malú Castellanos, and Ridhiman Ghosh. Exploiting burstiness in reviews for review spammer detection.

- In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*, 2013.
- [163] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. Spotting opinion spammers using behavioral footprints. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 632–640, 2013.
- [164] Parisa Kaghazgaran, James Caverlee, and Anna Cinzia Squicciarini. Combating crowdsourced review manipulators: A neighborhood-based approach. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 306–314, 2018.
- [165] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 939–948, 2010.
- [166] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [167] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

- [168] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *International semantic Web conference*, pages 351–368. Springer, 2003.
- [169] Sini Ruohomaa and Lea Kutvonen. Trust management survey. In *International Conference on Trust Management*, pages 77–92. Springer, 2005.
- [170] Giorgos Zacharia and Pattie Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–907, 2000.
- [171] Shearer and Gottfried. News use across social media platforms 2017, 2017.
- [172] Zimdars. False, misleading, clickbait-y, and satirical news sources. available at. 2016.
- [173] Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164, 2009.
- [174] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.
- [175] S Bird. Nltk: The natural language toolkit steven. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72, 2006.
- [176] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

- [177] CHE Gilbert and Erric Hutto. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, volume 81, page 82, 2014.
- [178] Saif Mohammad. Word Affect Intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018.
- [179] Ashlee Milton, Leveeson Batista, Garrett Allen, Siqu Gao, Yiu-Kai Ng, and Maria Soledad Pera. "don't judge a book by its cover": Exploring book traits children favor. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pages 669–674. ACM, 2020.
- [180] Thomas T Hills. The dark side of information proliferation. *Perspectives on Psychological Science*, 14(3):323–330, 2019.
- [181] Anu Shrestha, Francesca Spezzano, and Indhumathi Gurunathan. Multi-modal analysis of misleading political news. In *Disinformation in Open Online Media - Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26-27, 2020, Proceedings*, volume 12259 of *Lecture Notes in Computer Science*, pages 261–276. Springer, 2020.
- [182] Elisa Shearer and Amy Mitchell. News use across social media platforms in 2020. *Pew Research Center*, 2020.
- [183] Pamela J Shoemaker, Tim P Vos, and Stephen D Reese. Journalists as gatekeep-

- ers. In Karin Wahl-Jorgensen and Thomas Hanitzsch, editors, *The handbook of journalism studies*, volume 73, page 15. 2009.
- [184] Axel Bruns, Tim Highfield, and Rebecca Ann Lind. Blogs, twitter, and breaking news: The produsage of citizen journalism. *Producing theory in a digital world: The intersection of audiences and production in contemporary theory*, 80(2012):15–32, 2012.
- [185] Lauren Vogel. Viral misinformation threatens public health, 2017.
- [186] Emma S Spiro, Sean Fitzhugh, Jeannette Sutton, Nicole Pierski, Matt Greczek, and Carter T Butts. Rumoring during extreme events: A case study of deep-water horizon 2010. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 275–283, 2012.
- [187] Jim Isaak and Mina J Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.
- [188] Edelman Trust Barometer. Edelman trust barometer global report. *Edelman*, available at: https://www.edelman.com/sites/g/files/aatuss191/files/2019-02/2019_Edelman_Trust_Barometer_Global_Report_2.pdf, 2019.
- [189] Francesca Spezzano, Anu Shrestha, Jerry Alan Fails, and Brian W. Stone. That’s fake news! investigating how readers identify the reliability of news when provided title, image, source bias, and full articles. *Proceedings of the ACM on Human Computer Interaction journal*, 5(CSCW1, Article 109), 2021.
- [190] Long Ma, Chei Sian Lee, and Dion H Goh. Understanding news sharing in social media from the diffusion of innovations perspective. In *2013 IEEE In-*

- ternational Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pages 1013–1020. IEEE, 2013.
- [191] Chei Sian Lee and Long Ma. News sharing in social media: The effect of gratifications and prior experience. *Computers in human behavior*, 28(2):331–339, 2012.
- [192] Yarimar Bonilla and Jonathan Rosa. # ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the united states. *American ethnologist*, 42(1):4–17, 2015.
- [193] Regina Rini. Fake news and partisan epistemology. *Kennedy Institute of Ethics Journal*, 27(2):E–43, 2017.
- [194] Anu Shrestha and Francesca Spezzano. Textual characteristics of news title and body to detect fake news: A reproducibility study. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 120–133. Springer, 2021.
- [195] Julio CS Reis, Haewoon Kwak, Jisun An, Johnnatan Messias, and Fabricio Benevenuto. Demographics of news sharing in the us twittersphere. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 195–204, 2017.
- [196] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic inference and rep-

- representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067, 2019.
- [197] Libby Hemphill, Aron Culotta, and Matthew Heston. # polar scores: Measuring partisanship using social media content. *Journal of Information Technology & Politics*, 13(4):365–377, 2016.
- [198] Joshua M Chamberlain, Francesca Spezzano, Jaclyn J Kettler, and Bogdan Dit. A network analysis of twitter interactions by members of the us congress. *ACM Transactions on Social Computing*, 4(1):1–22, 2021.
- [199] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *ICWSM*, 13:1–10, 2013.
- [200] Anu Shrestha, Edoardo Serra, and Francesca Spezzano. Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. *NetMAHIB*, 9(1):22, 2020.
- [201] Ashlee Milton and Maria Soledad Pera. What snippets feel: Depression, search, and snippets. In Iván Cantador, Max Chevalier, Massimo Melucci, and Josiane Mothe, editors, *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020)*, Samatan, Gers, France, July 6-9, 2020, volume 2621 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.
- [202] Wei Wang, Ivan Hernandez, Daniel A Newman, Jibo He, and Jiang Bian. Twitter analysis: Studying us weekly trends in work stress and emotion. *Applied Psychology*, 65(2):355–378, 2016.

- [203] Yair Neuman. *Computational personality analysis: Introduction, practical applications and novel directions*. Springer, 2016.
- [204] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository*. Retrieved January, 8:2018, 2016.
- [205] Common Sense Media. News and america's kids: How young people perceive and are impacted by the news, 2017.
- [206] Jeffrey Gottfriedd and Elizabeth Grieco. Younger americans are better than older americans at telling factual news statements from opinions. *Pew Research Center*, 2018.
- [207] Christine Benesch. An empirical analysis of the gender gap in news consumption. *Journal of Media Economics*, 25(3):147–167, 2012.
- [208] Paula Poindexter and Dustin Harp. The softer side of news. In Paula Poindexter, Sharon Meraz, and Amy Schmitz Weiss, editors, *Women, men and news: Divided and disconnected in the news media landscape*, pages 85–96. Routledge New York, 2008.
- [209] Hyun Jung Oh, Elif Ozkaya, and Robert LaRose. How does online social networking enhance life satisfaction? the relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Computers in Human Behavior*, 30:69–78, 2014.
- [210] Jeffrey A Hall, Natalie Pennington, and Allyn Lueders. Impression management

- and formation on facebook: A lens model approach. *New Media & Society*, 16(6):958–982, 2014.
- [211] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In William W. Cohen and Andrew W. Moore, editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML, volume 148 of ACM International Conference Proceeding Series*, pages 233–240. ACM, 2006.
- [212] Jakab Buda and Flora Bolonyai. An ensemble model using n-grams and statistical features to identify fake news spreaders on twitter. In *CLEF*, 2020.
- [213] Juan Pizarro. Using n-grams to detect fake news spreaders on twitter. In *CLEF*, 2020.
- [214] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [215] Shearer Elisa Mitchell, Amy and Galen Stocking. News on twitter: Consumed by most users and trusted by many. *Pew Research Center*, 2021.
- [216] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.
- [217] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by

- exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [218] Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Harald Schoen, Peter Gloor, Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 2013.
- [219] Anu Shrestha and Francesca Spezzano. Textual characteristics of news title and body to detect fake news: A reproducibility study. In *ECIR, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 120–133. Springer, 2021.
- [220] Anu Shrestha and Francesca Spezzano. Characterizing and predicting fake news spreaders in social networks. *International Journal of Data Science and Analytics*, pages 1–14, 2021.
- [221] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.
- [222] Yunfei Long. Fake news detection through multi-perspective speaker profiles. Association for Computational Linguistics, 2017.
- [223] Kai Shu, H Russell Bernard, and Huan Liu. Studying fake news via network analysis: detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, pages 43–65. Springer, 2019.

- [224] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [225] Mediabias/factcheck apps/extensions.
- [226] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. Deep graph library: Towards efficient and scalable deep learning on graphs. 2019.
- [227] Ross P Kindermann and J Laurie Snell. On the relation between markov random fields and social networks. *Journal of Mathematical Sociology*, 7(1):1–13, 1980.
- [228] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.
- [229] Paulo Shakarian, Abhivav Bhatnagar, Ashkan Aleali, Elham Shaabani, Ruocheng Guo, et al. *Diffusion in social networks*. Springer, 2015.
- [230] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society*, 1(2):2056305115610141, 2015.
- [231] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.
- [232] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, pages 137–146, 2003.
- [233] Maryam Maleki, Esther Mead, Mohammad Arani, and Nitin Agarwal. Using an

- epidemiological model to study the spread of misinformation during the black lives matter movement. *arXiv preprint arXiv:2103.12191*, 2021.
- [234] Taichi Murayama, Shoko Wakamiya, Eiji Aramaki, and Ryota Kobayashi. Modeling the spread of fake news on twitter. *Plos one*, 16(4).
- [235] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.
- [236] *A Sociology of Monsters: Essays on Power, Technology, and Domination*. Sociological Review Monograph. Routledge, 1991.
- [237] Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE, 2018.
- [238] Saif M Mohammad. Word affect intensities. *arXiv preprint arXiv:1704.08798*, 2017.
- [239] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.
- [240] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.
- [241] Olaf Zorzi. Granovetter (1983): The strength of weak ties: A network theory revisited. In *Schlüsselwerke der Netzwerkforschung*, pages 243–246. Springer, 2019.

- [242] G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.
- [243] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1852–1872, 2018.
- [244] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *ACM WSDM*, pages 241–250, 2010.
- [245] W. van Osch and C. Coursaris. Social media research: An assessment of the domain’s productivity and intellectual evolution. *Communication Monographs*, 81(3):285–309, 2014.
- [246] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [247] Pen America. Faking news: Fraudulent news and the fight for truth, 2018.
- [248] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3):598–608, 2017.
- [249] William Yang Wang. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *ACL*, pages 422–426, 2017.
- [250] Tanushree Mitra and Eric Gilbert. CREDBANK: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267, 2015.

- [251] Symeon Papadopoulos Duc-Tien Dang-Nguyen Giulia Boato Michael Riegler Yiannis Kompatsiaris et al. Christina Boididou, Katerina Andreadou. Verifying multimedia use at mediaeval 2015. In *MediaEval*, 2015.
- [252] Chun-Nan Chou, Chuen-Kai Shie, Fu-Chieh Chang, Jocelyn Chang, and Edward Y Chang. Representation learning on large and small data. *arXiv:1707.09873*, 2017.
- [253] David Barton. *Literacy: An introduction to the ecology of written language*. John Wiley & Sons, 2017.
- [254] James P. Fairbanks, Nathan Knauf, and Erica Briscoe Georgia. Credibility assessment in the news : Do we need to read? In *MIS2: Misinformation and Misbehavior Mining on the Web Workshop*, 2018.
- [255] Elisa Shearer and Katerina Eva Matsa. News use across social media platforms 2018. *Pew Research Center*, 2018.
- [256] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.
- [257] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*, pages 1–9, 2013.
- [258] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro

- Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.
- [259] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1), 2019.
- [260] Kaisong Song, Wei Gao, Ling Chen, Shi Feng, Daling Wang, and Chengqi Zhang. Build emotion lexicon from the mood of crowd via topic-assisted joint non-negative matrix factorization. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016*, pages 773–776. ACM, 2016.
- [261] Jonah Berger. Arousal increases social transmission of information. *Psychological science*, 22(7):891–893, 2011.
- [262] Arman Cohan, Sydney Young, and Nazli Goharian. Triaging mental health forum posts. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 143–147, 2016.
- [263] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*, 2017.
- [264] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107, 2015.

- [265] Minsu Park, Chiyong Cha, and Meeyoung Cha. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, volume 2012, pages 1–8. ACM New York, NY, 2012.
- [266] Minsu Park, David W McDonald, and Meeyoung Cha. Perception differences between the depressed and non-depressed users in twitter. *ICWSM*, 9:217–226, 2013.
- [267] Ronghua Xu and Qingpeng Zhang. Understanding online health groups for depression: social network and linguistic perspectives. *Journal of medical Internet research*, 18(3), 2016.
- [268] Johannes Zimmermann, Timo Brockmeyer, Matthias Hunn, Henning Schauenburg, and Markus Wolf. First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical psychology & psychotherapy*, 24(2):384–391, 2017.
- [269] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.
- [270] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.
- [271] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as

- a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.
- [272] Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych@NAACL-HTL, New Orleans, LA, USA, June 2018*, pages 168–173, 2018.
- [273] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.
- [274] Shervin Malmasi, Marcos Zampieri, and Mark Dras. Predicting post severity in mental health forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 133–137, 2016.
- [275] Jacopo Staiano and Marco Guerini. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.
- [276] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.
- [277] Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cécile Paris. Data61-csiro systems at the clpsych 2016 shared task. In *Proceedings of the 3rd Workshop*

- on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 128–132, 2016.
- [278] Shervin Malmasi, Marcos Zampieri, and Mark Dras. Predicting post severity in mental health forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 133–137, 2016.
- [279] Chris Brew. Classifying reachout posts with a radial basis function SVM. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 138–142, 2016.
- [280] Doug Millen. Reachout annual report 2013/2014, 2015.
- [281] Atari Metcalf and Victoria Blake. Reachout. com annual user survey results, 2013.
- [282] Lisa Lustberg and Charles F Reynolds III. Depression and insomnia: questions of cause and effect. *Sleep medicine reviews*, 4(3):253–262, 2000.
- [283] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.
- [284] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *SIGKDD*, pages 855–864, 2016.

- [285] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.
- [286] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F Cooper, and Gilles Clermont. Outlier detection for patient monitoring and alerting. *Journal of biomedical informatics*, 46(1):47–55, 2013.
- [287] Jiawei Han and Micheline Kamber. *Data mining: Concepts and techniques*. 2012.
- [288] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [289] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- [290] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.
- [291] Anu Shrestha and Francesca spezzano. Detecting depressed users in online forums. In *International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (HI-BI-BI 2019), in conjunction with ASONAM'19*, pages –, 2019.

APPENDIX A:

QUALITATIVE CODES

This appendix consists primarily of Table A.1 below that identifies the qualitative codes inductively generated by three reviewers. It also includes the explanation/description of the codes and some examples.

Table A.1: Thematic codes (and examples) that were inductively developed by analyzing the question “why did you identify the news item as real or fake”. The table identifies the code, provides a description, and example quotes [with the associated accuracy]. The accuracy is indicated as: FN = false negative (identified fake news as real); FP = false positive (identified real news as fake); TN = true negative (identified real news as real); TP = true positive (identified fake news as fake).

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Plausible	Believable; possible; reasonable; could happen	<ul style="list-style-type: none"> • “This is said about many politicians” [FN] • “I haven’t heard about this but it seems real enough that it could be believable” [TN] • “It seems like something our former president could have said” [FN] • “I think that article is real because this seems like something that could definitely happen” [TN] • “Nothing pops out at me in a way that would make me think it’s fake. All the information seems like it could be legitimate” [FN]

Continued on next page

Continued from previous page

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Implausible	Unlikely; unbelievable; unrealistic	<ul style="list-style-type: none"> • “Seems [...] like something to be seen on Instagram” [FP] • “This seems unrealistic and like one of those click bate stories” [FP] • “I’m not sure Trump is very articulate and able to pull off a speech like this” [FP] • “Seemed like someone made this up and didn’t really seem like an actual article” [FP]
Familiar	Heard about it; sense of recognition; seen similar stories	<ul style="list-style-type: none"> • “I saw something like this on NBC news at some point” [TN] • “The content of the article seems vaguely familiar” [FN] • “I feel like I heard this somewhere but not sure” [FN] • “I’ve heard similar stories” [TN]
Unfamiliar	Not recognized; haven’t heard of it	<ul style="list-style-type: none"> • “I haven’t heard about it before” [TP] • “I am not familiar with the material” [TP] • “I don’t remember this much discussion back and forth during the last election on the subject” [FP]
Title Professional	Proper grammar and punctuation; has quotes or statistics; lacks obvious bias	<ul style="list-style-type: none"> • “Because it uses quotes and statistics in their title” [FN] • “I think it is real because this title has a lot of contexts no bias and it’s asking a question” [TN] • “I said this was real because it doesn’t seem to have a political title” [TN] • “Used a quote” [FN]

Continued on next page

Continued from previous page

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Title Unprofessional	Poor grammar; all caps or exclamations; clickbait; title emotional (see below)	<ul style="list-style-type: none"> ● “Unprofessional language” [FP] ● “The title contains one bolded word which you would not normally see” [TP] ● “The title seemed like click bate” [FP] ● “The word lead being in all caps is sketchy to me as well as the title being so uninformative” [TN]
Title Emotional	Fear-mongering; aggressive; alarmism; evocative	<ul style="list-style-type: none"> ● “The title is a sneak diss” [TP] ● “This sounds too dramatic to be a real news story” [FP] ● “I think this is fake because the title is trying to evoke people that she is a liar and not being truthful” [TP]
Picture Professional	Neutral image; direct image of the story/event; realistic image (e.g., video still)	<ul style="list-style-type: none"> ● “Picture looks to be taken from a security camera” [TN] ● “The picture is video evidence of it actually happening” [TN] ● “The photo used matches the topic well and seems to be of high quality” [TN]
Picture Unprofessional	Photoshop; meme; poor image quality or editing; picture emotional (see below)	<ul style="list-style-type: none"> ● “The picture looks like a meme” [FP] ● “Pictures look fake and photoshopped” [TP] ● “The text has nothing out of the ordinary. But, the picture makes Obama look like he’s a dictator or something” [TP] ● “The photo uses photoshop which typically is not used within a news article” [TP] ● “The image is throwing me off and looks badly edited” [TP]

Continued on next page

Continued from previous page

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Picture Emotional	Negative; aggressive; fear-mongering; biased selection of image	<ul style="list-style-type: none"> ● “It is using an aggressive picture to try and get you to feel a certain way about him” [FP] ● “The photo associated looks to be manipulative and biased suggesting the article’s intentions are emotionally skewed and attempting to insight fear and confusion in the political environment” [TP]
Text Professional	Neutral/calm tone; uses statistics; includes quotes, citations, and/or sources; good grammar or punctuation; detailed; focuses on events rather than opinion and editorial	<ul style="list-style-type: none"> ● “No punctuation errors” [TN] ● “This article has facts and quotes that lead me to believe this is real news. It is also professional written and speaks on more facts than opinions” [FN] ● “The above news is real because the diction is neutral. In addition, in the end it allows the readers to see the documents as proof that this had really happened” [FN] ● “They use facts and statistics that seems to be legit” [FN] ● “With the direct quotes and details that this excerpt used, I have the feeling that it is real” [FN] ● “It has a calmer tone which makes me think that it is real” [FN] ● “Provides clear background information and location and sources” [TN]

Continued on next page

Continued from previous page

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Text Unprofessional	Emotional, fear-mongering, or attacking tone; heavy on opinion and editorial; poor grammar and punctuation	<ul style="list-style-type: none"> ● “This article is clearly biased and persuasive, with only using pathos as its rhetoric” [TP] ● “The wording seems entirely exaggerated, like an attack” [TP] ● “The content is good, but the writing and editing is super unprofessional. It’s very opinionated, so it’s fake news” [TP] ● “This is fake because some of the English in this article is not as professional as one found on a reliable source” [FP] ● “It seems like this article is used for propoganda because in the end they advertise for Ted Cruz” [TP]
Source Untrusted or Unfamiliar	Skepticism about where the article itself is found (e.g., Facebook) or sources cited within it (e.g., “as reported by...”); skepticism about the type of media that did the reporting	<ul style="list-style-type: none"> ● “It states that Hillary’s website is HillaryClinton.com. Which doesn’t sound credible at all” [FP] ● “The source comes straight from Facebook so new stories can easily be changed and published by anyone” [TP] ● “This is believable, but I do not know the source” [FP] ● “Seems to be a radicalized right wing website” [TP] ● “If it is on live television don’t trust the news to quote someone the right way” [FP]

Continued on next page

Continued from previous page

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Perception of Person	Bases reasoning on strong opinions the person who is the subject of the news item; mentions trust or lack of trust in an individual person	<ul style="list-style-type: none"> • “Unlikely Donald Trump, we can trust what Obama told us is actually real news” [FN] • “I believe this to be true because Hillary has been caught in a lot of issues so this doesn’t surprise me that this is something she would do” [FN] • “It just does not sound like the President. He cites unity and I feel that he is not for unity to the public” [FP] • “Don’t believe it was the Illuminati or anything. I do believe Obama worked for the world not the USA. Reasoning is he gave USA enemies nuclear rights and a pallet full of cash. Why? Would you like ever provide your enemy” [FN]
Pre-existing Beliefs	Bases reasoning on information/beliefs not in the article itself (and not specifically about a person in the news item)	<ul style="list-style-type: none"> • “I feel like this isn’t too far from the beliefs of the democrats” [TN] • “This article doesn’t seem to be real or logical in my eyes because of how important the internet is in the lives of any U.S. citizen” [TP] • “Military news is usually pretty real” [FN]
Missing Information	Not enough evidence to consider it true; not enough evidence to consider it false; lack of sources, quotes, statistics or details	<ul style="list-style-type: none"> • “I don’t have enough information to give a real or fake” [FP] • “Not enough information” [FP] • “It’s difficult to believe something when there is zero explanation behind it” [TP] • “Doesn’t have any solid evidence that this is true” [FP] • “Probably is true but not backed by solid evidence” [TP]

Continued on next page

Continued from previous page

Code	Explanation of Code	Example Quotes [and associated accuracy of that case]
Guess / Unsure	Explicitly mentions not being sure or making a guess	<ul style="list-style-type: none"><li data-bbox="686 499 1383 569">• “This could be fake or real. There is no way for me to gauge its fakeness” [TN]

APPENDIX B:

MULTI-MODAL ANALYSIS OF MISLEADING POLITICAL NEWS

The internet is a valuable resource to openly share information or opinions. Unfortunately, such internet openness has also made it increasingly easy to abuse these platforms through the dissemination of misinformation. As people are generally awash in information, they can sometimes have difficulty discerning misinformation propagated on these web platforms from truthful information. They may also lean too heavily on information providers or social media platforms to curate information even though such providers do not commonly validate sources. In this chapter, we focus on political news and present an analysis of misleading news according to different modalities, including news content (headline, body, and associated image) and source bias. Our findings show that hyperpartisan news sources are more likely to spread misleading stories than other sources and that it is not necessary to read news body content to assess its validity, but considering other modalities such as headlines, visual content, and publisher bias can achieve better performances.

B.1 Introduction

The volume of misleading news present in current media has grown in popularity in recent years through social media and online news sources. In 2017, the Pew Research

Center found that 67% of American adults (ages 18+) get news from social media, which was a 5% increase since 2016 [171]. An analysis of news leading up to the 2016 election conducted by BuzzFeed, found that there was more engagement with the leading misleading news stories than real news stories [46]. News is becoming more accessible and widespread than ever before. However, information proliferation has also contributed to the spread of misleading news, which has fostered the advancement of various methods to determine the validity of news. One such method is developed upon evaluating linguistic attributes such as features determining readability and lexical information [32, 1, 31]. These methods often mimic that of what would generally be considered the most effective of all: reading through the news with the purpose of evaluating their accuracy. However, with the spread of misleading news, it is unlikely, if not impossible, for everyone to spend large quantities of time reading through multiple newspapers and sources. Of course, the news sharing process occurs rapidly, necessitating effective methods to recognize signals of misleading content. In fact, reading the news body content may be time-consuming, and often people are exposed to news through their snippet on social media, where only the news headline and images are shown.¹ This trend of showing only some flimsy cuts of news with catchy headline and visuals in social media news feeds has made people share such news frequently without having deep reading and monitoring. A recent study by Gabielkov et al. [19] found evidence that the number of news shares is an inaccurate measure of actual readership. Thus, people are immersed in information across social media, which is often shared without reading and validating the content, thus leading to possible consequences of its diffusion.

¹There are also some browser extensions that checks the source and further add the publisher bias to the news appearing in the social media feed [225].

In this chapter, we use machine learning and multi-modal content analysis to detect misleading political news. To the best of our knowledge, we present the *first* content-based study considering the headline, body content, visual, and source bias modalities together for misleading news detection. Because the news trends continuously evolve, we analyze news text (from body and headline) by focusing on linguistic style, text complexity, and psychological aspects of the text, rather than topic-dependent representations of documents (e.g., [109]). Moreover, we consider *new features* that have not been explored before such as to capture emotions in images and the political bias of the news publisher. Our analysis, conducted on two state-of-the-art political news datasets, namely FakeNewsNet [58] and BuzzFeed-News [31], reveals that:

- News headlines are more informative than news body content, suggesting that we can avoid to “read” the news excerpt and focus on other modalities to better detect misleading news.
- By comparing news headline and excerpt content, we observe that headline characteristics are more consistent than excerpt ones across datasets (e.g., punctuation features are the most important group of features in both datasets considered), and, in general, the headline focuses more on briefly drawing the attention of the reader, while a higher number of emotional/ psychological words is more a characteristic of an excerpt than the headline, for misleading news.
- Publisher bias is a strong predictor of news validity. In fact, by analyzing information collected from mediabiasfactcheck.com (“the most comprehensive media bias resource on the Internet”), we show that hyper-partisan news sources are more likely to spread misleading stories than other sources.

- Image features improve the automatic detection of misleading news with the most important features being the ones highlighting the expressions and emotions of depicted people.

- It is possible to detect misleading news from its snippet (news headline, image, and source bias) more accurately than looking into the body content: AUROC 0.91 vs. 0.78 on FakeNewsNet and 0.81 vs. 0.77 on BuzzFeedNews.

Overall, this chapter contributes to determining effective and explicable multi-modal factors to recognize misleading news, that can be taught to people to recognize misleading news from its snippet and possibly decrease the unconscious spread of misinformation in social media [247].

B.2 Related Work

To detect misleading news, many works have considered news content (headline, body, image), the social network between the users and their social engagement (share, comment, and discuss given news), or a hybrid approach that considers both [30]. Regarding misleading news detection from news content (which is the focus of this chapter), Potthast et. al [31] attempted to classify news as real or fake based on its style as being part of hyperpartisan news, mainstream news, or satire. This study used a dataset composed of 1,627 articles from a Buzzfeed dataset. Features such as n-grams, stop words, parts of speech, and readability were considered in this study. Although there was higher F1-measure in determining the hyperpartisan vs. mainstream articles (0.78 F1-measure based on stylistic features and 0.74 for topic) the research was limited in deciphering between fake and real news (0.46 F1-measure for style-based features).

Horne and Adali [1] considered both news body and headline for determining the

validity of news. They included three datasets: a dataset created by BuzzFeed leading to the 2016 U.S. elections, one created by the researchers containing real, fake and satire sources, and a third dataset containing real and satire articles from a previous study. Based on textual features extracted from body and headline, they found out that the content of fake and real news is drastically different as they were able to obtain a 0.71 accuracy when considering the number of nouns, lexical redundancy (TTR), word count, and the number of quotes. Further, the study found that fake titles contain different sorts of words (stop words, extremely positive words, and slang, among others) than titles of real news articles resulting in a 0.78 accuracy. Pérez-Rosas et al [32] collected two new datasets, the FakeNewsATM dataset covering seven different news domains (education, business, sport, politics, etc.) and the Celebrity dataset regarding news on celebrities. They analyzed the news body content only and achieved an F1-measure up to 0.76 in detecting misleading content. They also tested cross-domain classification obtaining poor performances by training in one dataset and testing in the other one, but better accuracies (ranging from 0.51 to 0.91) in training on all but the test domain in the FakeNewsATM dataset.

Images in news articles also play a role in misleading news detection [33, 34, 35, 36]. Fake images are used in news articles to provoke emotional responses from readers. Images are the most eye-catching type of content in the news; a reader can be convinced of a claim by just looking at the title of the news and the image itself. So, it's crucial to include image analysis in fake news detection techniques. For instance, Jin et al. [248] showed that including visual and statistical features extracted from news images improves the results for microblogs news verification up to an F1-measure of 0.83 on a dataset collected from Sina Weibo on general news events and

associated images. Wang et al. [37] proposed a deep-learning-based framework to extract features from both text and image of the tweets about news not related to specific events to detect misleading content. Results show an F1-measure ranging from 0.72 on Twitter to 0.83 on Sina Weibo.

In contrast with previous work, this chapter provides a comprehensive study of four different content-based modalities to detect misleading political news. Other works have considered a single modality (e.g., either body content or images) or a subset of the modalities we considered (e.g., headline and body, or body, and image) but all these modalities together have not been investigated so far. Also, work involving image analysis [248, 37] focused on micro-blog content rather than proper news content.

B.3 Datasets

In this section, we discuss the lack of a large scale misleading news dataset (especially in the political domain) and present the datasets we use in this chapter, including a *new* dataset containing publisher bias and credibility we crawled from the Media-Bias/FactCheck website.

Available Datasets and Limitations

There exist several datasets containing political news that have been used for fake news detection, as shown in Table B.1.

Horne and Adali used two datasets in their paper [1]. The first dataset, DS1, contains 36 real news stories and 35 fake news stories, while the second one, DS2, contains 75 real, misleading, and satire news (75 for each category). The main drawback of these two datasets is that labels are assigned according to the credibility of the news source, instead of via fact-checking. However, a news source can have

Table B.1: Available datasets for misleading news detection.

Dataset	Size	Text	Images
BuzzFeedNews [31]	1,627	✓	
Horne and Adali DS1 [1]	71	✓	
Horne and Adali DS2 [1]	225	✓	
Pérez-Rosas et al [32]	480	✓	
FakeNewsNet [58]	384	✓	✓

mixed credibility and publish both factual and misleading information. Pérez-Rosas et al [32] collected a dataset of 480 news where 240 are fact-checked real news belonging to six different domains (sports, business, politics, etc.) and 240 are fake news collected via crowdsourcing, i.e., they asked AMT workers to write a fake news item based on one of their real news item and by mimic journalist style (hence these are unrealistic news articles). In this chapter, we use two datasets (described later in the section) to conduct our analysis, namely FakeNewsNet [58] and BuzzFeedNews [31] (the largest available dataset). FakeNewsNet is the only state-of-the-art dataset containing information beyond the news content modality and in the political domain.

As Table B.1 shows, there is generally limited availability of large scale benchmarks for fake news detection as collecting labels requires fact-checking, which is a time-consuming activity. As reported in [30], other datasets have been used for related tasks, but they are not suitable for our analysis as they do not contain proper news articles. For instance, LIAR [249] contains human-labeled short statements, while CREDBANK [250] contains news events, where each event is a collection of tweets. Finally, the MediaEval Verifying Multimedia Use benchmark dataset [251] used in [37] contains images and tweets instead of news articles.

FakeNewsNet Dataset

This dataset consists of details about the news content, publisher information, and social engagement information [58]. The ground truth labels are collected from journalist experts such as Buzzfeed and the fact-checking website Politifact. The dataset is divided into two networks, Buzzfeed and Politifact, and the news contents are collected from Facebook web links. We downloaded all the available images related to the news in this dataset. The publishers' bias is retrieved from the dataset described in the next section. We merged together the news from both Politifact and Buzzfeed to have a larger dataset to work with. After cleaning the dataset from missing news bodies or headlines, we obtained a total of 384 news, 175 misleading and 209 factual.

BuzzFeedNews dataset

It contains news regarding the 2016 U.S. election published on Facebook by nine news agencies [31]. This dataset labels 356 news articles as left-leaning and 545 as right-leaning articles, while 1264 are mostly true, 212 are a mixture of true and false, and 87 are false.

MediaBias/FactCheck Dataset

To exploit the partisan information of the news source, we crawled the website mediabiasfactcheck.com, whose main goal is to educate the public on media bias and deceptive news practices. This website contains a comprehensive list of news sources, their bias, and their credibility of factual reporting scores. Here, the publisher's political bias is defined by using seven degrees of bias: *extreme-right*, *right*, *right-centered*, *neutral*, *left-centered*, *left*, and *extreme-left*. We collected the factual

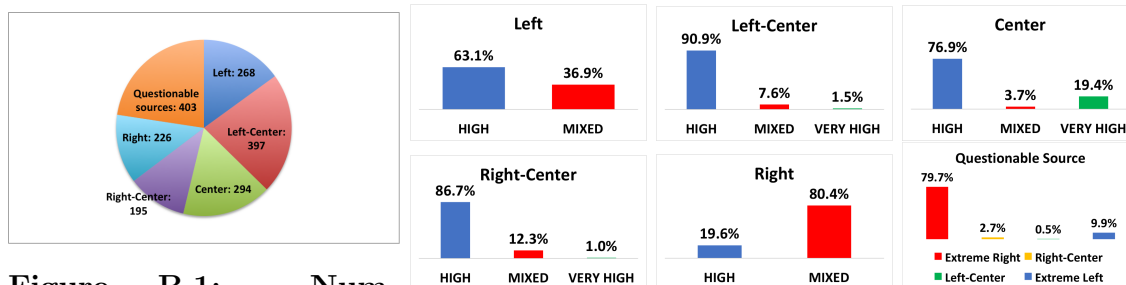


Figure B.1: Number of publishers per category in the MediaBias/FactCheck dataset. **Figure B.2:** Publisher credibility per bias and category in the MediaBias/FactCheck dataset.

reporting score of all the news sources under five categories: *Left bias* (moderately to strongly biased toward liberal causes), *Left-center* (slight to moderate liberal bias), *Least* (minimal bias), *Right-Center* (slightly to moderately conservative in bias), and *Right bias* (moderately to strongly biased toward conservative causes). The credibility score of these publishers falls into three categories: *Very high* (which means the source is always factual), *High* (which means the source is almost always factual) and *Mixed* (which means the source does not always use proper sourcing or sources to other biased/mixed sources). We also collected the publisher bias under the category *Questionable Sources*, which contains extremely biased publishers, mainly doing propaganda and/or writing misleading news. The number of publishers in each category considered is reported in Figure B.1. We retrieved a total of 1,783 publishers. The relationship between the source bias and its credibility is analyzed in Section B.4.3.

B.4 Multi-modal Features

We now describe the set of features we used in the chapter to analyze misleading political news. We consider four modalities, namely news content, and headline, images, and source bias.

B.4.1 Textual Features

Several approaches have been developed to extract features from text, from the widely used bag-of-words to the most recent BERT [109] deep learning-based approach. Although these approaches are popular in text analysis, they generate topic-dependent feature representation of documents that are not suitable for the dynamic environment of news where stories' topics change continuously. Therefore, in our analysis, we consider features that focus on linguistic style, text complexity, and psychological aspect to detect misleading news, such as Linguistic Inquiry and Word Count (LIWC) and text readability measures. Another approach is the Rhetorical Structure Theory (RST) which captures the writing style of documents [58]. However, as research has shown that the performance of LIWC is comparatively better than RST [58], we did not use RST in our analysis. Thus, to analyze the text of news body and headline, we consider the following groups of features (we also consider the number of stop words and upper case word count as additional features for news headline).

Linguistic Inquiry and Word Count (LIWC) LIWC is a transparent text analysis tool that counts words in psychologically meaningful categories. We use the LIWC 97 measures for analyzing the cognitive, affective, and grammatical processes in the text. To examine the difference between the factual and misleading news writing style, we divide the LIWC features into four categories [174]:

Linguistics features (28 features) refer to features that represent the functionality of text such as the average number of words per sentence and the rate of misspelling. This category of features also includes negations as well as part-of-speech (Adjective, Noun, Verb, Conjunction) frequencies.

Punctuation features (11 features) are used to dramatize or sensationalize a news story that can be analyzed through punctuation types used in the news such as Periods, Commas, Question, Exclamation, and Quotation marks, etc.

Similarly, *psychological features* (51 features) target emotional, social process, and cognitive processes. The affective processes (positive and negative emotions), social processes, cognitive processes, perceptual processes, biological processes, time orientations, relativity, personal concerns, and informal language (swear words, nonfluencies) can be used to scrutinize the emotional part of the news.

Summary features (7 features) define the frequency of words that reflect the thoughts, perspective, and honesty of the writer. It consists of Analytical thinking, Clout, Authenticity, Emotional tone, Words per sentence, Words more than six letters, and Dictionary words under this category.

Readability Readability measures how easily the reader can read and understand a text. Text complexity is measured by using attributes such as word lengths, sentence lengths, and syllable counts. We use popular readability measures in our analysis: Flesh Reading Ease, Flesh Kincaid Grade Level, Coleman Liau Index, Gunning Fog Index, Simple Measure of Gobbledygook Index (SMOG), Automatic Readability Index (ARI), Lycee International Xavier Index (LIX), and Dale-chall Score. Higher scores of Flesch reading-ease indicate that the text is easier to read, and lower scores indicate difficult to read. Coleman Liau Index depends on characters of the word to measure the understandability of the text. The Gunning Fog Index, Automatic Readability Index, SMOG Index, Flesh Kincaid Grade Level are algorithmic heuristics used for estimating readability, that is, how many years of education is needed to understand the text. Dale-Chall readability test uses a list of words well-known

for the fourth-grade students (easily readable words) to determine the difficulty of the text. We use this group of 9 readability features to measure news writing style complexity.

B.4.2 Image Features

To analyze the image associated with the news, we consider several tools, including (1) the ImageNet-VGG19 state-of-the-art deep-learning-based techniques to extract features from the images, (2) features describing face emotions, and (3) features referring to image quality such as noise and blur detection. Details regarding the features extracted to analyze images are reported in the following.

ImageNet-VGG19 We used a VGG19 pre-trained model from Keras for the visual feature extraction, which demonstrated a strong ability to generalize the images outside the ImageNet dataset via transfer learning [252]. We removed the classification layer of the VGG19 model and used the last fully connected layer of the neural network to generate a vector of latent features representing each input image. We used PCA to reduce the number of extracted features to 10.

Face Emotions Images associated with factual news articles typically depict a figure speaking, whereas the misleading news articles contain more images of people with only expressions on their faces. Further, images in real news usually portray people with more positive expressions than people depicted in misleading news images. Thus, to capture face emotions in images, we used Microsoft Azure Cognitive Services API to detect faces in an image ² which extracts several face attribute features. Among all the features extracted, we consider face emotion (anger, contempt, disgust,

²<https://docs.microsoft.com/en-us/azure/cognitive-services/face/quickstarts/csharp>

Table B.2: Feature ablation for FakeNewsNet (left) and BuzzFeedNews (right) datasets.

Features	AUROC	F1	Avg. Prec.
News Content			
Readability	0.622	0.520	0.530
Punctuation (LIWC)	0.744	0.625	0.662
Linguistic (LIWC)	0.732	0.599	0.642
Psychological (LIWC)	0.728	0.623	0.634
Summary (LIWC)	0.666	0.550	0.542
All LIWC	0.751	0.615	0.666
All (Feature reduction (30))	0.784	0.663	0.697
Headline			
Upper Case WC	0.630	0.536	0.525
Stop Word Count	0.640	0.577	0.514
Readability	0.680	0.589	0.579
Punctuation (LIWC)	0.716	0.570	0.639
Linguistic (LIWC)	0.679	0.544	0.561
Psychological (LIWC)	0.604	0.520	0.503
Summary (LIWC)	0.674	0.557	0.596
All LIWC	0.675	0.547	0.639
All (Feature reduction (30))	0.801	0.657	0.756
Bias	0.868	0.739	0.670
Image			
Face Emotions	0.559	0.415	0.431
ImageNet-VGG19	0.534	0.420	0.419
Image Quality	0.551	0.430	0.400
All (Feature reduction (10))	0.595	0.479	0.466

Features	AUROC	F1	Avg. Prec.
News Content			
Readability	0.638	0.355	0.306
Punctuation (LIWC)	0.735	0.453	0.342
Linguistic (LIWC)	0.706	0.416	0.332
Psychological (LIWC)	0.741	0.446	0.400
Summary (LIWC)	0.675	0.399	0.302
All LIWC	0.762	0.477	0.410
All (Feature reduction (30))	0.771	0.477	0.410
Headline			
Upper Case WC	0.700	0.454	0.316
Stop Word Count	0.668	0.408	0.293
Readability	0.672	0.388	0.319
Punctuation (LIWC)	0.686	0.403	0.348
Linguistic (LIWC)	0.639	0.367	0.276
Psychological (LIWC)	0.631	0.357	0.298
Summary (LIWC)	0.621	0.347	0.265
All LIWC	0.734	0.445	0.386
All(Feature reduction (30))	0.794	0.520	0.420
Bias	0.708	0.563	0.386

fear, happiness, neutral, sadness, and surprise) and smile features. Each of these features ranges in $[0,1]$ and indicates the confidence of observing the feature in the image.

Image Quality Misleading news images are more likely to have been manipulated (e.g., via photoshop) and have a lower quality than factual news images typically. Thus, to capture news image quality to some extent, we computed the amount of blur in an image by using the OpenCV blur detection tool ³ implementing a method based on the Laplacian Variance [111] along with noise level of face pixels provided by Microsoft Azure Cognitive Service API.

³<https://www.pyimagesearch.com/2015/09/07/blur-detection-with-opencv/>

B.4.3 Source Bias

Several studies in the field of journalism have theorized a correlation between the political bias of a publisher and the trustworthiness of the news content it distributes [114, 115]. To validate this assumption, we examine the relationship between the political bias of a news source and its credibility by analyzing the information about 1,785 publishers in the MediaBias/FactCheck dataset.

Figure B.2 shows the distribution of the credibility score per political bias category (from Left to Right) and the bias distribution in the questionable sources. The plots show that when the news source is moderate to strongly biased (either conservative or liberal), then the source is more likely to publish misleading news than other news sources that are more moderate and declared as left-centered, right-centered, or neutral. Also, we see that *Extreme-right* (or strongly conservative) is the predominant bias among the questionable sources. Thus, we also use the news source bias as another modality in our analysis.

B.5 Multi-modal Analysis

We used each group of features described in the previous section in input to a logistic regression classifier with L2 regularization (with 5-fold cross-validation) to compute the performance of these features in classifying factual vs. misleading stories. We also tried other classifiers such as Support Vector Machine (SVM) and Random Forest, but Logistic Regression achieved the best results. Hence, we report in the chapter Logistic Regression results only. We used class weighting to deal with class imbalance. The results for logistic regression are reported in Table B.2 according to the area under the ROC curve (AUROC), F1-measure (F1), and average precision (AvgP) and discussed

Table B.3: Top-30 most important news body content features and their corresponding logistic regression coefficients for the FakeNewsNet (left) and BuzzFeedNews (right).

FakeNewsNet				BuzzFeedNews			
Factual		Misleading		Factual		Misleading	
-0.97	assent	1.77	death	-1.08	affect	0.97	posemo
-0.87	hear	1.02	discrep	-0.71	fleschkincaid	0.86	negemo
-0.86	interrog	0.85	sexual	-0.61	dalechallknown	0.77	smog
-0.84	risk	0.82	informal	-0.61	nonflu	0.62	ari
-0.83	sad	0.81	motion	-0.55	dalechallscore	0.48	bio
-0.83	Parenth	0.69	shehe	-0.46	Dash	0.46	male
-0.61	relativ	0.68	family	-0.44	percept	0.45	filler
-0.54	compare	0.68	swear	-0.43	SemiC	0.43	female
-0.54	gunningfog	0.67	bio	-0.43	body	0.36	see
-0.52	auxverb	0.65	QMark	-0.43	ingest	0.35	affiliation
-0.51	i	0.54	colon	-0.41	gunningfog	0.34	anx
-0.51	drives	0.53	they	-0.40	swear	0.33	relig
-0.50	cogproc	0.51	netspeak	-0.29	shehe	0.28	Colon
-0.45	social	0.51	tentat	-0.25	friend	0.26	adverb
-0.45	you	0.51	adj	-0.25	netspeak	0.26	assent

in the following.

News Body Content The first modality we analyze is the news body content. Here, we see that the LIWC features are better than the readability features for both the datasets: 0.75 vs. 0.62 AUROC for FakeNewsNet and 0.76 vs. 0.64 for BuzzFeedNews. Also, performances are comparable for both the dataset, according to AUROC. One difference between the two datasets is the most important group of features within the LIWC features: punctuation features are the most important ones for FakeNewsNet (0.74 AUROC, 0.63 F1, 0.66 AvgP) whereas psychological features (0.69 AUROC, 0.40 F1, 0.35 AvgP) are the best predictors for the BuzzFeedNews dataset. As the latter has a higher class imbalance than FakeNewsNet (19% vs. 45% of misleading news), we obtain lower values of F1-measure and average precision.

Combining both readability and LIWC features (and by performing feature reduction to avoid overfitting) classification results improve with respect to each group of features individually: AUROC of 0.78 for FakeNewsNet and 0.77 for BuzzFeedNews. Feature reduction consists of the most informative features in the news body content computed by using the coefficients of a logistic regression model (30 features in total, 15 for factual news, and 15 for misleading ones). Table B.3 shows these most important features for FakeNewsNet and BuzzFeedNews and the corresponding coefficients from the logistic regression model. We see that readability features appear within the most important features in both datasets. By comparing the readability of factual and misleading news, we observe that factual news is harder to understand. We have, on average, higher values of readability scores in factual than misleading news, indicating higher text complexity (cf. Figure B.3). On the other hand, misleading news uses more informal language and tentative words evoking uncertainty than factual ones. As we see in Figure B.3, on average, misleading news has higher scores for these language features on both datasets: higher frequency of informal words (e.g., ‘thnx’, ‘hmm’, ‘youknow’), swear words, and netspeak (words

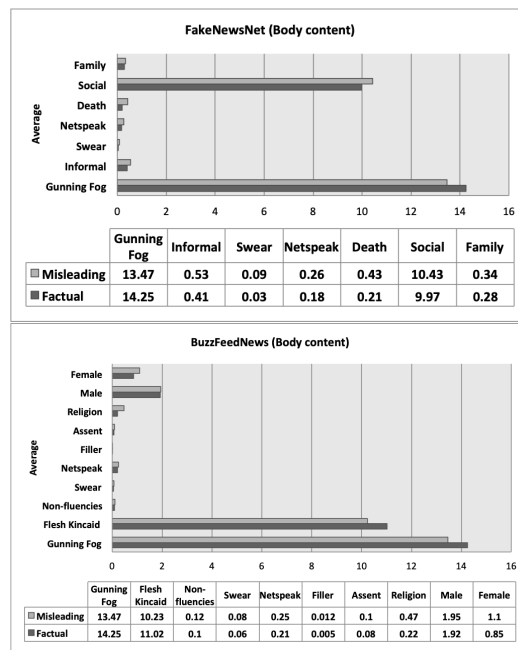


Figure B.3: Most important features for news body content with average values for factual and misleading news: FakeNewsNet (top) and BuzzFeedNews (bottom).

corresponding coefficients from the logistic regression model. We see that readability features appear within the most important features in both datasets. By comparing the readability of factual and misleading news, we observe that factual news is harder to understand. We have, on average, higher values of readability scores in factual than misleading news, indicating higher text complexity (cf. Figure B.3). On the other hand, misleading news uses more informal language and tentative words evoking uncertainty than factual ones. As we see in Figure B.3, on average, misleading news has higher scores for these language features on both datasets: higher frequency of informal words (e.g., ‘thnx’, ‘hmm’, ‘youknow’), swear words, and netspeak (words

Table B.4: Top-30 most important headline features and their corresponding logistic regression coefficients for FakeNewsNet (left) and BuzzFeedNews (right) datasets.

FakeNewsNet				BuzzFeedNews			
Factual		Misleading		Factual		Misleading	
-1.13	colemanliau	1.47	ari	-0.62	dalechallknown	0.35	# uppercase words
-1.12	Paranth	1.10	friend	-0.42	swear	0.22	ari
-1.10	affiliation	1.04	we	-0.39	nonflu	0.17	informal
-0.89	negate	0.67	Exclam	-0.36	# stopwords	0.17	fleshkincaid
-0.83	fleshkincaid	0.94	sexual	-0.32	assent	0.15	WPS
-0.76	# stopwords	0.79	motion	-0.22	netspeak	0.15	Exclam
-0.60	shehe	0.60	tentat	-0.20	dalechallscore	0.15	health
-0.48	relativ	0.57	family	-0.18	colemanliau	0.14	hear
-0.43	lix	0.55	space	-0.11	home	0.13	relig
-0.39	i	0.46	netspeak	-0.10	drives	0.13	female
-0.38	home	0.46	differ	-0.09	time	0.12	they
-0.33	male	0.45	they	-0.08	i	0.12	affiliation
-0.33	nonflu	0.45	reward	-0.08	WC	0.10	ingest
-0.32	bio	0.41	time	-0.08	Apostro	0.09	male
-0.30	Colon	0.37	body	-0.08	social	0.08	power

frequently used in social media and text messaging in FakeNewsNet, and higher frequencies of non-fluencies (e.g. ‘er’, ‘umm’, ‘uh’, ‘uh-huh’), swear words, netspeak, filler words and assent words in BuzzFeedNews. The above analysis clearly shows that factual news in both datasets is written with complex constructions of texts, which is mostly seen in the field of journalism [253], unlike the misleading ones which are written informally showing non-professional character.

Also, misleading news in both datasets has higher frequencies of psychology related words such as personal concerns (death in FakeNewsNet and religion-related words in BuzzFeedNews) and social words (e.g., social and family-related words in FakeNewsNet and male and female related words in BuzzFeedNews).

News Headline Among all the features we considered to analyze the news headline, we see in Table B.2 that, LIWC punctuation features are the best group of features in

both datasets achieving an AUROC of 0.72 (resp. 0.69), an F1-measure of 0.57 (resp. 0.40) and an average precision of 0.64 (resp. 0.35) on FakeNewsNet (resp. BuzzFeedNews) dataset. This shows that the headline’s features are more consistent across datasets than news body content. Similarly to the news body content, by combining both readability and LIWC features (and by performing feature reduction to avoid overfitting as we did for excerpt features), classification results improve with respect to each group of features individually: AUROC of 0.80 for FakeNewsNet and 0.79 for BuzzFeedNews.

Table B.4 shows the most important headline features in our datasets. Figure B.4 shows the average values for factual vs. misleading news of the best features discussed in the following. Again, readability measures appear among the most important features in both datasets. Comparing the average values of readability features between factual and misleading news provides evidence that factual news headlines are written professionally than misleading ones. Also, factual news headlines of both datasets have a higher average value of stopwords count, while BuzzFeedNews misleading news headlines are written using more capital

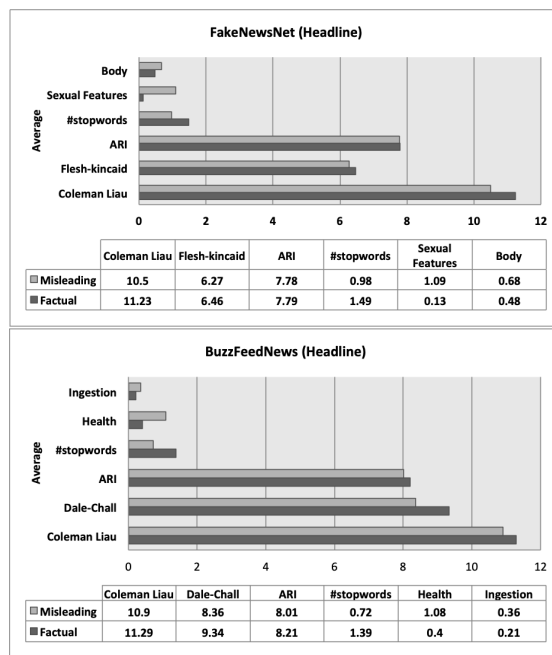


Figure B.4: Most important features for news headline with average values for factual and misleading news: FakeNewsNet (top) and BuzzFeedNews (bottom).

letters.

In addition, we see that the misleading news headlines have higher frequency of words related to biological processes (e.g., ‘eat’, ‘blood’, ‘pain’), namely sex (e.g., ‘love’, ‘incest’, ‘beauty’) and body lexicon (e.g., ‘cheek’, ‘hands’, ‘lips’) in FakeNewsNet, and health related words (e.g., ‘clinic’, ‘pill’, ‘ill’) and ingestion (e.g., ‘eat’, ‘dish’) in BuzzFeedNews.

This analysis shows that the orientation towards the feelings, body, and health lexicon is a very strong characteristic of a misleading news headline. Observing such biological words occurring significantly more in misleading news than in factual ones indicates that the former is made more sensational along with more uppercase letters for exaggerations to catch the reader’s attention.

News Source Bias The news source bias is a strong predictor for news credibility in both the datasets considered, and it achieves AUROC of 0.87 (resp. 0.71), F1-measure of 0.74 (resp. 0.56), and average precision of 0.67 (resp. 0.39) in the FakeNewsNet (resp. BuzzFeedNews) dataset. This result further confirms the correlation between source bias and the credibility of the news it distributes. It is worth noting that the publisher’s information is independent of the news labels as the former is collected from MediaBias/FactCheck, while the latter from Buzzfeed and Politifact.

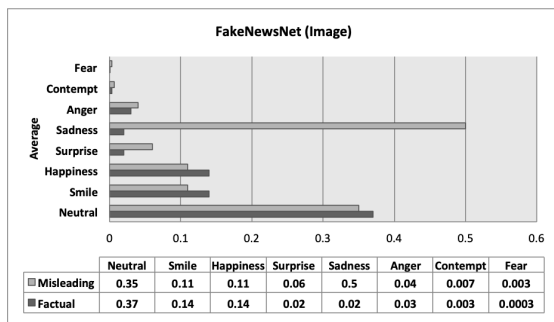


Figure B.5: Most important features for news image and average values for factual and misleading news.

Table B.5: Top-10 most important image features and corresponding logistic regression coefficients for FakeNewsNet.

Factual		Misleading	
-0.16	happiness	1.02	surprise
-0.16	smile	0.61	sadness
-0.14	noise	0.29	anger
-0.07	neutral	0.09	contempt
-0.03	VGG19	0.08	fear

News Image Image features are not as good as other modalities in detecting misleading news in the FakeNewsNet dataset. However, when we use the image associated with the news to determine the news validity, we see that features describing face emotions achieve best results according to AUROC (0.56) and average precision (0.43), while image quality features are the best according to F1-measure. Moreover, by combining all the image features (and performing feature reduction by considering only the top-10 most important features according to the coefficients of the logistic regression), we improve the classification results up to 0.60 AUROC, 0.48 F1-measure, and 0.47 average precision. The top-10 most important image features are reported in Table B.5. As expected, we see the face emotion-based features to be the most important ones. Figure B.5 shows the average values for factual vs. misleading news of the best image features. Here, we see that, on average, images associated with factual news depict people with more neutral-positive emotions (neutral, smile, happiness) than images associated with misleading news. On the other hand, misleading news is paired with more provocative images showing people expressing, on average,

Table B.6: Results comparing news snippet feature combination (headline, image, and source bias) with news body content for FakeNewsNet (left) and BuzzFeedNews (right).

Features	AUROC	F1	Avg. Prec.
Headline	0.801	0.657	0.756
Headline + Image	0.821	0.678	0.725
Headline + Image + Bias	0.908	0.783	0.817
News Content	0.784	0.663	0.697

Features	AUROC	F1	Avg. Prec.
Headline	0.794	0.520	0.420
Headline + Bias	0.812	0.534	0.462
News Content	0.771	0.477	0.410

more surprise, sadness, anger, contempt, and fear. Also, only one ImageNet-VGG19 feature appears in the top-10, where we find the noise level of face pixel feature as well.

B.5.1 Do We Need to “Read”?

Here, we address the question of whether we need to look at the news body content to detect misleading news, or we can achieve better results by using other modalities. Fairbanks et al. [254] posed and investigated this question for the first time and found that exploiting web links within news articles’ bodies outperforms body text-based features for misleading news detection. To address the question in our case, we can refer to the first part of our analysis and Table B.2. We see that, in both datasets, we get better information from the news headline to determine whether it is factual or not: AUROC of 0.80 vs. 0.78 in FakeNewsNet and 0.79 vs. 0.77 in BuzzFeedNews. This result confirms and generalizes by using larger datasets the finding of Horne and Adali [1] that the news title is more informative than the body content. Moreover, in the case of the FakeNewsNet dataset, considering the publisher bias achieves a better AUROC of 0.87.

B.5.2 Can we Detect Misleading News from its Snippet?

Next, we address the question of whether combining headline, bias and image features, hence considering the news snippet and mimic how news is distributed on social networks, can further improve misleading news detection results. Table B.6 report the combined results for FakeNewsNet (left) and BuzzFeedNews (right). For headline, image, and content, we consider the most important features previously computed via feature reduction (30, 10, 30 features, respectively). The first observation is that, even if the image features alone are not enough to differentiate between factual and misleading news (AUROC of 0.60 in the FakenewsNet, cf. Table B.2), we see from Table B.6 (left) that they help in improving classification results when combined with the headline features (2% improvement for AUROC and F1-measure). Moreover, adding the source bias further improves up to 0.91 AUROC, 0.78 F1-measure, and 0.82 average precision. In the case of the BuzzFeedNews dataset, we do not have image information, but Table B.6 (right) shows that adding the bias to the headline features achieves 0.81 AUROC, 0.53 F1-measure, and average precision 0.46, which is better than only considering the news body content. It is worth noting that, as reported in Section B.2, Potthast et. al [31] addressed the problem of automatically detecting misleading stories in the BuzzFeedNews dataset achieving an F1-measure of 0.46. They only analyzed news content with a different set of style-based features. However, their experimental setting was different from the one of this chapter. Thus, for a fair comparison with the methods used in this chapter, we reproduced their setting (considering only the left-wing articles and the right-wing articles of the corpus and balancing the dataset via oversampling) and computed classification results. We achieve an F1-measure of 0.58 with the news body content (best 30 features from

readability and LIWC) and F1-measure of 0.61 when we consider the combination of the best 30 headline features and source bias. In both cases, we improve their proposed method.

Thus, our analysis reveals that looking at the news snippet by considering the headline characteristics from Table B.4, checking the publisher bias and putting more attention on the associated images provides user-friendly tools that can be taught to people via media literacy to warn them about possible misleading news and can hopefully prevent people from massively spreading non-factual news through online social media.

B.6 Conclusion

We presented an analysis of the relative importance of different news modalities (body, headline, source bias, and visual content) in detecting misleading political news. In particular, our findings demonstrate a strong correlation between political bias and news credibility and the importance of image emotion features. Moreover, we showed that it is not necessary to analyze the news body to assess its validity, but comparable results can be achieved by looking at alternative modalities, including headline features, source bias, and visual content.

APPENDIX C:

AN ANALYSIS OF PEOPLE’S REASONING FOR SHARING REAL AND FAKE NEWS

The increase in the volume of fake news and its widespread over social media has gained massive attention as most of the population seeks social media for daily news diet. Humans are equally responsible for the surge of fake news spread. Thus, it is imperative to understand people’s behavior when they decide to share real and fake news items on social media. In an attempt to do so, we performed an analysis on data collected through a survey where participants (n= 363) were asked whether they were willing to share the given news item on their social media and explain the reasoning for their decision. The results show that the analysis presents several commonalities with previous studies. Moreover, we also addressed the problem of predicting whether a person will share a given news item or not. For this, we used intrinsic features from participants’ open-ended responses and demographics attributes. We found that the perceived emotions triggered by the news item show a strong influence on the user’s decision to share news items on social media.

C.1 Introduction

Social media has emerged as popular information source people rely on for events, breaking news, and emergencies. Indeed, it has become a source of daily news diet

for the increasingly large population. Statistics show that majority of the population (71% of American adults) ever get news through social media in 2020 [182] which was increased by 3% since 2018 [255]. The landscape of news consumption and information flow has drastically changed with the popularity of social media. It has transformed how news content is created, how people engage with news items, and share information, blurring the journalists' boundary in traditional media that is first verifying and then disseminating only the accurate news items [183]. Moreover, users in social media (both organizations and individuals) actively participate in creating and sharing news items with friends, families, and other readers due to its ease of use, lower cost, and convenience of further sharing [184, 30]. This shift of the news paradigm has led to an unprecedented transformation in both news quality and quantity that users encounter in social media, increasing the probability of potential encounters and the spread of fake news, fostering social media as a fertile ground for the production and propagation of fake news.

The sheer volume of fake news being observed in social media has recently become an obvious cause of concern. Many studies have highlighted the characteristics of fake news through linguistic and psychological attributes [32, 1, 31, 194], writing styles [216, 1, 217], network-based attributes [218] and hybrid attributes considering both linguistic and network [30].

Despite several studies illustrating cues to identify fake news and mitigate its spread, there is a worrisome amount of fake news widely spreading over social media. Fake news has been identified as more likely to go viral than real news, spreading faster and wider [64]. Additionally, an analysis of news about the 2016 election conducted by BuzzFeed, also found more engagement with fake news than real news

[46]. Earlier studies analyzed the potential reason behind this rapid diffusion of news in social media, focusing on various factors, including polarized communities of users with common belief (echo-chambers) [256], epidemiological models [257]. Some studies highlighted the actors responsible for spreading fake news, including bots and cyborgs [258]. Although bots are equally responsible for spreading real and fake news, the considerable spread of fake news is caused by human activity [64, 66] as people are generally not able to accurately identify which news item is fake and which is real [189]. Thus, it is crucial to understand the people's sharing behavior of fake and real news on social media to minimize fake news diffusion.

In this context, this study seeks to better understand how people reason when they decide to share real news and fake news. In particular, we surveyed 363 undergraduate students and asked participants to report and explain their willingness to share given news items (with headline and image) on their social media. We also leveraged the demographic attributes of participants like gender and political orientation in our study. We performed a comprehensive data analysis to investigate the pattern of news sharing behavior, the role of demographics in news sharing decisions, and why people share real and fake news. Furthermore, we addressed the problem of predicting whether a person will share a given news item or not according to emotion, psychological, and demographics features as a binary classification task.

Our experiments show several commonalities with previous findings regarding news-sharing behavior.

- News sharing is rare as only a small percentage (19.2% to 28.2%) of users expressed the willingness to share news in social media, regardless of news veracity.

- Female participants are prone to share more news than male participants regardless of news veracity.
- Left-leaning participants tend to share real news more than fake news, independently of the news source’s political orientation, and right-leaning participants were instead more prone to share news items from sources with the same political-leaning, independently of news veracity.
- The prominent themes illustrated by the approaches used by participants to make their sharing decisions fall under subjectivity and the focus on others’ interest or disinterest in news topic.
- Emotion features are more effective in predicting people’s willingness to share a given news item.

C.2 Related Work

Several studies have been conducted to understand the characteristics of users that are likely to contribute to spreading fake news on social networks. Vosoughi et al. [64] revealed that the fake news spreaders had, on average, significantly fewer followers, followed significantly fewer people, and were significantly less active on Twitter. Moreover, bots tend to spread both real and fake news, and the considerable spread of fake news on Twitter is caused by human activity. Shrestha and Spezzano showed that social network properties help in identifying active fake news spreaders [65]. Shu et al. [66] analyzed user profiles to understand the characteristics of users that are likely to trust/distrust fake news. They found that, on average, users who share fake news tend to be registered for a shorter time than the ones who share real news and

that bots are more likely to post a piece of fake news than a real one, even though users who spread fake news are still more likely to be humans than bots. They also show that real news spreaders are more likely to be more popular and that older people and females are more likely to spread fake news. Guess et al. [67] also analyzed user demographics as predictors of fake news sharing on Facebook and found out political-orientation, age, and social media usage to be the most relevant. Specifically, people are more likely to share articles they agree with (e.g., right-leaning people tended to share more fake news because the majority of the fake news considered in the study were from 2016 and pro-Trump), seniors tend to share more fake news probably because they lack digital media literacy skills that are necessary to assess online news truthfulness, and the more people post in social media, the less they are likely to share fake news, most likely because they are familiar with the platform and they know what they share.

Shrestha et al. [68] analyzed the linguistic patterns used by a user in their tweets and personality traits as a predictor for identifying users who tend to share fake news on Twitter data [69, 68]. Likewise, Giachanou et al. [75] proposed an approach based on a convolutional neural network to process the user Twitter feed in combination with features representing user personality traits and linguistic patterns used in their tweets to address the problem of discriminating between fake news spreaders and fact-checkers.

Ma et al. [190] went beyond the user and news characteristics and analyzed the characteristics of diffusion networks to explain users' news sharing behavior. They found opinion leadership, news preference, and tie strength to be the most important factors at predicting news sharing, while homophily hampered news sharing in users'

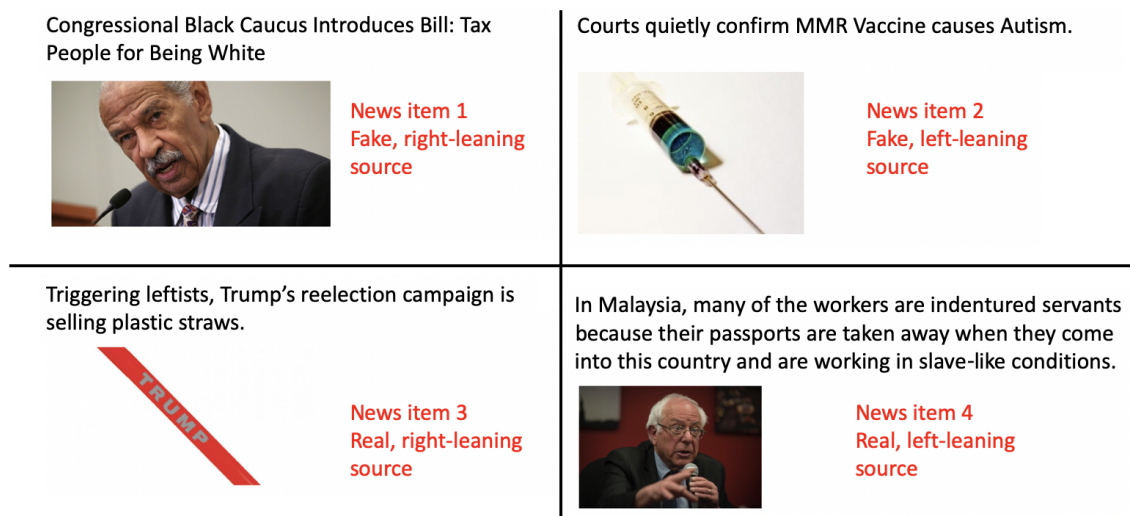


Figure C.1: News items used in our survey instrument.

local networks. Also, people driven by gratifications of information seeking, socializing, and status-seeking were more likely to share news on social media platforms [191].

C.3 Data Collection

We conducted an online survey delivered via Qualtrics. Through this online survey, participants were given four news headlines and accompanying images. For each news item, participants were asked whether they were willing to share the given news item on their social media and write an explanation of the reasoning for their decision. We considered the four news items shown in Figure C.1 and gathered from politifact.com. In this news set, two are real news items, and two are fake news items, as fact-checked by politifact.com. Both real and fake news items are one from a left-leaning source and one from a right-leaning source. News source political-leaning has been gathered from mediabiasfactcheck.com.

	Percentage of Sharing
News Item 1 (Fake)	19.2%
News Item 2 (Fake)	22.9%
News Item 3 (Real)	20.0%
News Item 4 (Real)	28.2%

Table C.1: News Sharing Behavior.

We recruited undergraduate students ($n = 363$) from a volunteer pool in general education social science courses (Psychology 101) to participate in our survey (258 F, 101 M, 4 Other; mean age 19.7, $SD = 4.25$). The research was approved by the university IRB. Participants were compensated with course credit (volunteering for studies being one option for a research experience requirement). Participants received no training.

C.4 Data Analysis

News sharing is rare. We start the analysis of our data by observing that only a small percentage of users expressed the willingness to share news in social media, independently of the veracity of the news. As shown in Table C.1, this percentage ranges between 19.2% and 28.2% among the news considered in our survey. Previous research [259] has shown that sharing news articles from fake news domains on Facebook was a rare activity during the 2016 U.S. presidential campaign. Our data on fake news sharing is aligned with this result, but our respondents also showed some preliminary evidence that this pattern may be true for real news sharing as well.

The role of demographics in news sharing. We collected demographic data from our survey participants, including gender, political orientation, and age. As most participants are in the same age range (18-25), we did not consider age in our

analysis.

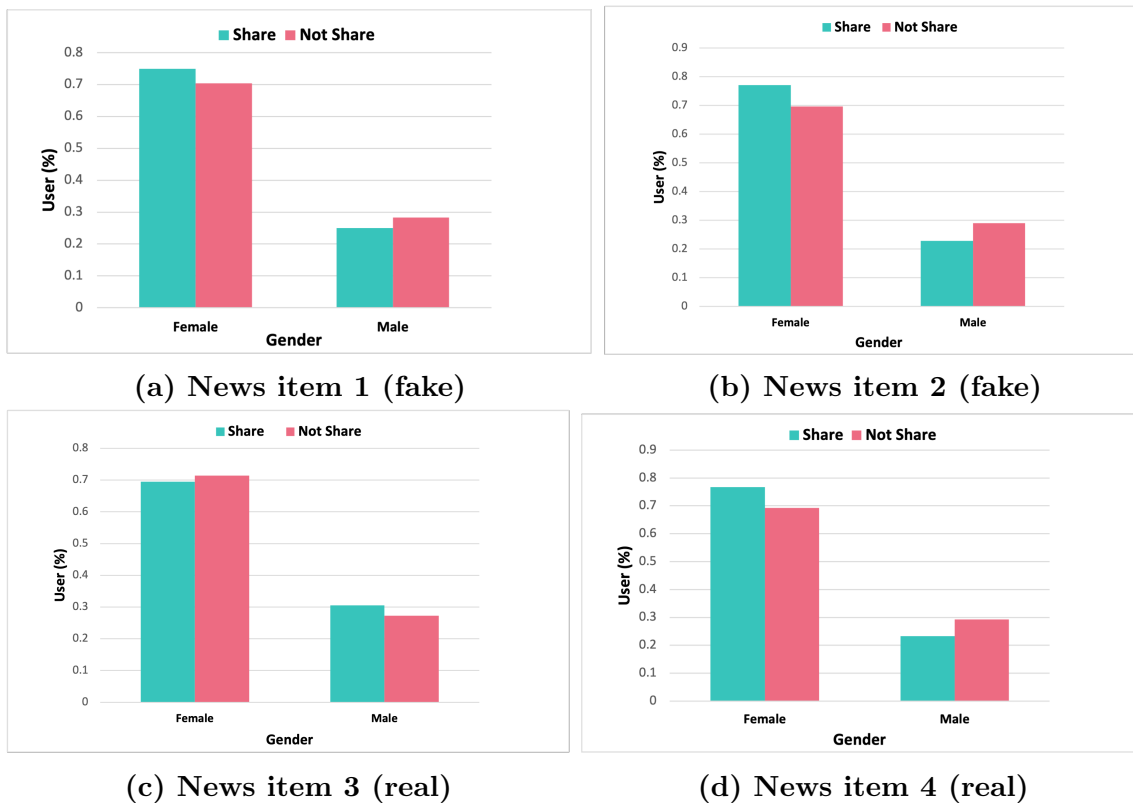


Figure C.2: Distribution of participant's gender.

When looking at differences in sharing behavior according to gender (see Figure C.2), we observe that the female participants were more prone to share both the fake news items considered than male participants who were more skeptical about the same news items. Shu et al. [237] in their studies have shown a similar result where female users tend to trust fake news more than male users. In general, females were more prone to share more news items than males (three vs. one).

Regarding participants political orientation, we see two interesting patterns as reported in Figure C.3: (1) left-leaning participants were more prone to share real news than fake news, independently of the political orientation of the news source; (2)

right-leaning participants were instead more prone to share news items from sources with the same political-leaning (news items 1 and 3), independently of news veracity. Similarly, Guess et al. [259] have shown that, in 2016, conservatives were more likely to share articles from pro-Trump fake news domains than liberals or moderates.

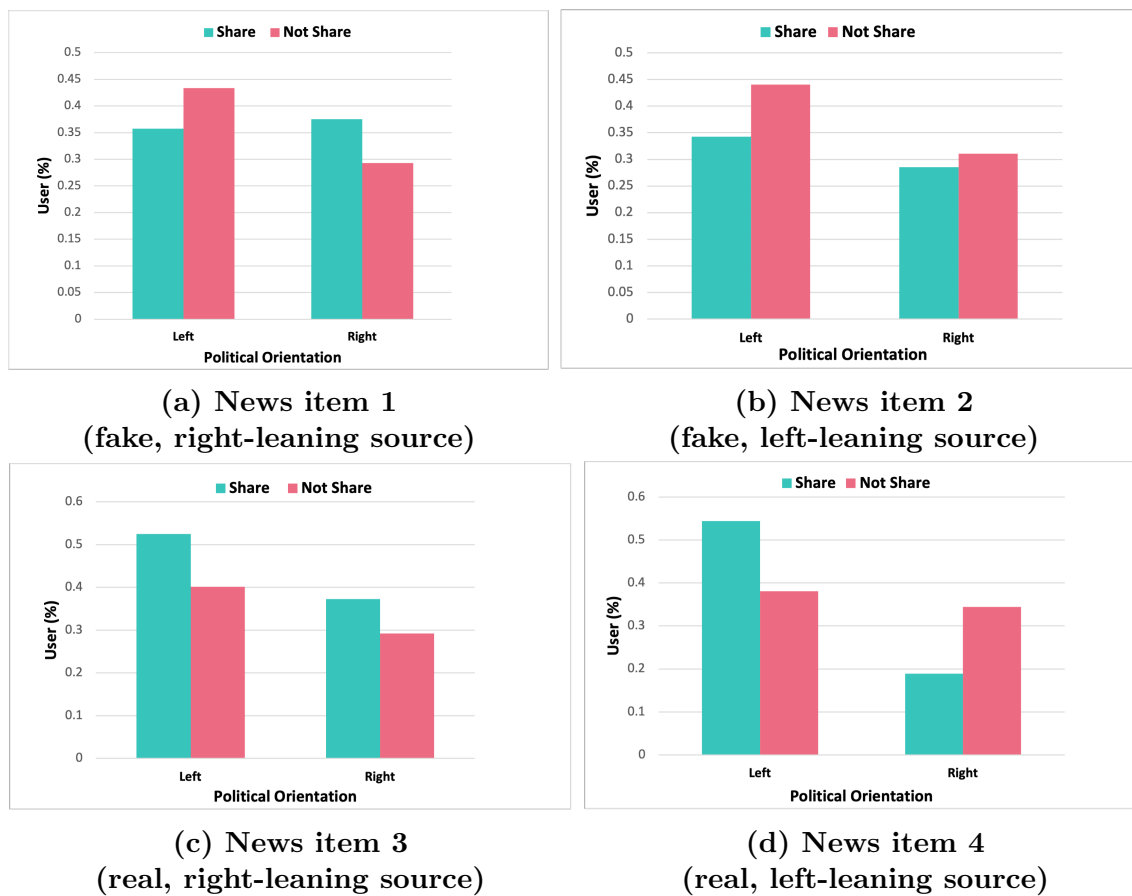


Figure C.3: Distribution of participant's self-identified political orientation.

Why people share real and fake news? Yaqub et al. [102] analyzed open-ended responses of participants in the study where they explained the reason behind their intention to share true, false, and satire headlines. In their study, the most frequent

rationales behind sharing/not sharing news were (1) the interest/non-interest towards the news, (2) the potential of generating discussion among the friends, (3) the fact that the news is not relevant to the user's life, and (4) the perceived news credibility, especially as a motivation for not sharing news.

We conducted a similar analysis on a sample of our data (n=25). Specifically, we conducted a thematic analysis to identify the prominent themes that illustrated the approaches used by participants to make their sharing decisions. We followed an inductive approach to generating codes [108]. We found out the principal codes to be focused on potential others ("My friends would/would not be interested in this"), interest or disinterest in the news topic, and subjectivity/the self ("I would/wouldn't share this because...", "I would call that fake/real") and are mostly aligned with the finding by Yaqub et al. [102].

Regarding performing credibility assessment before making the sharing decision, we also found in our sample data that this was performed more often for fake news (28% of the times for news item 1 and 56% for news item 2) than for real news (24% of the times for news item 3 and 16% for news item 4). Moreover, when performed, the credibility assessment was much more correct in the case of fake news (100% of the times for news item 1 and 93% for news item 2) than real news (67% of the times for news item 3 and 25% for news item 4).

Overall, the data analysis performed in this section shows that our collected data presents several commonalities with previous studies, ensuring we have quality data suitable for further investigations.

C.5 Predicting News Sharing

In this section, we address the problem of predicting whether a person will share or not a news item according to emotion and psychological features generated when they consider a news item and demographics (gender and political orientation) as well. We modeled the problem as a binary classification task where we computed emotion and psychological features from participants' open-ended responses to the survey question asking for an explanation of their decision to share or not the given news item.

C.5.1 Textual Features Extraction

Emotion Features (Emotion)

In order to compute a vector of scores quantifying participants' emotions when deciding whether or not to share a news item, we considered their open-ended survey responses and proceeded as follows. We started by cleaning responses' text by expanding contraction words, correcting misspellings and grammatical mistakes using LanguageTool¹ and replacing negated words with their WordNet antonym. Next, we extracted emotions from the text by using the Emotion Intensity Lexicon (NRC-EIL) [178] and EmoLex [260]. Emotion features computed via NRC-EIL include anger, joy, sadness, fear, disgust, anticipation, surprise, and trust, while Emolex² features include happy, sad, angry, don't care, inspired, afraid, amused, and annoyed. Feature vectors have been computed by using the approaches proposed in [179, 201]. In addition, we also considered emotion-related features as computed by the 2015 Linguistic Inquiry and Word Count (LIWC) [174] tool, which includes effective processes like anxiety, anger, positive and negative emotion.

¹<https://pypi.org/project/language-tool-python/>

²<https://sites.google.com/site/emolexdata/>

Psycho-linguistic Features (LIWC)

To understand the relationship between psychological states and the participants' decision-making, we considered the set of psycho-linguistic features computed by the Linguistic Inquiry and Word Count (LIWC) tool [174]. LIWC is a transparent text analysis tool that counts words in psychologically meaningful categories. Specifically, we considered psychological processes that include social processes (e.g., family, friends), cognitive processes (e.g., think, cause, perhaps), perceptual processes (e.g., see, heard, felt), biological processes (e.g., eat, pain, love), relativity (e.g., area, move, day) and personal concerns (e.g., work, leisure, achieve, home, money, religion, death).

Demographics (Demog)

As explicit features, we used participants' self-identified gender and political orientation to understand if the demographic attributes provide potential cues in predicting users' sharing decisions.

C.5.2 Experimental Setting and Results

We used each group of features described in the previous section as input to a random forest classifier to compute the performance of these features in predicting whether a reader of a news item (a participant of our survey) is willing to share or not the given news item on their social networks. We also tried other classifiers such as Support Vector Machine (SVM) and logistic regression, but random forest achieved the best results. Hence, in the chapter, we report the results of random forest only. We used class weighting to deal with the class imbalance and performed 5-fold cross-validation.

The results are reported in Table C.2 according to the area under the ROC curve

	Features	AUROC	AvgP	F1
News Item 1 (Fake)	LIWC	0.611	0.247	0.166
	Demog	0.518	0.207	0.228
	Emotion	0.720	0.403	0.228
	All	0.722	0.382	0.129
News Item 2 (Fake)	LIWC	0.608	0.307	0.175
	Demog	0.565	0.250	0.325
	Emotion	0.706	0.416	0.162
	All	0.707	0.421	0.122
News Item 3 (Real)	LIWC	0.586	0.310	0.257
	Demog	0.617	0.258	0.300
	Emotion	0.771	0.578	0.477
	All	0.796	0.585	0.439
News Item 4 (Real)	LIWC	0.611	0.397	0.302
	Demog	0.590	0.317	0.356
	Emotion	0.784	0.564	0.423
	All	0.786	0.562	0.359

Table C.2: Comparison of emotion, psycho-linguistic, and demographic features to predict whether a news item will be shared or not. We used a random forest classifier. Best results among feature groups considered separately are in bold. Best overall results are shaded.

(AUROC), average precision (AvgP), and F1-measure (F1). As can be seen, when each feature group is considered separately, emotion features are the best performing features compared to LIWC features and demographics with 72% vs. 61% and 52% AUROC and 40% vs. 25% and 20% average precision for news item 1, 71% vs. 61% and 57% AUROC and 42% vs. 31% and 25% average precision for news item 2, 77% vs. 59% and 62% AUROC and 58% vs. 31% and 26% average precision for news item 3 and 78% vs. 61% and 59% AUROC and 56% vs. 40% and 42% average precision for news item 4 (bold in Table C.2). We further considered a combination of all feature groups to see if combining demographics, psychological and emotional features can provide complementary information that can help improve the

prediction. We observed that when the combination of all feature groups is considered, the performance remained more or less the same if not improved according to AUROC (shaded in Table C.2). This demonstrates that emotion features are more effective than other groups of features considered in our study for predicting people’s sharing behavior. Hence, one of the motivations for potential news-sharing behavior in social media could be emotional persuasion. It will not be inaccurate to say that being persuaded by strong emotions like anger, fear, surprise, joy, etc., triggered by news content, people tend to get involved and share more news on social media. This finding aligns with the previous research by Berger et al. [261] which also states that emotional arousal tends to increase the likelihood of sharing news on social media.

C.6 Conclusion and Future Work

To sum up, this chapter presents findings from studying people’s reasoning when they decide to share real and fake news items provided with headlines and images. This chapter investigates the correlation between the user’s sharing decision and explicit attributes provided by participants like demographics and political orientation. Furthermore, we addressed the problem of predicting whether a person will share a given news item or not using intrinsic features like psychological and emotion from participants’ open-ended responses explaining their willingness to share given news item along with demographics attributes.

The results show that news sharing is rare, and among the participants expressing willingness to share, females are prone to share more news in general. Participants’ political orientation exerts a significant pattern on news sharing behavior that is left-leaning participants’ news sharing behavior is motivated by news veracity rather than political orientation. In contrast, it is the other way around for right-leaning

participants. Likewise, it shows the possibility of users sharing news items depends on the perceived relevance of news interest among friends and families. Moreover, this chapter also highlights that the perceived emotions triggered by the news item show a strong influence on user's news sharing behavior in social media.

One potential limitation of our study is that we have considered only four news of each political leaning (2 fake and 2 real). Considering a bigger set of news items could have shown significant patterns and support to our findings. Furthermore, this work focuses on a younger sample of the limited range of age, due to which we did not consider age in demographic attributes. It could have added some more insights regarding news sharing behavior among different age groups if we could consider participants of a wide range of ages (from younger to older population). We will address these limitations in our future work.

Acknowledgements

This work has been supported by the National Science Foundation under Award no. 1943370. We thank Brian Stone for facilitating the data collection and Ashlee Milton and Maria Soledad Pera for providing us with the code used in their papers [179, 201] to compute emotional features.

APPENDIX D:

MULTI-MODAL SOCIAL AND

PSYCHO-LINGUISTIC EMBEDDING VIA

RECURRENT NEURAL NETWORKS TO

IDENTIFY DEPRESSED USERS IN ONLINE

FORUMS

Depression is the most common mental illness in the U.S., with 6.7% of all adults who have experienced a major depressive episode. Unfortunately, depression extends to teens and young users as well, and researchers observed an increasing rate in recent years (from 8.7% in 2005 to 11.3% in 2014 in adolescents and from 8.8% to 9.6% in young adults), especially among girls and women. People themselves are a barrier to fight this disease as they tend to hide their symptoms and do not receive treatments. However, protected by anonymity, they share their sentiments on the Web, looking for help.

In this chapter, we address the problem of detecting depressed users in online forums. We analyze user behavior in the ReachOut.com online forum, a platform providing a supportive environment for young people to discuss their everyday issues, including depression. We propose an unsupervised technique based on recurrent neu-

ral networks and anomaly detection to detect depressed users. We examine the linguistic style of user posts in combination with network-based features modeling how users connect in the forum. Our results on detecting depressed users show that both psycho-linguistic features derived from user posts and network features are good predictors of users facing depression. Moreover, by combining these two sets of features, we can achieve an F1-measure of 0.64 and perform better than baselines.

D.1 Introduction

Depression is a mental illness commonly seen in people (6.7% of all U.S. adults have experienced at least one major depressive episode), which negatively affects their thoughts and behaviors. Depression causes mood fluctuations and impermanent emotional responses to the challenges of everyday life. Especially when lasting for a while and with moderate or severe intensity, depression may become a serious health condition. It can cause the affected person to suffer greatly and perform poorly at work, at school, and in the family. It has been one of the common problems seen in tens of millions of people. At its worst, depression can lead to suicide. Close to 800,000 individuals die due to suicide every year. According to a 2015 report by the World Health Organization, more than 300 million people are affected by depression. Unfortunately, depression extends to teens and young users as well, and researchers observed an increasing rate in recent years (from 8.7% in 2005 to 11.3% in 2014 in adolescents and from 8.8% to 9.6% in young adults), especially among girls and women. Very few people in the world receive the treatments provided for depression. In many countries, fewer than 10% of people in need receive such treatments. One of the barriers to this is the people themselves. They tend to hide their symptoms to avoid being known as psychiatric patients or because people are unaware of the

condition and what is happening with them. Online forums and social media are platforms where people, protected by anonymity, can share their thoughts freely and publicly and look for help. Thus, the content of online posts is a valuable source of information to analyze in order to infer the presence of mental illness in these users and take timely actions.

In this chapter, we address the problem of detecting users at risk of depression in online forums. Indeed, online posts provide a mean to infer an individual’s mood and socialization behavior. Our research contributes to automatically retrieving forum users that are potentially at risk and suggest them to the forum administrators for further investigation so that they can promptly act to take care of these people, eventually.

We formulate the problem as a binary classification task and use unsupervised techniques with (i) psycho-linguistic features describing the linguistic style of the user posts and the emotions expressed in them, and (ii) user networking behavior in the “who replies to whom” network extracted from the forum posts. Specifically, we propose a multi-modal methodology where a user embedding is first computed from the sequence of their posts via recurrent neural networks in an unsupervised fashion and, then, combined with user networking behavior. Finally, unsupervised anomaly detection is performed on these features to classify users as depressed or not.

We test our approach on a dataset extracted from ReachOut.com: an Australian non-profit online forum established in 1996 to support young people in addressing problems common to their generation, including alcohol and drug addiction, gender identity, sexuality, and mental health concerns. This dataset is made available by CLPsych’17 shared task. Related work on this dataset has analyzed user posts to

automatically triage them by their risk of being written by users suffering depression [262, 263]. Our results on detecting depressed users show that both psycholinguistic features derived from user posts and network features are good predictors of users facing depression. Moreover, by combining these two sets of features, we can achieve an F1-measure of 0.64 and perform better than baselines.

The chapter is organized as follows. Section D.2 summarizes related work, Section D.3 describes the dataset we used in this chapter, Section D.4 presents our proposed unsupervised technique to identify depressed users, Section D.5 reports on our experimental evaluations and, finally, conclusions are drawn in Section D.6.

D.2 Related Work

Researchers have been analyzing the online behavior of users in social media to detect depression. Resnik [264] studied topic models in the analysis of linguistic signals for detecting depression. These depression detection efforts demonstrated that it is possible to analyze depressed users on social media on a large scale. Preliminary research done by Park et al. [265] explored the use of language to describe depression utilizing real-time moods captured from Twitter users. Further, Park et al. [266] conducted face-to-face interviews with 14 active Twitter users to explore their behavior. They found that depressed users perceive Twitter as a tool for social awareness and emotional interaction. Using social network and linguistic patterns, Xu and Zhang [267] attempted to explain how Web users discuss depression-related issues. They found that depressed users have an intensive use of self-focus words and negative affect words. Zimmermann et al. [268] looked at how first-person pronoun use might be a predictor of future depressive symptoms. Computerized analysis of written text through LIWC features has also been applied to understand predictors of neurotic

tendencies and psychiatric disorders [269].

De Choudhury et al. [199] explored the potential of using Twitter to detect and diagnose major depressive disorders in an individual. Thus, to detect depressed users, they considered both linguistic and network features and achieved an F1-measure of 0.68. Similarly to our work, they found out that depressed users on social media exhibit lower reciprocity and a higher clustering coefficient than non-depressed ones (cf. Subsection D.4.2), but observed a different posting activity as measured by the insomnia index (cf. Subsection D.3.1). To predict depression, Eichstaedt et al. [270] performed a linguistic analysis of the history of Facebook statuses posted by patients visiting a large urban academic emergency department.

De Choudhury et al. [271] proposed a statistical metric named Social Media Depression Index (SMDI), which is used to predict indicative depressive posts on Twitter and also helps to categorized depression levels. MacAvaney et al. [272] addressed the problem of detecting posts on a dataset of annotated Reddit posts by including temporal information about the diagnosis and achieved an F1-measure of 0.55.

Many researchers have used the CLPsych'17 dataset (the same we use in this chapter) to perform linguistic analysis of the user posts and triage them by author level of risk. These works have used TF-IDF weighted unigrams, post embeddings using sent2vec [273], LIWC lexicon, as a measure of emotion [262] and sentiment [274]. Other works leveraged DepecheMood [275] to identify emotions associated with a post, and the MPQA subjectivity lexicon [276] to distinguish between objective and subjective posts.

Yates et al. [263] addressed both the post and user detection problems via deep-learning-based text classification. They used a Reddit dataset for user classification

and achieved an F1-measure of 0.65. For the user post detection problem, they used the ReachOut CLPsych'16 dataset (a previous version of the dataset we use in this chapter) and achieved the best F1-measure of 0.61, while the F1-measure for other linguistic methods they compared with ranges between 0.53 and 0.5 [277, 278, 279, 262]. Yates et al. also tested their proposed methodology on the ReachOut CLPsych'17 dataset (the same we use in this chapter) and reported an F1-measure of 0.50 for post detection.

Overall, several works have addressed the problem of (1) detecting depressed users in social media platforms such as Twitter and Reddit, and (2) identifying posts that may indicate a risk of depression. Experimental results reported in these works show that both user and post detection are challenging problems. Moreover, previous work has focused on supervised detection while, in this chapter, we propose an unsupervised technique to identify depressed users in online forums by using both psycho-linguistic and network features.

D.3 The ReachOut Forum

ReachOut.com¹ is an Australian non-profit online forum available for free, which is well reached by common people in Australia (1.58 million of visitors each year [280]). This forum provides mental health services along with information and environment to support the youth of age 14-25 so that they can share their mental issues and experiences anonymously. Based on the communications through posts, young people are provided with resources, help, and proper guidance from well-trained moderators. The practical support and tips provided by this organization make it easier for parents to help their children facing mental illness.

¹<https://au.reachout.com/>

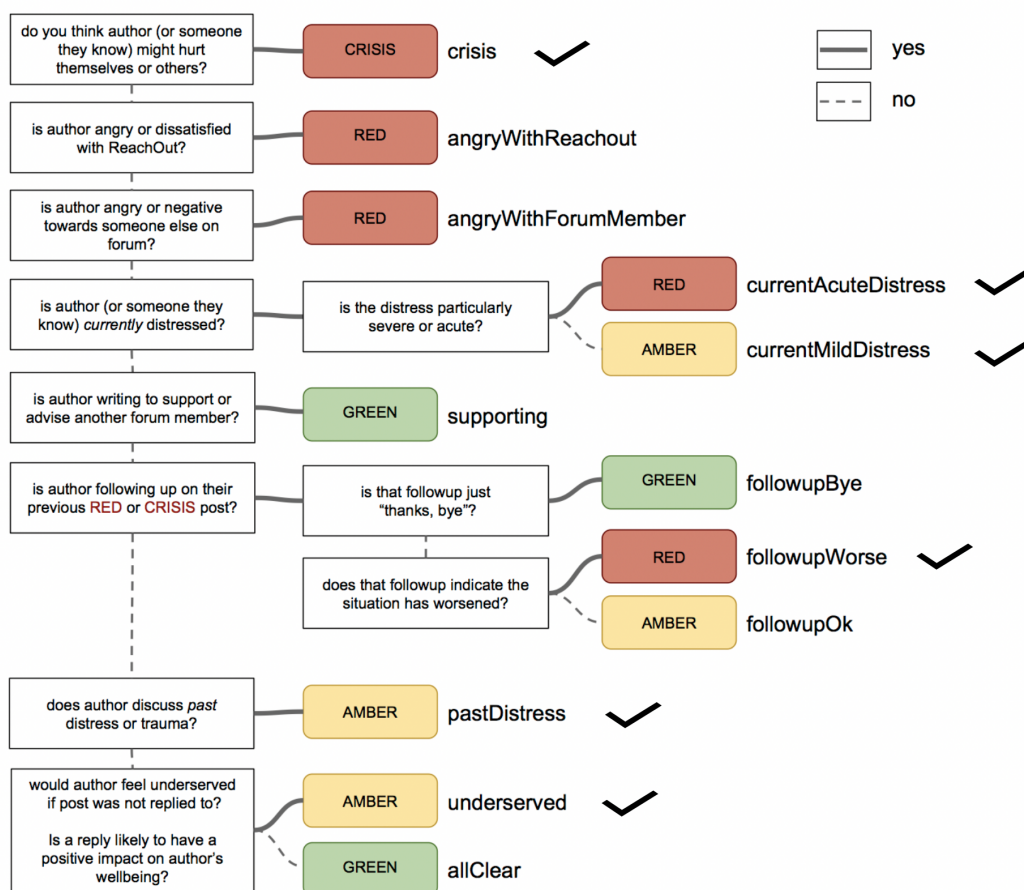


Figure D.1: The triage annotation decision tree [3].

In 2013, a survey was conducted among the users of the forum, showing that 33% of Australian young people are aware of the site and proving that the forum was beneficial to support people with mental disorders [281]. The survey results showed that “77% of participants reported experiencing high or very high levels of psychological distress”, and that 46% of these distressed visitors “were more likely to seek help from at least one professional source after visiting ReachOut.”

D.3.1 Dataset

We used the dataset provided by the 2017 CLPsych shared task [3] containing labeled forum posts from the ReachOut.com platform. The dataset contains a total of 147,619 forum posts, out of which, 1,588 were manually annotated by three separate judges according to the following categories (indicating how urgently the post requires moderator’s attention):

1. ***Crisis*** indicates that the author is at imminent risk of being harmed, themselves, or others. It should be prioritized above all others.
2. ***Red*** indicates that a moderator should respond to the post as soon as possible.
3. ***Amber*** indicates that a moderator should address the post at some point, but they do not need to do so immediately.
4. ***Green*** identifies posts that do not require direct input from a moderator and can safely be left for the wider community of peers to respond.

Moreover, the annotators added further information regarding the motivation of why the post may or may not need attention, according to the flowchart shown in Figure D.1. We considered the types of annotated posts marked with the tick symbol in the above figure as posts dealing with users who have a mental disease. Therefore, as our task is user-oriented: among all the users who authored at least one annotated post, we consider a user as depressed if they have posted or commented at least one post that is annotated as *crisis*, *currentAcuteDistress*, *currentMildDistress*, *followupWorse*, *pastDistress*, *undeserved* and non-depressed, otherwise. Overall, we marked 65 users as depressed and 94 users as non-depressed. Table D.1 details the size of our dataset.

Table D.1: Summary of the CLPsych 2017 dataset.

Users	Posts	Depressed	Non-Depressed	Unknown
1,716	62,036	65	94	1,557

Posting Activity

Insomnia is one of the major symptoms of depression, and literature on depression indicates that users showing depression signs tend to be more active during the evening and night, indicating insomnia as a promising feature for depression detection [282]. Thus, we analyzed the posting behavior of the users as in [199]. We divided the time into day and night and considered the ‘night’ window as ‘9PM-6AM’ and the ‘day’ window as ‘6:01AM-8:59PM’ (we used the local time of the user), and analyzed the average number of posts during these windows for depressed and non-depressed users.

De Choudhury et al. [199] showed that depressed users in Twitter tend to post more at night and have a higher insomnia index, defined as the normalized difference in the number of posts made during the night window and the day window. Conversely, in our dataset, we observe that, in general, all the users tend to post more during the day than at night, and that depressed users tend to post more than non-depressed ones during both night and day time. On average, depressed users write 1.75 posts at night (vs. 0.76 posts for non-depressed users) and 4.79 posts during the day (vs. 2.79 posts for non-depressed users). One reason could be that the forum is specifically open to provide help, so depressed people feel free to post there at any time of the day, while they engage more with other online activities such as Twitter at night when their symptoms worsen [282]. Thus, we did not find the insomnia index as an important feature in our dataset, and then, we did not use it for identifying depressed users.

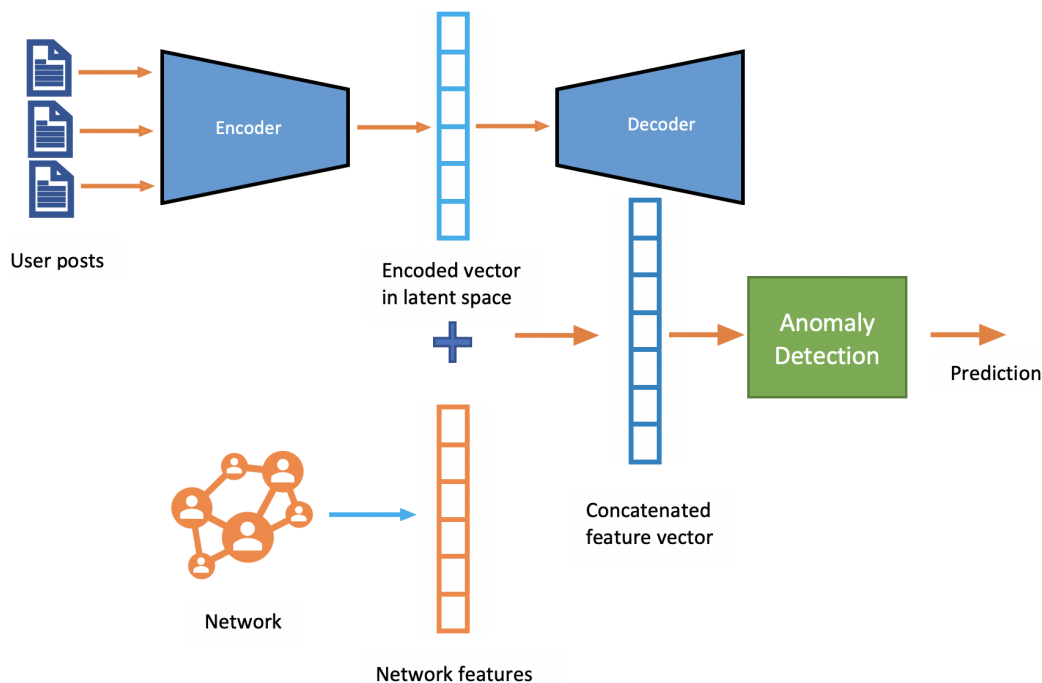


Figure D.2: Overview of the proposed unsupervised technique to identify depressed users in online forums.

D.4 Methodology

In this section, we describe our proposed methodology to identify depressed users in online forums. Given the scarcity of labeled users, we propose an unsupervised technique, as shown in Figure D.2. Given a user u , the first step is to compute a latent representation of u given the temporal sequence of posts they contributed to the forum. This user latent representation is learned in an unsupervised way by using a Long Short Term Memory (LSTM) autoencoder, as explained in Subsection D.4.1. Next, we consider how forum users interact among them. Thus, we build a “who replies to whom” network and compute network-based features for each user as described in Subsection D.4.2. These network features are concatenated to the user

latent representation extracted from their post sequence and then used in input to an anomaly detection algorithm to identify depressed users. The different unsupervised algorithms we used and compared to perform the anomaly detection task are detailed in Subsection D.4.3.

D.4.1 Unsupervised Learning of User Representation from Their Posts

We propose to compute, for each forum user u , a set of latent features given the sequence of comments $\langle p_1^u, \dots, p_n^u \rangle$ they posted in the forum. These latent features are learned by the stacked Long Short-Term Memory (LSTM) autoencoder shown in Figure D.3. An LSTM is a recurrent neural network [166] where each cell C has the architecture shown in Figure D.4 (adapted from [2]). Here, each LSTM cell outputs the next state h_t ($1 \leq t \leq n$) by taking in input the previous state h_{t-1} and the next vector x_t . The operations done by the single LSTM cell C are described by the following equations:

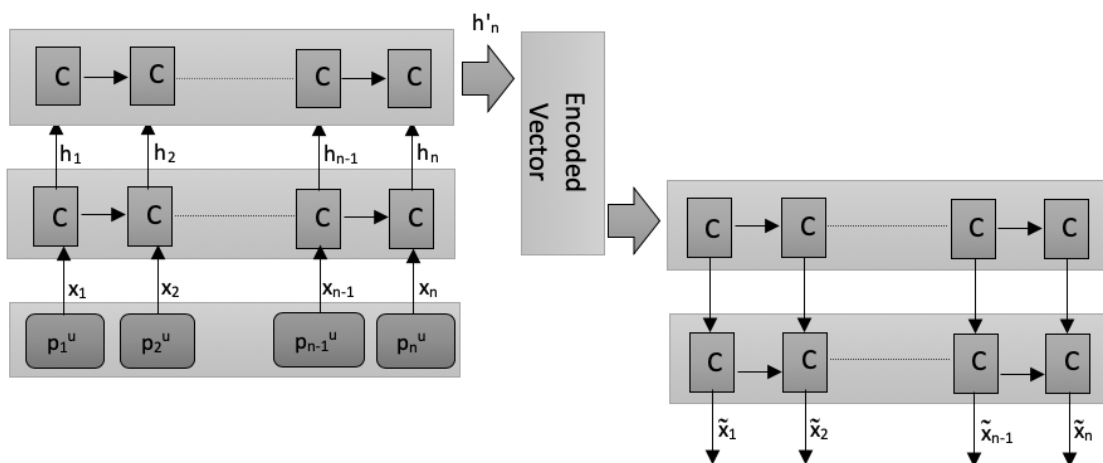


Figure D.3: Autoencoder architecture

$$a_t = \rho(W_a \cdot [h_{t-1}, x_t]) = 0 \quad (\text{D.1})$$

$$b_t = \rho(W_b \cdot [h_{t-1}, x_t]) \quad (\text{D.2})$$

$$y_t = \tanh(W_y \cdot [h_{t-1}, x_t]) \quad (\text{D.3})$$

$$g_t = \rho(W_g \cdot [h_{t-1}, x_t]) \quad (\text{D.4})$$

$$c_t = c_{t-1} \cdot a_t + b_t \cdot y_t \quad (\text{D.5})$$

$$h_t = \tanh(c_t) \cdot g_t \quad (\text{D.6})$$

where the W_a, W_b, W_y and W_g are the weights representing the LSTM cell C and the entire LSTM neural network.

The encoder part in Figure D.3 takes in input the sequence of user posts, where each post p_t^u is represented by using a vector x_t of linguistic features extracted from the post. As described in Subsection D.4.1, we used the LIWC linguistic features. The subsequence $\langle x_1, \dots, x_t \rangle$ is converted by the first LSTM into a single vector representation h_t of size k_1 . The second LSTM takes in input the vectors $\langle h_1, \dots, h_n \rangle$ from the first LSTM and further reduces their size to $k_2 < k_1$. The output h'_n of the last cell of the second LSTM gives the user u latent representation (or encoded vector) and represents the entire sequence of user u 's posts.

The decoder part takes in input h'_n and reconstructs the input autoencoder sequence $\langle x_1, \dots, x_t \rangle$ by using the inverse of the architecture used by the encoder. We used the root mean square error (RMSE) loss function that measures the error between the input sequence $\langle x_1, \dots, x_t \rangle$ and the reconstructed one $\langle \tilde{x}_1, \dots, \tilde{x}_t \rangle$. We train our LSTM autoencoder by considering all the users in our dataset (depressed,

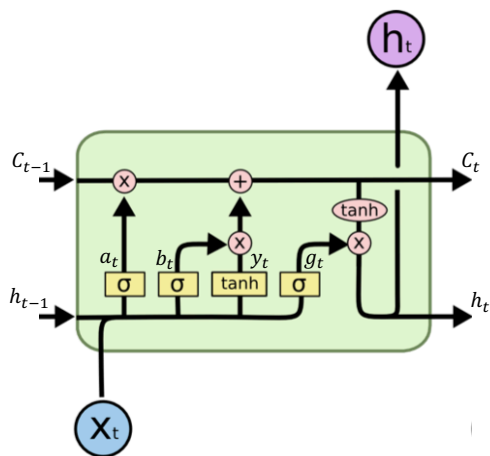


Figure D.4: Description of an LSTM cell C . Figure adapted from [2].

non-depressed, and unknown).

Psycho-Linguistic Features for Modeling User Posts

The linguistic style captures how language is used by individuals and provides information about their behavioral characteristics subject to their social environment. Language can be quantified to unveil clues about the underlying psychology of the individual. Thus, to represent each post p_i^u in input to the LSTM autoencoder described in the previous section, we compute a vector x_i of Linguistic Inquiry and Word Count (LIWC) features from the post text. LIWC is a transparent text analysis tool that counts words in psychologically meaningful categories. It reads text files in batches and counts the percentage of words that belong to each category, which can be grouped as Linguistic, Punctuation, Psychological, and Summary features [174].

Linguistics features refer to features that represent the functionality of text, such as the average number of words per sentence and the rate of misspelling. This category of features also includes negations as well as part-of-speech (Adjective, Noun, Verb, Conjunction) frequencies. There are a total of 28 features under this category.

Punctuation features are used to dramatize or sensationalize a post that can be analyzed through types of punctuation used in the posts such as Periods, Commas, Colons, Semicolons, Question marks, Exclamation marks, Dashes, Quotation marks, Apostrophes, Parentheses, and Other punctuation. There are a total of 11 features under this category.

Similarly, *psychological features* target emotional, social process, and cognitive processes. The affective processes (positive and negative emotions), social processes, cognitive processes, perceptual processes, biological processes, time orientations, relativity, personal concerns, and informal language (swear words, nonfluencies) can be used to scrutinize the emotional part of the posts. There are a total of 51 features under this category.

Summary features define the frequency of words that reflect the thoughts, perspective, and honesty of the writer. This category consists of features such as Analytical thinking, Clout, Authenticity, Emotional tone, Words per Sentence (WPS), Words with more than six letters, and Dictionary words. There are a total of 7 features under this category.

We used all the LIWC features for analyzing the cognitive, affective, and grammatical processes in the text, which helps in examining the difference between the writing style of posts among depressed and non-depressed users.

D.4.2 Network Features

Since most of the work in depression or mental illness detection via social media has been done by analyzing user posts (especially on the ReachOut forum [262, 263]), it would be interesting to analyze the users also from a networking point of view. Thus, to extract network-based features, we built a “who replies to whom” network

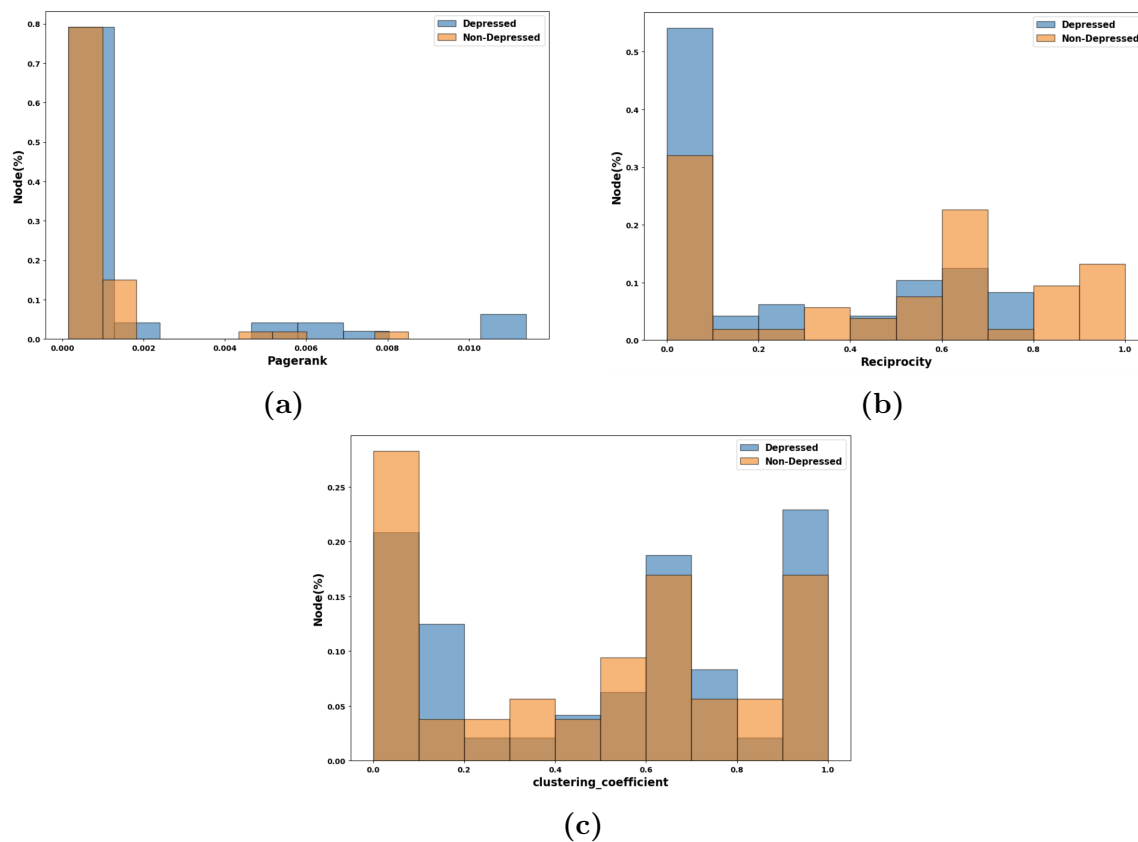


Figure D.5: PageRank (a), Reciprocity (b), and Local Clustering Coefficient (c) distribution of depressed (blue) and non-depressed (orange) users.

as follows. We considered each user as a node in the network and added an edge from node u to node v if u wrote a post in reply to v 's post. We denote this network as a directed graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. In this chapter, we use the following network features.

PageRank

The PageRank (PR) is a popularity measure for nodes in a network G and it is defined as

$$PR(u) = \frac{1 - \beta}{|V|} + \beta \sum_{v \in M(u)} \frac{PR(v)}{|L(v)|} \quad (\text{D.7})$$

where β is the damping factor usually set to 0.85 [283], $M(u)$ is set of nodes that link to u , and $L(v)$ is the set of nodes pointed by v .

Figure D.5(a) shows the distribution of PageRank for depressed and non-depressed users in our dataset. Here, we observe that depressed users tend to have a higher PageRank than non-depressed ones (0.0017 vs. 0.0008 on average).

Reciprocity

Reciprocity captures a basic way in which interaction on online sites takes place. When two users u and v interact, one expects that comments will be exchanged between them, i.e., users reply to each other. The reciprocity of a single node u is defined similarly. It is the ratio of the number of edges in both directions to the total number of edges involving the node u .

Figure D.5(b) shows the distribution of reciprocity for depressed and non-depressed users in our dataset. As we can see, depressed users tend to have lower reciprocity than non-depressed ones (average reciprocity of 0.24 vs. 0.46), meaning that depressed users tend to reply (or their posts are replied) less than non-depressed users.

Clustering Coefficient

It is observed that people who share connections in a social network tend to form clusters. The Local Clustering Coefficient (LCC) measures the probability that the neighbors of a node are connected and it is equal to

$$LCC(u) = \frac{2 \times |\{(v_1, v_2) \in E \mid v_1, v_2 \in \Gamma(u)\}|}{|\Gamma(u)| \times (|\Gamma(u)| - 1)} \quad (\text{D.8})$$

where $\Gamma(u) = M(u) \cup L(u)$.

Figure D.5(c) shows the distribution of the local clustering coefficient for depressed and non-depressed users in our dataset. We observe that depressed users have a higher local clustering coefficient value than non-depressed ones (average LCC of 0.52 vs. 0.47), meaning that depressed users' neighbors are more connected among them than the neighbors of non-depressed ones.

Node2Vec

Network embedding is a technique for mapping graph nodes in a geometric high dimensional space. Once the embedding is obtained for each node, its geometric representation can be used as features in input to machine learning algorithms. Node2Vec [284] is an embedding technique based on random walks. It computes the embedding in two steps. First, the context of a node (or neighborhood at a distance d) is approximated with biased-random walks of length d that provide a trade-off between breadth-first and depth-first graph searches. Second, the values of the embedding features for the node are computed by maximizing the likelihood of generating the context by the given node.

D.4.3 Anomaly Detection

Anomaly detection is the task of identifying the outlier or anomaly or the entity that does not comply with the normal behavior. The observation that significantly deviates from other observations is called an anomaly [285]. The task of anomaly detection is not limited, for instance, to finding suspicious behavior in networks (intrusion detection) or finance applications but this technique can be leveraged for uncovering rare events such as symptoms of a new disease or unusual symptoms and rare diseases [286]. Thus, we use anomaly detection in our work to identify depressed users by assuming that their behavior deviates from the one of normal users. As reported in Table D.1, we have scarce labeled data; thus, in this chapter, we apply unsupervised anomaly detection techniques ² to identify non-depressed (normal) and depressed (abnormal) users and use the available ground truth for evaluations purposes only. We apply and compare the following anomaly detection techniques.

Clustering-based Anomaly Detection Techniques. Clustering is an unsupervised machine learning technique that groups the observations into K clusters. Clustering can be used to performing anomaly detection under the assumption that “normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster.” [285] In this chapter we use K -means, Gaussian Mixture Model (GMM), and DBSCAN algorithms [287] to cluster the data. K -means and GMM have the problem of finding the optimal number of clusters and Gaussian components, respec-

²Unsupervised anomaly detection is used when the data is unlabelled, i.e., the class of an instance (normal or anomaly) is not known. This approach does not require the training or testing data, which makes it more flexible and widely applicable. The main idea of unsupervised anomaly detection is to provide a score for each instance by learning intrinsic properties such as distance or density. This score is called the anomaly score that determines whether the instance is normal or anomalous.

tively. In order to find these parameters, we used the Elbow method for K -means and the Bayesian Information Criterion (BIC) for GMM.

Once the observations are clustered by using K -means, we perform anomaly detection as follows. First, we compute the Euclidean distance between each data point and the centroid of the respective cluster. Secondly, we calculate the maximum cluster radius to identify the outliers. The maximum cluster radius determines the association of the data point to the particular cluster. For this, we use a percentile distance value as the threshold τ (i.e., for each cluster, the α th percentile of the distribution of all distances between data points and their respective centroid) to distinguish to which class the data point lies (if the distance is greater than τ , then we classify the user as depressed (anomaly), non depressed otherwise).

Regarding GMM, once the model is fitted, it provides the weighted log of probability density function values for each data point. This probability density function can be used to understand which sample belongs to which class. For this, we used the percentile of the weighted log probability distribution values as the threshold τ (i.e., the α st percentile of the distribution of all weighted log probability) in order to determine which class the data point lies: if the weighted log probability is less than τ then we classify the user as depressed (anomaly), non-depressed otherwise.

The DBSCAN algorithm is a density-based clustering algorithm that directly labels the data points as normal or anomaly, so no further steps are required to perform the anomaly detection task [288].

One-class classification based anomaly detection techniques. These techniques assume that all the data instances have only one class label. Under this category of algorithms, we used One-Class SVM in our work. The One-Class SVM

algorithm [289] learns the intrinsic properties of the normal cases (non-depressed users in our case) and uses these properties to understand which data point deviates from normal behavior (the known class). The data point that shows the abnormal behavior or that deviates from the normal are classified as anomalies (depressed users in our case) by the algorithm.

Ensemble-based anomaly detection techniques. We considered Isolation Forest under this category. The isolation forest algorithm is based on the fact that the anomalous observations are very rare and have different properties than the normal ones, and using these properties, the anomalous observations can be isolated from the normal ones in a more effective way. The basic idea of isolation forest is to separate each data point by randomly creating a separation line between the data point and the others. Since the anomalous data points are different than normal ones and are few in number and scattered, they can be segregated in a few numbers of splitting. Whereas, normal data points that are closer takes a significant number of splittings [290].

D.5 Experiments

This section reports on our experimental results of using linguistic and network-based features to identify depressed users in an unsupervised fashion.

D.5.1 Experimental Setting

Since our methodology is based on unsupervised anomaly detection, we compute the user representation from their posts and the network features by considering all the users in the dataset (depressed, non-depressed, and unknown). Next, anomaly detection is performed by using all the techniques presented in Section D.4.3, namely

K-means, Gaussian Mixture Model, DBSCAN, Isolation Forest, and One-class SVM. Once the users are labeled as normal (non-depressed) or abnormal (depressed) by any of the anomaly detection methods, we evaluate the prediction by using the ground truth available in our dataset: 65 depressed users and 94 non-depressed users. As evaluation measures, we use precision (Pr), recall (Re), and F1-measure (F1), similarly to related work.

Parameter Setting. The parameters of the algorithms used in our experimental evaluation have been set as follows. For Node2Vec, we set the number of features to 32, the random walk length to 20, and the number of walks to 100. For DBSCAN, we set *eps* to 0.1 and *min_samples* to 2. For One-Class SVM, we used RBF kernel with γ (kernel coefficient) equal to the inverse of the number of features and the ν parameter as the ratio of anomalous observations that we assume is present in the dataset. For isolation forest, we used 100 estimators and contamination as a proportion of outlier that we assume is present in the dataset. Finally, the autoencoder proposed in Section D.4.1 learns a user representation h'_n of size 32.

D.5.2 Baselines for comparison

We compare our proposed method from Section D.4 with the following network and linguistic based baselines:

- **PageRank + Reciprocity + Local Clustering Coefficient:** we perform the unsupervised anomaly detection task by considering these network features only.
- **Node2Vec:** we perform the unsupervised anomaly detection task by considering the Node2Vec features only.

Table D.2: Precision (Pr), Recall (Re), and F1-measure (F1) of anomaly detection with social network features, psycho-linguistic features and combination.

Features	K-means			GMM			DBSCAN			Isolation Forest			OC-SVM		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
PageRank + Reciprocity + Clustering Coeff.	0.41	0.54	0.43	0.56	0.56	0.56	0.49	0.12	0.16	0.59	0.60	0.52	0.61	0.59	0.59
Node2Vec	0.58	0.60	0.49	0.56	0.56	0.56	0.55	0.31	0.32	0.56	0.59	0.51	0.60	0.58	0.59
LIWC	0.67	0.62	0.51	0.55	0.55	0.55	0.39	0.27	0.23	0.62	0.62	0.54	0.58	0.56	0.56
Autoencoder	0.63	0.61	0.51	0.61	0.61	0.61	0.78	0.41	0.28	0.62	0.62	0.54	0.59	0.57	0.58
Proposed technique: Autoencoder + PageRank + Reciprocity + Clustering Coeff.	0.63	0.61	0.51	0.64	0.64	0.64	0.77	0.42	0.27	0.62	0.62	0.54	0.60	0.58	0.59
Autoencoder + Node2Vec	0.61	0.60	0.50	0.57	0.57	0.57	0.77	0.42	0.27	0.59	0.60	0.52	0.58	0.56	0.56
Autoencoder + All Network	0.58	0.60	0.49	0.58	0.58	0.58	0.77	0.42	0.27	0.62	0.62	0.54	0.60	0.58	0.58

- **LIWC**: for each user, we create a unique document by concatenating all their posts. Then, we compute the LIWC features of these documents (this is similar to the setting we had in our previous work [291]) and perform the unsupervised anomaly detection task with these features as input.

D.5.3 Results

Classification results are reported in Table D.2. When considering a singular modality (user post content or network), we have that our autoencoder based approach with the Gaussian Mixture Model (based on user posts only) provides the highest F1-score in comparison with the baselines PageRank + Reciprocity + Local Clustering Coefficient, Node2Vec and LIWC (cf. the first four rows of Table D.2). The 5% minimum improvement (w.r.t. network-based baselines such as Node2vec and PageRank + Reciprocity + Local Clustering Coefficient) is due to the fact that the autoencoder is able to (1) consider the temporal characteristics of the user’s mood

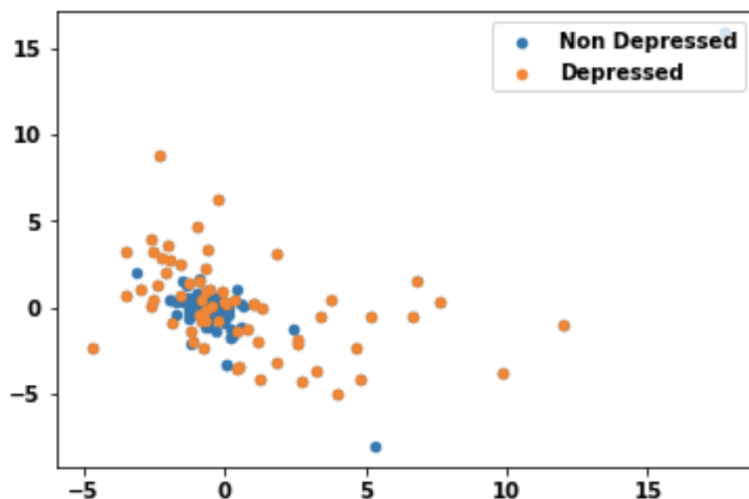


Figure D.6: Plot the user embeddings computed with the autoencoder and reduced in a 2-dimensional space via PCA.

through their posts (where the mood is computed by the LIWC features on each post), and (2) it is also able to recreate an embedding distribution that well pairs with standard anomaly detection techniques. Figure D.6 plots the user embeddings computed with the autoencoder and reduced in a 2-dimensional space via PCA. We clearly see that the majority of the orange points representing depressed users are far from the central cluster containing all the non-depressed users.

Our proposed methodology combines the user representation obtained by the autoencoder from the user post sequence with user network features. From the last three rows of Table D.2 we observe that PageRank + Reciprocity + Local Clustering Coefficient are better than Node2Vec features when combined with the autoencoder features. In fact, this combination provides a further 3% improvement in the F1-score (0.64) w.r.t the autoencoder features only. Please note that in the case of anomaly detection computed with DBSCAN, even if the precision for autoencoder is 0.78, the

recall is very low (0.41) by negatively impacting on the F1-score that is 0.28. For this motivation, we do not consider the DBSCAN as a good option.

The deep-learning-based approach we propose in this chapter is more complex (in terms of execution time) than traditional methods from Section D.5.2 we compare with. However, the additional complexity that exploits the temporal relationship among the user comments allows us to achieve better classification results than traditional methods. Moreover, once the model is trained, it can be used to infer the embedding for a new user without re-training (it is sufficient to pass in input the new user's sequence). Hence, our proposed approach can be applied to classify new users with the same complexity as traditional methods.

Qualitative analysis of unknown users. To further strengthen our experimental results, we considered the unknown users in our dataset, sorted them by the score provided by our proposed technique,³ and manually inspected the posts of the top-10 and bottom-10 unknown users. In the case of top-10 users, which are candidates for non-depressed users, we observed normal and positive comments similarly to common comments regarding travel or movies users post on social media: "Is that Castle House for real? I'd seen it before but assumed it was total CGI magic. I've been wasting.. err I mean spending a lot of time on Airbnb lately and have found so many places I want to travel largely because of how cool the accom is. Here is some igloo accom in Finland where you can see the Northern Lights through the ceiling!"; "My top five at the moment are: Game of Thrones...True Blood The Newsroom...Breaking Bad Also pretty excited about new season of Sons of Anarchy... Any fans of the above here?"

³The score is given by the weighed log probability obtained with GMM when this anomaly detection algorithm is applied to our autoencoder features plus PageRank, Reciprocity, and Local Clustering Coefficient network features.

Hence, we conclude these users do not show signs of depression.

Regarding the bottom-10 unknown users, which are candidates for depressed users, we found in the majority of them, at least one comment expressing discomfort. For instance, some comments were describing episodes where these people were crying without any reason: "when I'm alone again I get that 'empty' feeling and some days I feel kinda weak like I want to cry, and it's weird, this hasn't happened to me before." Other comments were about the fact that they noticed to be more aggressive than usual: "Hi sorry dumping All my problems on you again, but i have some, well, anger problems. Things piss me off easily and because i cant do vilance at school i take it out on my parents by yelling and fighting with them or crying for no reason."

This analysis confirms a good performance of our proposed technique also in the case of unlabeled users. In fact, crying without any reason and being more aggressive than usual are common symptoms of depression.

D.6 Conclusion

We addressed the problem of identifying depressed users in online forums in an unsupervised fashion. We analyzed user behavior in the ReachOut.com online forum by using psycho-linguistic features extracted from the sequence of user posts in combination with network-based features modeling how users connect in the forum. Our results showed the potential of these features in characterizing depressed users in online forums, especially user embedding extracted from user posts and network-based features such as Reciprocity and Local Clustering Coefficient. By combining both network and psycho-linguistic features, our proposed unsupervised approach achieved an F1-measure of 0.64 in detecting depressed users and performed better than baselines.

Future work will be devoted to (1) extending our results to the problem of early

detection of depressed users in online forums, and (2) exploit depressed user detection techniques to enhance risky post detection by including author network features.