PREDICTING FRIENDSHIP STRENGTH IN FACEBOOK

by

Nitish Dhakal

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

May 2017

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Nitish Dhakal

Thesis Title: Predicting Friendship Strength in Facebook

Date of Final Oral Examination: 7 March 2017

The following individuals read and discussed the thesis submitted by student Nitish Dhakal, 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.

Dianxiang Xu, Ph.D.	Chair, Supervisory Committee
Francesca Spezzano, Ph.D.	Co-Chair, Supervisory Committee
Edoardo Serra, Ph.D.	Member, Supervisory Committee

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

DEDICATION

This thesis work is affectionately dedicated to my loving family and my beloved wife. I have only come this far with their love and support.

ACKNOWLEDGEMENTS

I would not have come this far without the support and guidance of many individuals. I would like to express my heartiest gratitude to my wonderful advisor Dr. Dianxiang Xu. His proper guidance, timely suggestions, patience, prompt motivation and persisting help make this work possible. I would also like to thank my Co-Chair Dr. Francesca Spezzano for her anytime support and invaluable feedback and a keen interest in every stage of my research. I am also indebted to my committee member and instructor Dr. Edoardo Serra for his expert advice in my thesis work and his effort in teaching us data science. The seed of this dissertation was planted in his data science class.

I am also grateful to the department of computer science at Boise State University for providing me a full scholarship for my graduate studies. It would not have been possible to continue my graduate studies without the support from the department and most importantly my advisor Dr. Dianxiang Xu. I would also like to thank my colleagues Milson Munakami, Shuai Peng, Samer Khamaiseh and Dr.Yan Wu for their positive encouragements throughout the thesis work. Furthermore, I would also like to thank the participants of the survey for willingly sharing their invaluable time.

My special and sincere thanks go to my lovely wife and my entire family for their never-ending love and support for me in all my actions.

ABSTRACT

Effective friend classification in Online Social Networks (OSN) has many benefits in privacy. Anything posted by the user in social networks like Facebook is distributed among all their friends. Although the user can select the manual option for their post-dissemination, it is not feasible every time. Since not all friends are the same in a social network, the visibility access for the post should be different for different strengths of friendship for privacy. We propose a model with 24 features for finding friendship strength in a social network like Facebook. Previous works in finding friendship strength in social networks have used interaction and similarity based features but none of them has considered using linguistic features as the driving factor to determine the strength. In this paper, we developed a supervised friendship strength model to estimate the friendship strength based upon 24 different features comprising of structure based, interaction based, homophily based and linguistic-based features. We evaluated our approach using a real-world Facebook dataset that has 680 user-friend pairs and obtained accuracy of 85% across close and acquaintance friend classification. Our experiments suggest that features like *average comment length; likes, love, friend posts, mutual friends and closeness variable* consistently perform better in predicting friendship strength across different classifiers. In addition, combining language-based features with homophilic, structural and interaction features produces more accurate and trustworthy models to evaluate friendship strength.

TABLE OF CONTENTS

DEDICATION iv

ACKNOWLEDGEMENTS v

ABSTRACT vi

LIST OF TABLES xi

LIST OF FIGURES xiii

LIST OF ABBREVIATIONS xvi

CHAPTER ONE: INTRODUCTION 1

Background 1

Thesis Statement 5

Outline 6

CHAPTER TWO: RELATED WORK 8

Predicting Friendship Strength 8

Trust and Privacy in Social Network 10

Bag of Words Model (BOWM) 13

Sentiment Analysis 15

CHAPTER THREE: FRIENDSHIP STRENGTH MODEL 17

General Interaction Features 18

Linguistic Features 20

Comments Polarity 20

Contradiction Rank 21
Agreement Rank 22
Closeness Variable 22
Structural Features 24
Homophilic Features 25
CHAPTER FOUR: EXPERIMENTS 29
Datasets 29
Choosing Users 30
Fetching User Friends 31
Fetching Individual Posts 31
Fetching User Profiles and Profile Detail Information 31
Fetching User Interests and Groups 31
User Survey 32
Limitation in Datasets 34
Feature Selection 34
Pearson Correlation Coefficient 34
Recursive Feature Elimination (RFE) 38
Ensemble of Decision Trees for Feature Importance 38
Learning Curve 39
Data Subset Selection 43
Models to be Evaluated 46
All Features Only 47
Bag of Words Model (BOWM) on Comments 47

All Features Combined with BOWM 48
Feature Subset 1 49
Feature Subset 1 + BOWM 49
Linguistic Features 49
Feature Subset 2 49
Feature Subset 3 50
Feature Subset 4 50
Train and Test Sets 50
Results 50
Close and Good Vs Acquaintances and Causal 51
Close, Good and Causal Vs Acquaintances 56
Close Vs Acquaintance 61
Close Vs Acquaintance and Casual 65
Only Interaction Features 69
Linguistic Vs Other Features 71
Analysis 74
Limitation 77
CHAPTER FIVE: CONCLUSION AND FUTURE WORK 78
REFERENCES 80
APPENDIX A 83
Evaluation Metrics 83
APPENDIX B 88
Topic Modelling 88

APPENDIX C 94

Institutional Review Board 94

LIST OF TABLES

Table 1. Bag of Words Model 13

Table 2. Percentage of occurrence of the words among different friend categories 23

Table 3. Facebook user profiles URL 32

Table 4. Pearson correlation coefficient between vector of 24 features and output class for different combination of datasets 36

Table 5. Top five features according to RFE 38

Table 6. Top six features according to an extra trees ensemble 39

Table 7. Combined feature set of bag of words along with additional features 49

Table 8. Performance comparison of different regularization parameters for 9 different feature sets in “CL+GO Vs AC+CA” classification using logistic regression 53

Table 9. Comparison between classifiers for “feature subset 3” in “CL+GO Vs AC+CA” classification 55

Table 10. Performance comparison of different regularization parameters for 9 different feature sets in “CL+GO+CA Vs AC” classification using logistic regression 58

Table 11. Comparison between classifiers for “feature subset 1” in “CL+GO+CA Vs AC” classification 60

Table 12. Performance comparison of different regularization parameters for 9 different feature sets in “CL Vs AC” classification using logistic regression 62

Table 13. Comparison between classifiers for “feature subset 1” in “CL Vs AC” classification 64

Table 14. Performance comparison of different regularization parameters for nine different feature sets in “CL Vs AC+CA” classification using logistic regression 66

Table 15. Comparison between classifiers for “feature subset 1” in “CL Vs AC+CA”classification 68

Table 16. Confusion matrix 86

Table 17. Example of Topic Modelling of comments across different friend categories 90

LIST OF FIGURES

Figure 1. Friendship degree as the sum of interactions 3

Figure 2. Screenshot of interface to write and share posts in Facebook 4

Figure 3. Average normalized score of Facebook reactions among different categories of friends 19

Figure 4. Average comment length among different categories of friends 20

Figure 5. Average structural score among different categories of friends 24

Figure 6. Average similarity score between homophilic attributes among different categories of friends 28

Figure 7. Average percentage of occurrence of different categories of friends 30

Figure 8. Example of the interface implemented for acquiring the friend classification ground 33

Figure 9. Pairwise correlation coefficient between all the feature variables including class 37

Figure 10. Learning curve obtained using homophilic features 40

Figure 11. Learning curve obtained using linguistic features 41

Figure 12. Learning curve obtained using important features 41

Figure 13. Learning curve obtained using complete set of features 42

Figure 14. 2D scatter plot of 24 features showing all friend class pairs 44

Figure 15. 2D scatter plot of 24 features showing combination of close and good friends pair vs combination of casual and acquaintance friend pair 44

Figure 16. 2D scatter plot of 24 features showing combination of close, good and casual friend pair vs acquaintance friend pair 45

Figure 17. 2D scatter plot of 24 features showing only close and acquaintance friend pair 45

Figure 18. ROC curve of “feature Subset 3” in “CL+GO Vs AC+CA” classification using logistic regression..... 54

Figure 19. ROC curve of “feature Subset 3” in “CL+GO Vs AC+CA” classification using the random forest..... 55

Figure 20. ROC curve of “feature Subset 1” in “CL+GO+CA Vs AC” classification using logistic regression..... 59

Figure 21. ROC curve of “feature Subset 1” in “CL+GO+CA Vs AC” classification using the random forest..... 60

Figure 22. ROC curve of “feature Subset 1” in “CL Vs AC” classification using logistic regression 63

Figure 23. ROC curve of “feature Subset 1” in “CL Vs AC” classification using random forest 64

Figure 24. ROC curve of “feature Subset 1” in “CL Vs AC+CA” classification using logistic regression 67

Figure 25. ROC curve of “feature Subset 1” in “CL Vs AC+CA” classification using random forest 68

Figure 26. ROC curve in “CL+GO Vs AC+CA” classification with interaction features only..... 70

Figure 27. ROC curve in “CL+GO+CA Vs AC” classification with interaction features only..... 70

Figure 28. ROC curve in “CL Vs AC” classification with interaction features only. 71

Figure 29. ROC curve in “CL Vs AC +CA” classification with interaction features only 71

Figure 30. Comparison between ROC curves across “CL+GO Vs AC+CA” classification with and without linguistic features..... 73

Figure 31. Comparison between ROC curves across “CL+GO+CA Vs AC” classification with and without linguistic features..... 73

Figure 32. Comparison between ROC curves across “CL Vs AC” classification with and without linguistic features..... 74

Figure 33. Comparison between ROC curves across “CL Vs AC+CA” classification with and without linguistic features 74

Figure 34. Average sentiment intensity among different categories of friend 76

Figure 35. Sample ROC Curve 84

Figure 36. Inter topic distance map showing distribution of 50 different topics 91

Figure 37. Top-30 relevant terms for topic 1 identified as close topic 92

Figure 38. Top 30 relevant terms for topic 2 identified as a political topic. 93

LIST OF ABBREVIATIONS

OSN	Online Social Networks
BOWM	Bag of Words Model
TF	Term Frequency
IDF	Inverse Document Frequency
NLTK	Natural Language Toolkit
VADER	Valence Aware Dictionary and Sentiment Reasoner
FSM	Friendship Strength Model
SS	Structural Score
FTS	Friend Tag Score
UFTS	User Friend Tag Score
CPS	Comment Polarity Score
CR	Contradiction Rank
AR	Agreement Rank
IRB	Institutional Review Board
CL	Close
CA	Casual
AC	Acquaintance
GO	Good
RFE	Recursive Feature Elimination
PCA	Principal Component Analysis

SVM	Support Vector Machine
ROC	Receiver Operating Characteristic Curve
AUROC	Area under Receiver Operating Characteristic Curve
TPR	True Positive Rate
FPR	False Positive Rate
TNR	True Negative Rate
PPV	Positive Predictive Value
LDA	Latent Dirichlet al.location

CHAPTER ONE: INTRODUCTION

Background

Online Social Networks (OSN) are part of individual lives today. People are attached to many social networking sites such as Facebook, Twitter and LinkedIn to express their opinions, preferences, pleasure and experiences. In each of these social networks, an individual may have many friends with different friendship strengths [2] . This means that an individual's friendship network contains both strong and weak relationships [26] . Since it is difficult to put both kinds of relationship into the same category, we need a mechanism to differentiate between friends of weak strength and friends of strong strength. In OSN, the users can differentiate weak and strong friendships by creating their own virtual social circle [1] . Facebook offers services to categorize friends as acquaintances, family or spouses. Google Plus also allows users to create a desired circle. However, in both of these social networks, it is up to the users to differentiate their friends and group them into different circles.

It is natural that people try to maintain relationships by interacting with only those friends who they consider important. In Facebook, posting on a friend's wall, commenting or liking his/her posts are the most common modes of interactions. With the introduction of new Facebook reactions such as the love emoticon, ha-ha emoticon, wow emoticon, sad emoticon, and angry emoticon, people have nonverbal options to express their opinions. Among all the friends of an individual, it is highly likely that these kinds of interactions occur most frequently between friends with stronger ties. Consider a

Facebook user with a friend list of 800. Among them, there are only a few of number of friends with whom the user interacts on a regular basis and hence has a trustworthy/close personal relationship. There are friends who do not interact with the user at all. Studies [29] [30] suggest that an individual in a social network is tightly connected to a small set of friends and is loosely connected to a large group of friends. It is not beneficial for users to share information with a group of friends who they do not interact with at all [9] . People who participated in our survey (Chapter 4) mentioned this many times when they failed to notice the "friends" they have added on Facebook. They even unfriended the people later due to privacy concerns. This showed that people add friends in social networks without even knowing them. Due to the open nature and popularity of OSNs, users are getting more and more concerned towards their privacy [14] .

This thesis work tries to model these strongly and weakly bonded user-friends and rank them according to the four friendship categories named as casual, acquaintance, good and close for convenient automatic privacy assessment. In this study, we define close friends as the ones whom you trust in real life, share information with or someone with whom you are comfortable. In other words, close friends are the best friends. Similarly, we define good friends as the ones who are not among the closest ones but with whom you share a good bond. In the same way, acquaintances are the friends whom you have some familiarity with but not a personal relationship. They might be from your workspaces or from your schools. Finally, we define casual friends as the ones whom you have just met on the social network and do not have much information about the user. These four friendship categories represent the degrees of friendship.

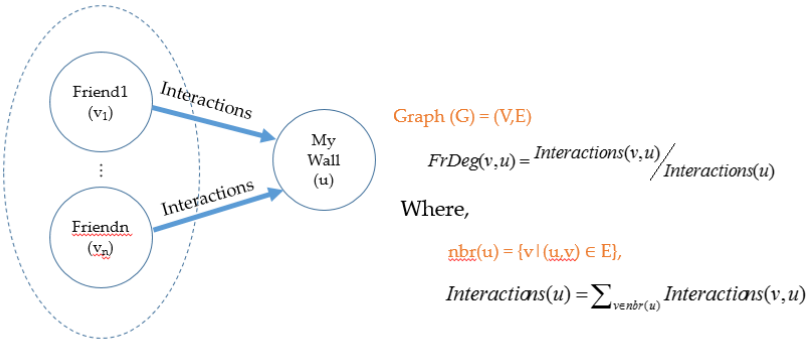


Figure 1. Friendship degree as the sum of interactions

In order to distinguish friends of different strengths, we first examined the total interactions made by all the friends *v* with the user *u*. Figure 1 shows a simple model where the sum of all the interactions determines the weight of friendship. The interaction here represents reactions, comments, posts, user friend tags etc. exchanged between a user and friends. It is worth noting that the interaction may be positive or negative. The polarity of the interactions between the user and a friend is determined by conducting linguistic analysis on the comments exchanged between friends with the user. Then, we evaluated the strength of the connection between the user and their friends. Existing research has modeled friendship strength in social media using interactions and similarity data [1] [4] [7] [9] [23] [24] [25] . These studies have used interaction data exchanged between the user and their friends to predict the relationship strength [27] [28] However, none of them has considered linguistic analysis for determining friendship strength. Previous studies conducted in social networks has also discovered that friendships can also be predicted from the network they are embedded in [8] . Therefore, besides linguistic features, we also take advantage of the network structural features shared between the users and their friends. Thus, this research utilizes previously used

interactions, similarity and structural features and combines them with linguistic features adding more reliability in predicting friendship strength.

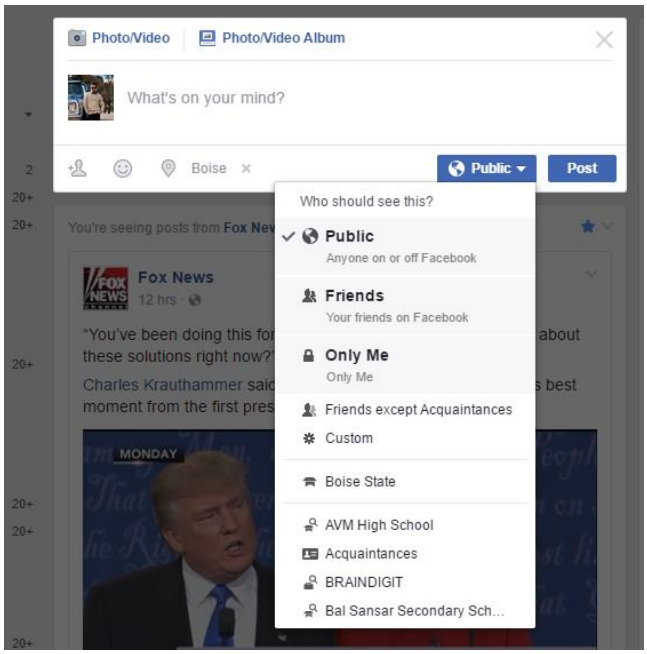


Figure 2. Screenshot of interface to write and share posts in Facebook

Effective means of automatic friends' classification has many benefits in privacy. Currently whenever the user posts in Facebook, it is disseminated to all of the user's friends (named as public). If the user is concerned about the privacy of the post, the user has the option to select the friends manually or post it to desired groups the user has already created. This is shown in Figure 2. In real life, the visibility access for the post should be different for close, good, casual and acquaintances. Consider a family photo posted by a user in Facebook. He would like his trustable friends to view and interact with the post. Consider another post, which is a general trending topic from the same user. In that case, he would want his casual friends to interact more. Therefore, if a system performs privacy assessment automatically then it would be easier from the user's perspective. Based upon the effectiveness and context of the model, the user could be

given a warning or could be given the option to set the visibility level before posting it on the social network.

Hence, to facilitate this automatic privacy assessment, we developed a model of friendship strength based upon 24 interactions, linguistic, structural and homophilic features. The complete list of the features is explained in Chapter 3. We considered 4 classes of friendship strength, i.e. casual, acquaintance, good and close. Then to test the model, we built a data set constructed from 34 users and their 680 friends. The dataset we obtained has 563 friends in average per user and 242 friends in average interacting with the user through comments, like, tags etc.

With the limitation on the number of users and their friends in our datasets, we combined different classes of a friend (close, good, casual and acquaintance) under two classes (weak and strong), treating this classification as a binary problem. Surprisingly, with the selected subset of features, our model predicted the friendship strength with an accuracy of about 85%.

Thesis Statement

In this thesis work, we developed a friendship strength model that is a function of interactions, linguistic, structural and homophilic variables. This model combines the existing interaction and similarities features with the new set of linguistic features generated from different existing literatures [2] , and used them to fit into different classifiers. The model seeks to find the features that work best in classifying strong relationships from weak ones on Facebook. Our main contribution in this thesis is threefold. Firstly, the utilization of linguistic features in determining friendship strength. Our study was the first one to initiate the use of linguistic features to help calculate

weight of friendship. Second, we studied friendship strength across our own self-created friendship categories. Third, we extracted a dataset of 680 user-friend pairs in Facebook with ground truths.

To evaluate our model we addressed two important research questions.

Q1. Are interactions good estimators of friendship strength?

Existing studies [9] [11] have shown that interactions alone have the ability to detect strong friends and weak friends. However, since we believe that not all friends are equal we wanted to see how these interaction variables treat different classes of friends, i.e. casual, acquaintance, good and close friends.

Q2. How does the linguistic features relate to the different friend categories?

We believe that sentiments in comments across users and their friends could be a factor in determining the friendship strength. For example, close friends tend to use more emoticons than acquaintances or casual friends. Similarly, there are specific words like ‘love you’, ‘miss you’, ‘honey’ etc. frequently used by close friends. Therefore, our goal was to extend our first research question by evaluating the friendship strength between users and their friends based on friend’s comments. We used sentiment analysis and other linguistic features to find the strength of comments. Using linguistic features to discuss the friendship strength is our novelty.

Outline

The paper is organized as follows. Chapter 2 discusses the related works in the detection of friendship strength, trust and privacy in social networks, a brief introduction on the bag of words model and the background on sentiment analysis. Chapter 3 formulates the friendship strength detection problem along with the brief explanation of

different features. Chapter 4 delves into the experiment section. This section contains information about Facebook datasets extraction, user survey, feature selection, data subset selection and models to be evaluated for detecting friendship strength. In addition, this section also includes experimental results obtained by combining different aspects of the datasets, analysis and limitation of this thesis work. Chapter 5 summarizes our research with a brief conclusion and discusses the possible future direction of this thesis work. This thesis paper also consists of three appendices. Appendix A discusses the evaluation metrics used in the research. Appendix B discusses topic modeling of Facebook comments in brief and finally Appendix C includes the IRB approval letter.

CHAPTER TWO: RELATED WORK

Predicting Friendship Strength

Finding a tie strength in social networks is not a new task. However, the use of linguistic features is a novel task. In 2009, Gilbert et.al. [4] tried to cover various aspects of relationships in the social networks; however, their model did not consider linguistic analysis as the emotional factor driving the friendship strength. Their model considered about 70 numerical indicators describing the tie strength among friends on Facebook on the dataset of about 2000 Facebook friends and classified with 85% accuracy. Besides the linguistic analysis, our model is different from them in terms of social distance variables (like occupation, education, political view). They have designed their model by taking the differences among these variables (between the user and a friend) as a feature but we assume the similarity between the variables as features. Since differences were more likely to happen, we thought similarity would make our feature strong.

Similarly, work done by Syed et.al. [7] tried to seek and rank influential friends. Zhang et.al [24] studied tie strength in mobile communication networks by taking reciprocity of calls between the users under consideration. Onnela et.al [23] also studied tie strength in mobile communication networks by considering network structure. There has also been a study on finding tie strength in musical social networks based upon the similarities of musical tastes in Last.fm [25] .

Likewise, there have been research projects in the past where data-mining techniques have been utilized to find a strong group of friends [18] , popular friends [18]

and significant group of friends across multiple domains in social networks [20] . These research projects take a number of messages posted by users in Facebook into account but do not consider other interactions among the user and their friends.

Mustafa et.al [1] did a similar project to our project but tried to generate friend rankings based upon the user interaction on mobile phones. They considered calls, voice messages, and chats as the factors of interaction. They have ranked the user friends by using a "Sports Ranking Algorithm" by giving weights to different types of interactions that happen in user's phones. Their research and this paper have similar problem formulation while giving weights.

Xiang et.al [9] research on Online Social Networks is similar to this paper in terms of the goal of distinguishing strong friends from weak ones in social media. They used an unsupervised approach (rather than our supervised approach) using a latent variable model to infer (hidden) relationship strengths. However, they did not consider polarity of interaction as a feature. Research done by West et.al [8] used both network structure and linguistic sentiment analysis to exploit social network structure. They tried to predict user A's relationship with B by taking both the network information and sentiment analysis of the evaluative text relating A to B. Their work and ours' is different in the sense that they tried to predict user A's opinion of B, and was more focused on opinion analysis rather than friendship strength.

Our work on finding tie strength closely relates with Jones et.al [9] where the researchers asked the number of individuals about who their close friends are in real life. Based on the interaction in between the survey users and their real world friends (ground truth) done on Facebook, they could successfully discriminate close friends from not

close friends. They achieved the area under ROC of 0.92 between strong and weak friends. They pointed to *media multiplexity* as the primary reason for their success. *Media multiplexity* assumes that if two people interact in one social media then they interact equally well in another medium as well. However, they limited their work in just finding closest friends among others. In addition, their study also lacked linguistic analysis.

Another work close to ours is the work done by Arnaboldi et.al [11] where the researchers analyzed the user friend relations among 28 users and their 7103 relationships. They did a survey like ours and asked the participants to rate their Facebook friends on a 0 to 100 score. Their model is limited to regression analysis of the friendship strength and statistical analysis of datasets.

Trust and Privacy in Social Network

The Cambridge dictionary describes trust as having confidence in something, or to believe in someone. Singh, S., & Bawa, S. (n.d.) [15] defined trust as the measure of confidence that a user would behave in a certain manner. This definition of trust is applicable to social networks. However, trust has different definitions in different contexts. Trust has been studied in different disciplines like psychology, economics, and sociology and computer science [14] . Specifically, in computer science, trust is classified as user and system trust. We will not discuss system trust here. Our research is applicable to online social networks hence it relates to user trust. According to Sherchan et.al [14] , user trust is inherently personalized and relational. This is because, in online social networks, when two users frequently interact with each other, their relationship becomes strong and it gradually evolves with experience. We selected different interaction, linguistic and structural variables based upon this notion of trust.

According to Sherchan et.al [14] , trust has different types. Following are the ones that relate to our research.

i. Calculative

This kind of trust assumes that a person would trust someone if the chances of gain are higher than the chances of loss. In online social networks, people try to establish this of kind of trust for private motives. User A would try to befriend with user B in social networks and show a high amount of interaction to get user B‘s attention and trust.

ii. Relational

This kind of trust assumes that trust develops incrementally over the interactions made between the trustor and the trustee. In a computer science perspective, this kind of trust where two parties interact gives rise to the ultimate trust called direct trust. Our research considers different interactions to find the relational trust among user friends on the social networks.

iii. Emotional

If an individual feels the notion of security and comfort with the trustee then it gives rise to emotional trust. Emotional trust makes the trustor have positive emotions towards the trustee. In an online social network, this kind of trust is seen in reactions and comments posted by the friends to the user. If user A has emotional trust with friend B, then user A expresses their emotional trust in various accounts of their interaction history in the social network either using emoticons or writing positive comments to express extreme love and affection.

Different online social networks have utilized the concept of trust in the past. Golbeck’s [16] work describes trust as non-transitive, personalized, unidirectional and

asymmetric in her research that involves trust in web-based social networks. In social networks, trust is not transferrable. If person A trusts person B and person B trusts person C then it is not necessary that person C also trust person A or vice versa. In addition, Person A trusting person B is entirely personal. The linguistic feature used in this thesis such as *contradiction rank* and *agreement rank* (see Linguistic Features) considers this.

In another paper, Golbeck [17] developed a recommendation system model called *FilmTrust* by utilizing trust among users as a principle factor in the algorithm. The trust among friends is generated from their interactions. In our model, we have used different facets of interactions and relationships between users and their friends to determine the trust.

Izzat et.al [13] have done research on trust and reputation models for social networks. Their work concentrated on finding a reputation score based on the personal and relational attributes of the user. Personal attributes include work, education, interest groups, favorites etc. Relational attributes include personal activities like quality and quantity of interactions with other friends. Relational attributes also contain characteristics associated with the individual's friends. These characteristics may be a number of interactions made by the user or their individual reputation level. Their proposed reputation model aims at preserving the user's privacy and is adaptive to the changes in a user's reputation score over time. Although the application of our model and their model seems to be the same, we have developed a machine-learning model with different classes of friends derived from ground truth. Our main works revolve around the model using different aspect of interactions, linguistic, structural and homophilic

features. One of the applications of our model is privacy assessment as explained in Chapter 1.

Bag of Words Model (BOWM)

To use sentences as a feature vector in our research, we have used a Bag of Words[1] Model (BOWM) approach. Bag of Words Model is a model representation used commonly in Natural Language Processing and Information Retrieval where the frequency of the occurrence of words is considered as a feature in training the classifier. In BOWM, each sentence is called a document and each unique word is called a term. The basic BOWM with n documents and m unique terms is shown in Table 1. For each term t in the column space, $f_{i,j}$ is the frequency of the unique term j in document i.

Table 1. Bag of Words Model

	t_1	t_2	...	t_m
d_1	$f_{1,1}$	$f_{1,2}$	...	$f_{1,m}$
d_2	$f_{2,1}$	$f_{2,2}$		$f_{2,m}$
:	:	:		:
:	:	:		:
d_n	$f_{n,1}$	$f_{n,2}$	...	$f_{n,m}$

In a BOWM model, the higher frequency represents higher relevancy of the term in the document while lower frequency represents lower relevance with the document. The lower frequency would also mean that the word has low significance or discriminative influence in the document. Although it is a naïve approach, BOWM is widely used and proven effective in many instances in classification. Despite its

[1] https://en.wikipedia.org/wiki/Bag-of-words_model

popularity, inability to consider ordering of the words and inability to consider the word semantics in the sentence is its major drawback [12] .

Therefore, instead of using just frequency of the terms as a feature in BOWM we use inverse document frequency. To do this $f_{i,j}$ of BOWM model is normalized. Normalization is carried out by dividing each frequency $f_{i,j}$ with the maximum frequency of the term i.e. max {fi} inside a document.

i.e. Normalized matrix $(tf_{i,j}) = \frac{f\ i,j}{\max\{fi\}}$ (1)

Then we calculate the document frequency matrix df_j,

Where df_j = Maximum number of the document containing each unique terms. Next step involves the computation of inverse document frequency idf.

Inverse document frequency (idf) = $\ln(\frac{N}{dfj})$ (2)

Where N is the total number of document and ln is the natural logarithm. The weighted matrix (tf-idf) is generated from the inverse document frequency matrix (2) and the normalized Matrix (1) as

Weighted matrix (tf-idf) $W_{i,j} = tf_{i,j} * idf$

$= tf_{i,j} * \ln(\frac{N}{dfj})$ (3)

These models are available in NLTK[2] library in python. NLTK library in python is a very good tool for Natural Language Processing. We have used NLTK for classification, tagging and stemming of the documents in our data.

[2] http://www.nltk.org/

Sentiment Analysis

Sentiment Analysis or opinion mining is a budding field of Natural Language processing which analyzes the sentiment, opinions, attitudes and emotions through the computational treatment of subjectivity in the text [5] . Due to the growth of social media, blogs, microblogs, discussion forums there is a huge chunk of opinion data available. It is very important to evaluate whether the sentiment expressed by people in different fields is negative or positive. Determining such polarity has different applications in multiple disciplines.

Different existing research has used lexicon-based features for sentiment analysis task [31] [32] [33] [34] . Building a lexicon requires a substantial amount of time and research. This is the main reason behind many researchers using the same lexicons in their research. Sentiment analysis in social media is different from other contexts. Use of abbreviation, shorter texts, and contextual level sentences make it difficult to analyze the social media texts than others. Words like “lol”, “BRB”,”WTH” etc. are quite widely used in social media. Standard sentiment lexicons does not provide such words and their polarity. In order to include these features, we need to take advantage of the lexicons that include the words frequently used in social media. Out of different techniques available for sentiment classification, we found VADER [5] to be quite effective in determining sentiments from social media context.

VADER is a parsimonious rule-based model, which works effectively with the social media texts and proves to be effective with slangs and acronyms very well. VADER is open source[3] and can be installed as a python package. The main reason for

[3] https://github.com/cjhutto/vaderSentiment

choosing VADER is its use of social media lexicons and mixture of ground truth from tweets, NY Times editorials, movie reviews, and Amazon reviews.

For a sentence or comment, VADER gives a negative, positive, neutral and compound score. For example, consider these 3 sentences in our datasets:

1. I miss you kid brother!
2. Such a great guy!!!
3. Great rewards, well earned!

For sentence 1, VADER outputs "*{'neg': 0.387, 'neu': 0.613, 'pos': 0, 'compound': -0.2244}"*. Where, *neg*, *neu*, *pos* and *compound* means negative, neutral, positive and compound respectively. For sentence 2, VADER outputs *"{'neg': 0.0, 'neu': 0. 287, 'pos': 0. 713, 'compound': 0. 7163}"*. Similarly, for sentence 3 VADER outputs "*{'neg': 0.0, 'neu': 0.094, 'pos': 0.906, 'compound': 0.8622}"*.

Sentence 1 has a negative sentiment, sentence 2 has a positive sentiment and sentence 3 has an even greater positive sentiment. From the scores itself the VADER library clearly detects these sentiments. The compound score gives the overall sentiment of the comment. The advantage of this score is its range. The range of the compound score is between -1 to 1. So we can easily change the range as per our need to accommodate extremely negative (-1 to -0.5), negative (-0.5 to -0.1), neutral (-0.1 to 0.1), positive (0.1 to 0.5) and extremely positive (0.5 to 1).

CHAPTER THREE: FRIENDSHIP STRENGTH MODEL

The interaction, linguistic, structural and similarity based data obtained as different features from Facebook has a different effect on determining whether the friend is close or just casual. For example, user A commenting positively to user B has more weights than user C who likes a post in user B's wall on Facebook. To encapsulate this, we suggest providing different weights to the interaction data assuming that different interactions have different effects. The model we use can be defined as the function of four different variables shown below.

$$\text{Friendship Strength Model } (FSM_{A,B}) = f(It, L, St, H) \quad (4)$$

Where,

It = General Interactions

L = Linguistic features like net comment polarity, contradiction rank, agreement rank and closeness variable

St = Structural feature like number of mutual friends

H = Homophily features

These four variables have different weights. These weights are determined by supervised learning methods, which are discussed in detail in the experiments section in chapter 4. Depending upon the nature of the social network these weights might be different with the network under consideration [1] .

General Interaction Features

Whether the user has strong or weak friends, they react to the posts that users share on their walls. It is the most common mode of friend interaction in Facebook. The reaction includes nonverbal ways of expressing the opinions like giving *likes*, *love*, *wow, haha, sad or angry* emoticons. Prior to this research, there have been no studies in "Facebook reactions". Higher numbers of "Facebook reactions" was assumed to signify higher intensity of friendship between the user and their friend. It is highly likely that weak friends have a weak level of such interactions.

Therefore, we take a number of such "Facebook reactions" as a set of features with different feature weights. From the dataset, we found that close friends usually tend to give more *love* emoticons and very few give *angry* emoticons, while *likes* were seen as the most common mode of interaction between all categories of friends. Therefore, we thought it was essential to give each of these reactions different weights. Figure 3 shows the average normalized score of different Facebook reactions across different categories of friends. From the figure, we see that reactions like *loves, likes, haha* and *sad* are high among close and good friends and low among casual friends and acquaintances.

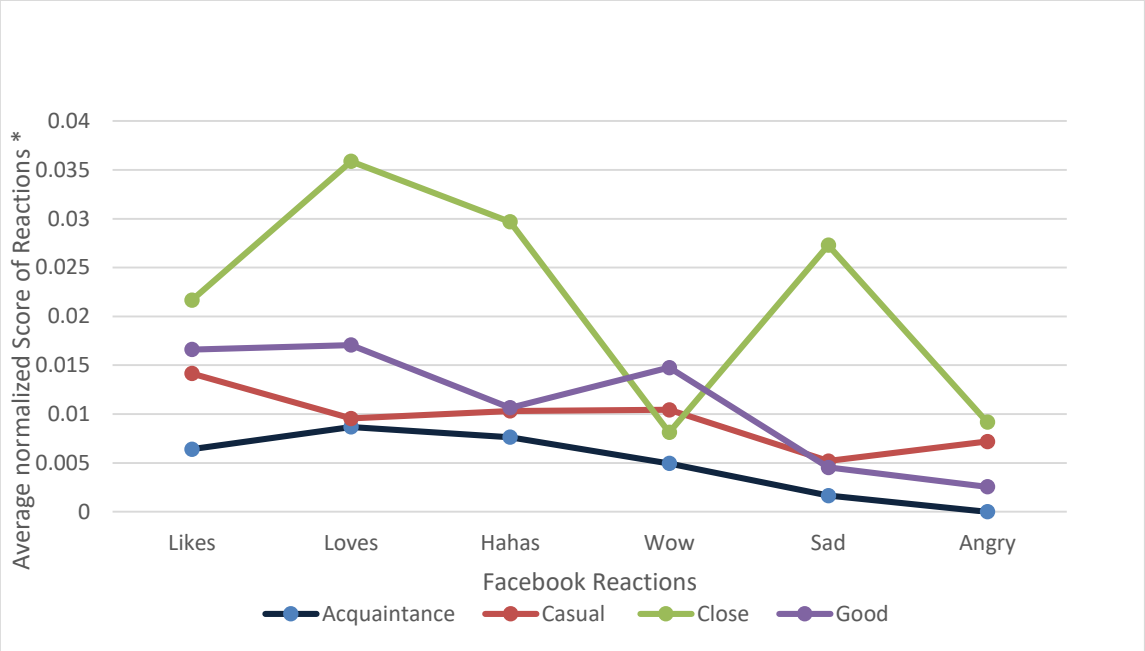
We represented the reaction value of each friend as:

$$\text{Reaction score for } i^{th} \text{ reaction for user-friend pair } (R_f) = \frac{|R_i|}{|R_{im}|} \quad (5)$$

Where,

R_i is the total number of reaction i given by particular friend to the user

R_{im} is the total number of reaction i given by all friends to the user



*Normalized score is calculated as in equation (5)

Figure 3. Average normalized score of Facebook reactions among different categories of friends

There is also a tendency of tagging friends or being tagged in posts on Facebook. We identify this as a separate set of feature with different weights in determining friendship strength. This is because, when a user tags his particular friend, it means that the friend is of high importance to the user. It was evident from our datasets that people often tag their close or good friends. On the other hand, when a friend tags the user, it does not mean the user thinks the same about the friendship. Our feature score for friend tags is given as the ratio of a number of tagging done by the friends *f* to the user *u* to the total number of tags done by all friends to the user *u*.

$$\text{Friend Tag Score } (FTS_{u,f}) = \frac{|Tags\ to\ the\ user\ u\ by\ f|}{|\ Tags\ to\ user\ u\ by\ all\ friends\ |} \quad (6)$$

Similarly, a user friend tag score is the ratio of a number of tags the user *u* has done to the particular friend in *u's* post to the total number of all the tags done by the user *u* to all *u's* friends.

$$\text{User Friend Tag Score } (UFTS_{u,f}) = \frac{|Tags\ to\ the\ friend\ f\ by\ user\ u|}{|\ Tags\ to\ all\ the\ friends\ by\ the\ user\ u\ |} \quad (7)$$

In social networks, we observe that different friends have different comment lengths. Figure 4 below shows the different average comment lengths across four different friend categories.

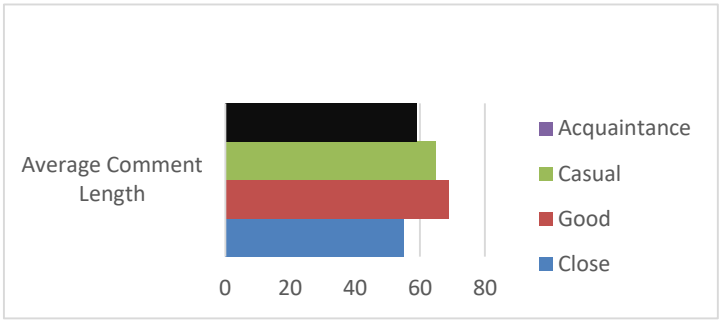


Figure 4. Average comment length among different categories of friends

According to the figure, the average comment length of good friends is slightly higher than other categories so we thought of using it as a feature in friend classification.

Linguistic Features

Another important mode of interaction in Facebook is by giving comments to friend posts. The polarity direction (positive or negative) of the comments given by a friend to the user determines whether there is a positive edge or negative edge between friend and user as per our second hypothesis. From the datasets available, we extracted four such features related to the comments.

Comments Polarity

We assumed that when friends give positive comments, there is positive interaction with the user and if they give negative comments then there is a negative interaction. Giving comments usually has higher weight than giving reactions alone. Therefore, we assumed that positive or negative comments would significantly determine

the strength of friendship. Comment polarity score is determined by the ratio of net comments (Positive – Negative) to the overall comments shared between users and their particular friends. This can be represented as:

$$\text{Comment Polarity Score } (CPS_f) = \frac{|Positive\ Comments\ by\ f| - |Negative\ Comments\ by\ f|}{|\ Comments\ given\ by\ all\ the\ User\ friends\ |} \quad (8)$$

Comment polarity is determined by passing each comment through theVADER library. As explained in sentiment analysis in Chapter 2, we have considered comments as positive if the VADER compound score is >0.1 and negative when the score is <-0.1. We consider the score between 0.1 and -0.1 as a neutral sentiment score. Inspired by this paper [2] , we extend our language-based features by introducing a new set of features called contradiction rank and agreement rank.

Contradiction Rank

Contradiction rank gives disagreement between users. Intuitively, the higher the contradiction, the higher the disagreement between the user pairs. Let x_u^+ be the fraction of the positive comments given by a friend *f* to the user *u*. Let x_u^- be the fraction of the negative comments given by the same friend to the user *u*. Similarly, we define y_u^+ and y_u^- as a fraction of all the positive and negative comments shared to user *u*. Hence, we define Contradiction Rank of the user-friend pair as:

$$CR\ (f, u) = x_u^+ y_u^- + x_u^- y_u^+ \quad (9)$$

For example, if a friend *f* gives 1 positive comment out of 4, then $x_u^+ = 1/4$ and $x_u^- = 3/4$. Suppose most of the friends think positively towards user *u*, making y_u^+ higher, say 5/7. Then, y_u^- is 2/7. Now, Contradiction Rank is CR (f, u) = (1/4) * (2/7) + (3/4) * (5/7) = 0.60. This means that there is high contradiction between the friend and the user.

Agreement Rank

Unlike the contradiction rank, agreement rank gives the degree of agreement between the users. Intuitively, the higher the Agreement Rank, the higher the agreement between users and hence the stronger the bond. For same parameters x_u^+, x_u^-, y_u^+ and y_u^-, we define agreement rank as:

$$AR(f, u) = x_u^+ y_u^+ + x_u^- y_u^- \quad (10)$$

For example, if a friend *f* gives 3 positive comments out of 4, then $x_u^+ = 3/4$ and $x_u^- = 1/4$. Suppose most of the friends think positively towards the user *u*, making y_u^+ higher, say 5/7. Then, y_u^- is 2/7. Now, Agreement Rank is AR (f, u) = (3/4) * (5/7) + (1/4) * (2/7) = 0.60. This means that there is a high degree of agreement between the friend and the user.

Closeness Variable

Upon evaluating the comments in our dataset, we saw that individuals have expressed their opinions like "*Love both of you fluffy goobs! ☺ Great picture honey!*", "*Our heart hurts with urs*","*A huge hug :)* ","*Ohh dear*"," *Wow honey so happy for you.. that's amazing ..!! Skype soon xxx ?!!*" etc. to show their attachments to their friends. We checked the smileys alone in 5919 comments containing smileys, 43% come from close friends, 32% from good friends, 16% from casual friends and 7% from the acquaintances. Most frequently used words in our dataset and their percentage used by different friend categories is listed in Table 2.

Table 2. Percentage of occurrence of the words among different friend categories

Words (Unigram/Bigram/Trigram)	**Close**	**Good**	**Casual**	**Acquaintance**
Love you	51.51%	38.38%	8.08%	4.04%
☺	43.36%	32.74%	16.81%	7.07%
Miss You	61.11%	24.07%	11.11%	3.70%
Birthday	27.14%	31.42%	25.71%	15.71%
Honey	72.72%	18.18%	9.09%	0%
lmfao or lmao	10%	80%	0%	10%
rofl	11.11%	88.88%	0%	0%
xoxo	44.44%	44.44%	0%	11.11%
I Love you	60.46%	23.25%	9.30%	6.97%

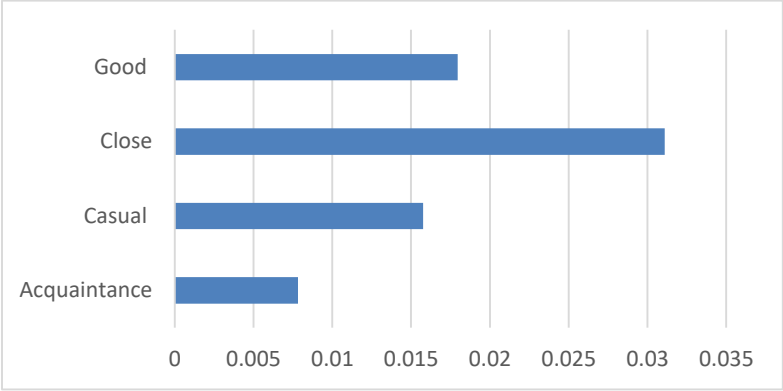
From Table 2 we see that the majority of close and good friends have used “love” in their comments. Although we see that acquaintances and casual friends have also used “*I love*” in their comments, they were distinctly referring to a third person/object instead of the user under consideration. For example, they have commented like “*I love "sparring" with you over political issues.*”, “*I love Hiking*” etc. We also see that (Table 2 blue boxes), close words like *love, miss, honey, lmfao, rofl* alone constitutes more than 80% of close and good friends combined. Besides the words listed in Table 2, we found exclusive use of words such as *babe* and *dude* in close and good friends but significantly lower in number in our other comments dataset.

From this dictionary of bigrams and trigrams, we produce another feature called *closeness feature* which would be a binary value representing the presence or absence of those closeness signifying words.

Structural Features

The strength of friendship between users also depends upon the network they are included. Intuitively, if the users share a large number of common friends between them then it is likely that they have stronger bonds and similarly low or zero common friends would mean that there is a weaker bond between the users. We define Structural Score (SS) as Jaccard Similarity of their number of friends in common which is:

$$\text{Structural Score (SS)} = \frac{|Friends\ of\ f| \cap |Friends\ of\ u|}{|Friends\ of\ f| \cup |Friends\ of\ u|} \quad (11)$$



* Structural score calculated as in equation (11)

Figure 5. Average structural score among different categories of friends

Figure 5 shows the average structural similarity scores across different categories of friends. We see that close friends tend to have more friends in common than good friends. Causal friends also have fewer friends in common than the acquaintance's, which has the least score. This is a general theory and our dataset confirms it too.

Homophilic Features

Homophily is the tendency of a user in the social network to associate or form a bond with similar users[4].We have defined six such homophilic attributes which would increase or decrease the strength of bonds between users in social media. We have used cosine similarity[5] as the primary measure to calculate the similarity score between these different homophilic attributes separately.

Cosine similarity between two vectors is a measure, which calculates the cosine of the angle between them. This measure takes orientation rather than the magnitude of word count (tf-idf) while comparing the documents in normalized space. Cosine similarity between two document vectors 'a' and 'b' is given as

$$\text{Similarity (a, b)} = \cos(\theta) = \frac{a.b}{||a|| \, . ||b||} \qquad (12)$$

Since the value of $\cos(\theta)$ ranges from -1 to 1, we assume that higher the cosine similarity score the stronger the bond between users and lower the score the weaker the bond between users. Following are the homophilic features used in our thesis work:

i. Political Similarity
ii. Education Similarity
iii. Religious Similarity
iv. "Interested In" Similarity
v. Hometown Similarity
vi. Workplace Similarity
vii. Profession Similarity
viii. Group Similarity
ix. User Interests Similarity

---

[4] https://en.wikipedia.org/wiki/Homophily

[5] http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Political similarity refers to a similarity in political ideologies between user friend pairs. A number of 200 user-friend pairs among 680 user-friend pairs in our dataset have their political ideologies set in their profile. People have defined their political ideologies as thinking, conservative, liberal, democratic etc. It is instinctive that if people have common political ideologies then it is more likely that they have a stronger bond. Hence, we have used this as a feature.

Another homophilic attribute is educational similarity. To compare the education similarity between users A and B, we took *institution name, title, location* and *page id* as a whole as a vector of words from the given education profiles and computed cosine similarity with the similar kinds of vectors of other users. We could have just taken the Facebook *page id* of the institution name to check education similarity. However, there could be many Facebook pages with same education institution name. Hence, the BOWM was a better choice here.

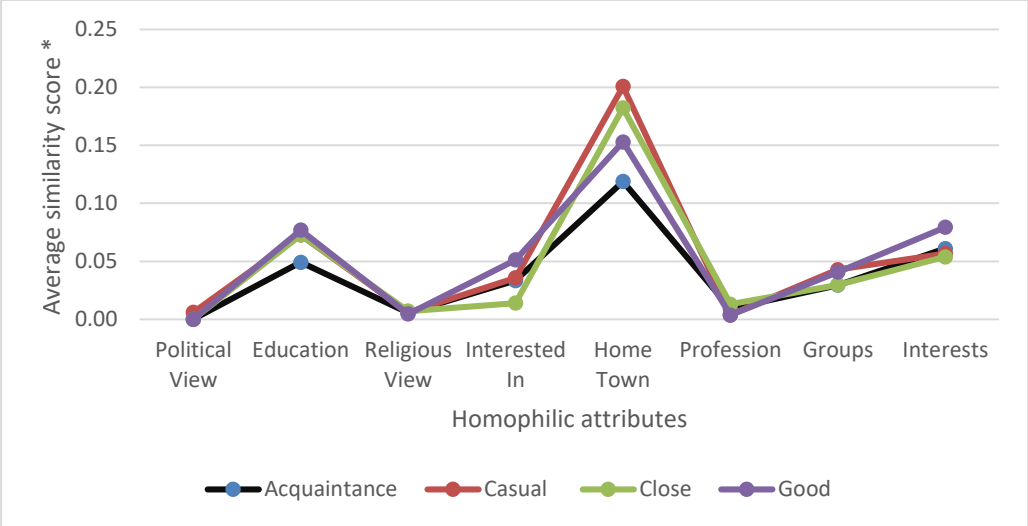
Besides educational similarity, we took religious similarity between user friend pairs as a homophilic attribute. A number of 150 user-friend pairs among 680 user-friend pairs have religious beliefs set in their profiles. People have defined their religious beliefs as Hindu, Christian, a follower of Jesus Christ etc. It is instinctive that people of similar faith could have a strong bond in their friendship. We took the religious beliefs in text form from the users and compared their similarity.

Another homophilic attribute is “Interested In” similarity. “Interested in” in Facebook refers to the sexual preference. Although the friend category may or may not be directly related to sexual preference, we added this as a feature to see if it has some correlation with the friend category.

Similar to education, people from the same hometown tend to be good friends. We used *current city, current hometown, current city page id* and *current hometown page id* of two friends as a vector of words and computed the cosine similarity between them to measure hometown similarity.

Users who work in the same place and have the same profession could have a stronger bond between them. Therefore, we considered this to find the similarity between the workplace and profession between users. To compute workplace similarity, we used *organization name, page id* and *location* of the organization and then used cosine similarity to calculate similarity. Similarly, to compute profession similarity, we used *profession title* between users and computed cosine similarity between them.

Users on Facebook are associated with different public and private groups. We assumed that two close friends are associated with similar groups. Hence, we computed group similarity between the user groups to find closeness. Finally, a close friend of a user might share similar interests with the user. For instance, they might like the same sport, movie etc. Hence, we took interest similarity as another homophilic attribute to compute similarity.



*Similarity score calculated as in equation (12)

Figure 6. Average similarity score between homophilic attributes among different categories of friends

Figure 6 shows the average of similarity score between eight different homophilic attributes among four different categories of friends. We see that there is a high degree of hometown similarity among all other similarities. Our dataset consisted of people from the Boise area so it was natural that they had a higher hometown similarity. Out of four friend categories, casual friend has the highest hometown similarity. We can say that people tend to add friends in Facebook when they have similar hometown regardless of knowing them offline. This also applies in the education and group similarities. Among all groups, acquaintances had the least similarity scores in most of the homophilic attributes. Moreover, we noted that there were fewer similarities in political views, religious views, professions and groups among the user-friend pairs in our dataset and it did not provide significant information.

CHAPTER FOUR: EXPERIMENTS

Datasets

Although Facebook provides good graph API[6] for a variety of processes, the interactions and user profile data that we needed for this project would not be possible with the use of graph API. We needed the wall feeds of the users, user friends, friends of friends, contact information, education information, hometown details, work information, religious and political beliefs and other personal details. Under unavailability of data, we decided to crawl the data ourselves. Crawling was one of the toughest parts of data extraction process since Facebook provides a very good security mechanism to prevent bots. In the due course of this research, we had to create 14 Facebook profiles. The toughest part of this process was to prevent Facebook from recognizing our automation by adding details like profile details, new friends, interaction to each new profile we created after each subsequent block. We were blocked every 3 days of our operation. With this daunting approach and by the use of a computer with local configuration, in time duration of about a month, we were able to extract information of 680 user-friend pairs. However some of the users - friend pairs were removed after the preprocessing steps.

[6] https://developers.facebook.com/

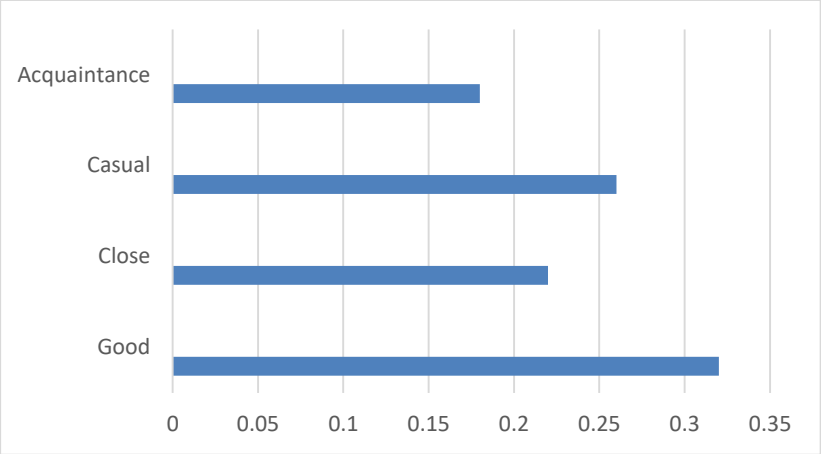


Figure 7. Average percentage of occurrence of different categories of friends

Figure 7 shows the average percentage of occurrence of a different category of friends among the 34 users in the dataset. From this figure, we can see that most of the users have a greater number of good friends compared to causal and acquaintance. The percentage of an average number of good friends is highest and is about 32%.While the percentage of average number of acquaintance is the lowest and is about 18%. The percentage of close friends and casual friends is 21% and 26% respectively. Following were the techniques implemented for data extraction process:

Choosing Users

In the first step, we chose English-speaking users studying at Boise State University through the survey described in the User Survey section. These users were asked to add us as a friend on Facebook. Then we used the Facebook API to obtain the Facebook ids of these participants.

Facebook id is a hex number associated with everyone's Facebook profile. One thing to be noted was that, since the crawling would result in data from the public profile of the user, it would miss the information that was kept private by the user. Therefore, despite our efforts, we could not get all the information needed in this research.

Fetching User Friends

After obtaining the user Facebook ids for the users in the survey, we visited the researcher's list of friends' URL given by *http://www.facebook.com/{user-id}/friends.* We did a depth wise search for a single level to obtain user friends and friend of friends Facebook ids.

Fetching Individual Posts

Similarly, to gather the profile posts, all the user ids of users were taken and each user's wall was visited. The URL of the wall is given by *http://www.facebook.com/{user-id}.* In order to maintain equality between the data and fast retrieval, the extraction process lasted for only 20 minutes for a single user profile.

Fetching User Profiles and Profile Detail Information

After gathering the ids of the user and their friends, we then extracted the user names of the Facebook ids by using graph API. Since API does not provide user profile information other than name, we visited different URLs to get this information as shown in Table 3.

Fetching User Interests and Groups

Since we had all the Facebook ids of the users and their friends under evaluation, we then fetched the interests and groups associated with the users by scraping the URL *http://www.facebook.com/{user-id}/about.*

Table 3. Facebook user profiles URL

URL	Information
http://www.facebook.com/{userid}/about?section=education	User education information
http://www.facebook.com/{userid}/about?section=living	User home town information
http://www.facebook.com/{userid}/about?section=contactinfo	User contact information
http://www.facebook.com/{userid}/about?section=relationship	User relationships information

User Survey

To answer our questions on friendship strength we went through the Institutional Review Board (IRB) at Boise State University. Upon getting the approval of IRB (Appendix C), we sent a mass email to about 2000 students studying at Boise State University, out of which 54 of them replied and agreed to be part of the survey. We created a Facebook page[7] and asked the participants to add us as a friend. In the period of about a month and a half, we extracted the datasets as described earlier. The principle reason for selecting these candidates was first they were studying in the same university as the researcher and second the primary speaking and written language of these candidates was English. This was done because our sentiment model was built with

[7] https://www.facebook.com/profile.php?id=100011647085665

English datasets. Our main goal here was to build a ground truth, which comprised of user-classified friends. To start with the survey, we created our own website called ”*http://nitishdhakal.com.sage.arvixe.com/”.* This website was built upon ASP.Net using C# as a backend with MS SQL Server as a data store. APIs were made to gather survey results with just a click from the running application. The website presented each participant with their random 20 friends and they had to classify their friends under close, good, acquaintance and casual category. These 20 friends were selected based upon their frequency of comments in descending order. The reasoning for this was some friends are close and they interact with the users well, while some do not. In addition, we wanted to evaluate the sentiment of the comments. Figure 8 shows the snippet of the user interface from the survey. Out of the 54 participants, only 34 of them responded within the timeframe of 15 days. From these participants, we had 680 pairs of user– friends.

One of the interesting things found from the survey was almost every user failed to recognize at least one of their Facebook friends presented to them. This tells us that we have a tendency to add unknown people on Facebook.

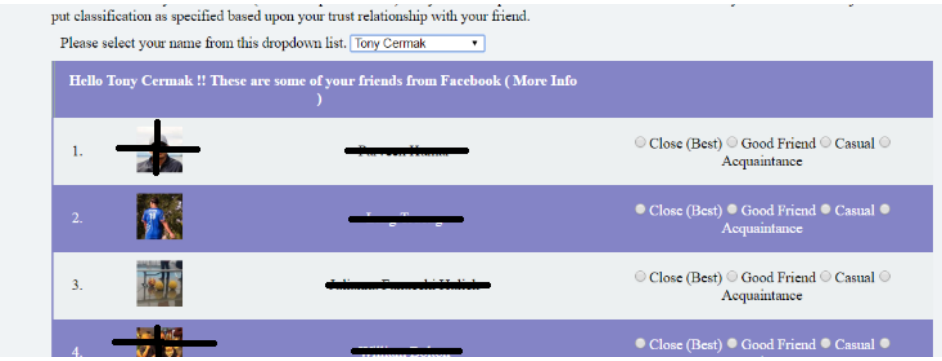


Figure 8. Example of the interface implemented for acquiring the friend classification ground

Limitation in Datasets

The dataset thus obtained contained the interactions observed in the "Facebook wall" of 54 users. A Wall[8] is a profile space of a particular user on Facebook, where all the users and their friends' posts appear as a feed ordered by recency. We crawled the information of only the "wall posts" that were public. Also due to the individual privacy settings, the user profile information like gender, contact information, profession, and user education information was limited to the ones that are public.

Feature Selection

A very good understanding of features leads to better performing models. Dataset containing many irrelevant features has a direct impact on accuracy. Our dataset of 24 features consisted of some features, which were very sparse. Therefore, we needed to perform feature selection for the following reasons:

i. To reduce the features, making generalization much better
ii. To have a better understanding of features and their relationship with the response variable
iii. Reduce misclassification
iv. Reduce computational cost because of redundant features

We used four different feature selection techniques explained as follows:

Pearson Correlation Coefficient

Pearson correlation coefficient[9] is a univariate feature selection method which examines the relationship between each individual features with the response variable. The resulting value of the coefficient lies between [-1, 1], where -1 means perfect negative correlation while +1 means perfect positive correlation. Table 4 shows the

[8] https://en.wikipedia.org/wiki/Facebook_features

[9] https://en.wikipedia.org/wiki/Pearson_product moment_correlation_coefficient

correlation between the vector of 24 features and the output classes. There are four different output classes in the table obtained by a different combination of CL, GO, CA and AC class. CL, GO, CA and AC represents close, good, casual and acquaintances respectively. The combination produces a binary output. The reason for combining these different classes is explained in Data Subset Selection section.

The blue boxes in Table 4 represent higher positive or negative correlations between features and the output class. Among the 24 features, we see that *average comment length, net comment ratio, contradiction rank, agreement rank, likes, loves, haha, sad, friend posts, mutual friends* and *closeness variable* have a high degree of correlation with the strength of output classes. It is noteworthy that since most of the values were sparse, our correlation was very low. N/A values represent features with zero instances.

Table 4. Pearson correlation coefficient between vector of 24 features and output class for different combination of datasets

Features	Correlation Coefficient			
	CL+GO Vs	**CL+GO+CA Vs**	**CL Vs AC**	**CL Vs AC+CA**
Average Comment	0.18	0.16	0.22	0.20
Net Comment Ratio	0.04	-0.14	-0.05	0.13
Contradiction Rank	-0.10	0.12	-0.02	-0.23
Agreement Rank	0.001	-0.20	-0.06	0.16
Likes	0.22	0.39	0.38	0.29
Loves	0.12	0.10	0.19	0.17
Hahas	0.08	0.10	0.09	0.10
Wow	0.047	0.05	0.03	0.02
Sad	0.06	0.13	0.10	0.08
Angry	-0.05	0.08	N/A	-0.07
User Tags	N/A	N/A	N/A	N/A
Friend Posts	0.23	0.22	0.35	0.33
Mutual Friends	0.37	0.43	0.52	0.41
Member of Family	N/A	N/A	N/A	N/A
Political Similarity	-0.04	N/A	N/A	N/A
Education Similarity	0.08	0.10	0.11	0.00
Religious Similarity	0.015	0.01	0.03	0.02
Interested in Similarity	0.005	0.09	-0.04	-0.06
Home Town Similarity	0.03	0.09	0.15	0.09
Workplace Similarity	0.02	-0.002	0.09	0.06
Profession Similarity	0.006	-0.01	0.03	0.05
Groups Similarity	0.001	0.09	0.02	-0.03
Interests Similarity	0.09	0.00	0.02	0.04
Closeness variable	0.20	0.23	0.28	0.24

* CL, GO, CA and AC represents close, good, casual and acquaintances respectively

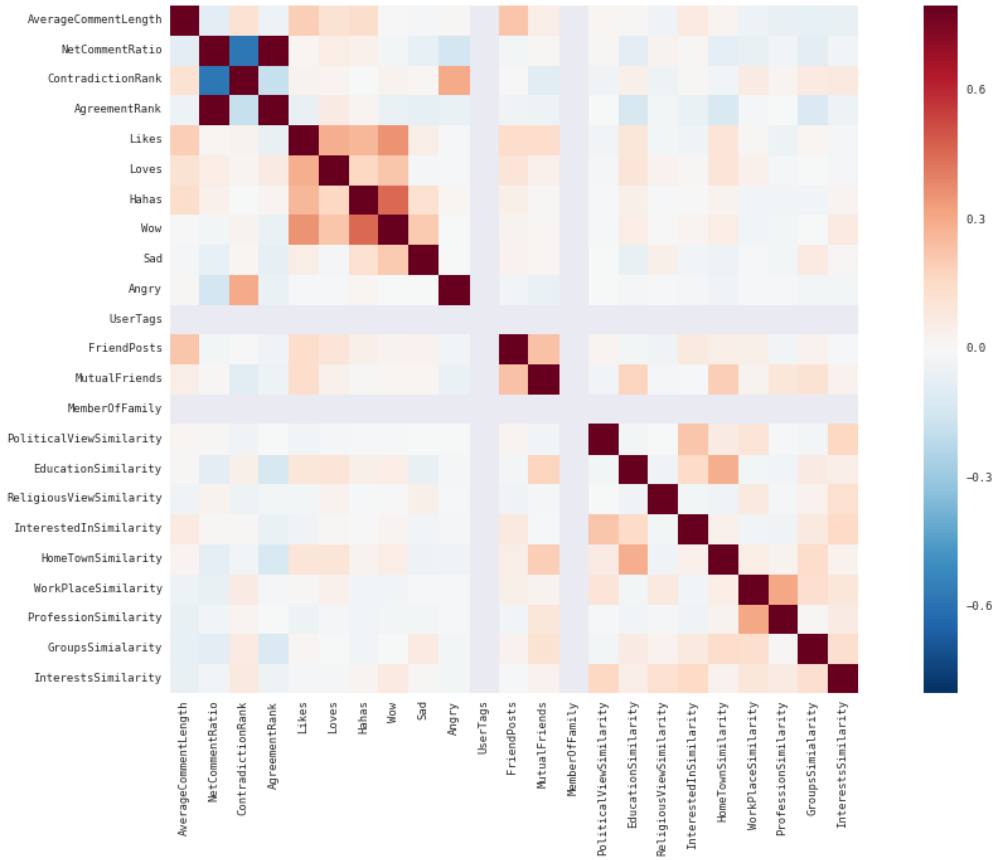


Figure 9. Pairwise correlation coefficient between all the feature variables including class

Figure 9 shows the pairwise correlation among the different features used in the dataset. We observed that *contradiction rank* and *net comment ratio* had a negative correlation, which was true according to the formula that we used for their calculations. Higher *net comment ratio* implies there is a high ratio of positiveness in the comments while higher *contraction rank* implies there is large dissimilarity among the user friend comments. The figure also shows a high degree of correlation between *contradiction rank* and *angry* emoticons. Usually, *angry* emoticon signifies contradiction to the post. It may also signify sarcasm between close friends. Similarly, we found a high degree of correlation between reactions like *likes* with other reactions like *loves*, *haha* and *wow*. The correlation between *education similarity* and *hometown similarity* and between

workplace similarity and *profession similarity* is also higher. This means that people with the same hometown tend to have a similar educational background.

Recursive Feature Elimination (RFE)[10]

This method is one of the feature selection approaches provided by scikit-learn Python library. This method recursively removes attributes and builds a model based on those attributes that remain. We have used logistic regression with this model and used model accuracy to identify attributes or combination of attributes, which contributed the most in predicting the output class. Table 5 shows top five features along with their rank given by RFE.

Table 5. Top five features according to RFE

Features	**Rank**
Average Comment length, Loves, Sad, Friend Posts, Mutual Friends	1
Closeness variable	2
Contradiction Rank	3
Interests Similarity	4
Sad	5

Ensemble of Decision Trees for Feature Importance

This method is another feature selection approach provided by scikit-learn Python library. This method utilizes the ensemble of decision trees like random forest or extra

[10]http://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.RFE.html#sklearn.feature_selection.RFE

trees classifier[11] to compute the relative importance of each feature/attribute. We used the extra trees ensemble on our dataset to find important features as shown in Table 6.

Table 6. Top six features according to an extra trees ensemble

Features	**Rank**
Likes	1
Friend Posts	2
Mutual Friends	3
Interests Similarity	4
Average Comment	5
Contradiction Rank	6

Learning Curve

Since the dataset contained sparse values for some features and these sparse values might have increased our classification error, we generated a learning curve[12] to see which features to keep and which to remove. A learning curve shows training and testing errors with varying numbers of training samples. It is a good tool to find whether our estimator suffers from high variance or high bias. Variance is sensitivity to small variations in the training set. High variance leads to overfitting of the data. Bias is the error obtained due to the incorrect model assumption. This error occurs because the model does not fit in the relevant data points[13]. Figure 10, 11, 12 and 13 shows the

[11]http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html#sklearn.ensemble.ExtraTreesClassifier

[12] http://scikit-learn.org/stable/modules/learning_curve.html

[13] https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

learning curves produced using homophilic features, linguistic features, important features selected from the feature selection process and a complete set of features respectively. The important features mentioned here are *average comment length, likes, loves, friend posts, mutual friends* and *closeness variable*, which we have chosen from the result of correlation analysis and RFE.

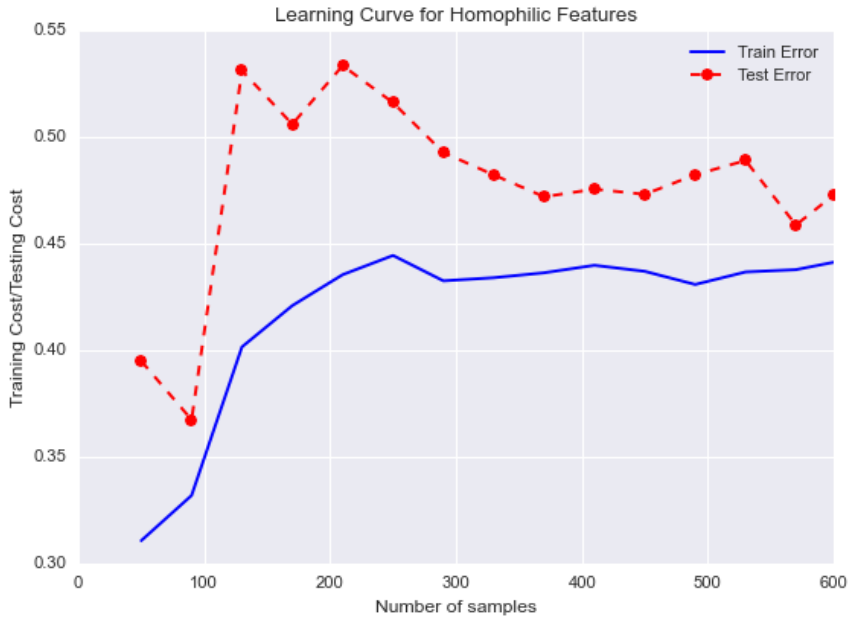


Figure 10. Learning curve obtained using homophilic features

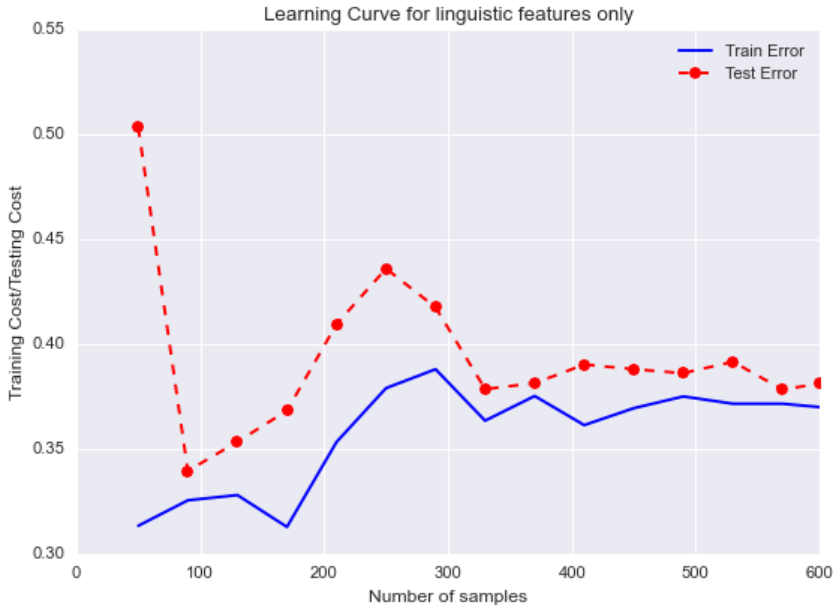


Figure 11. Learning curve obtained using linguistic features

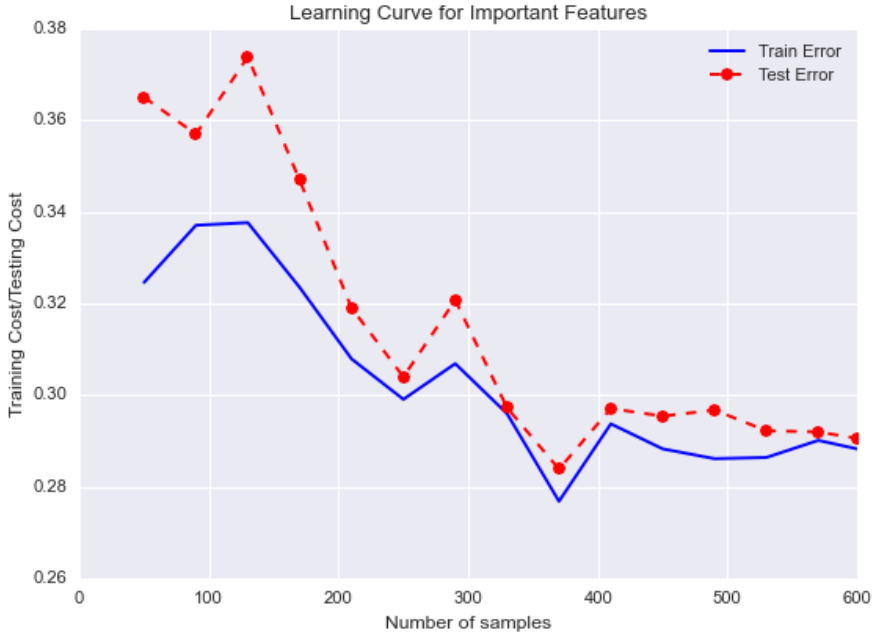


Figure 12. Learning curve obtained using important features

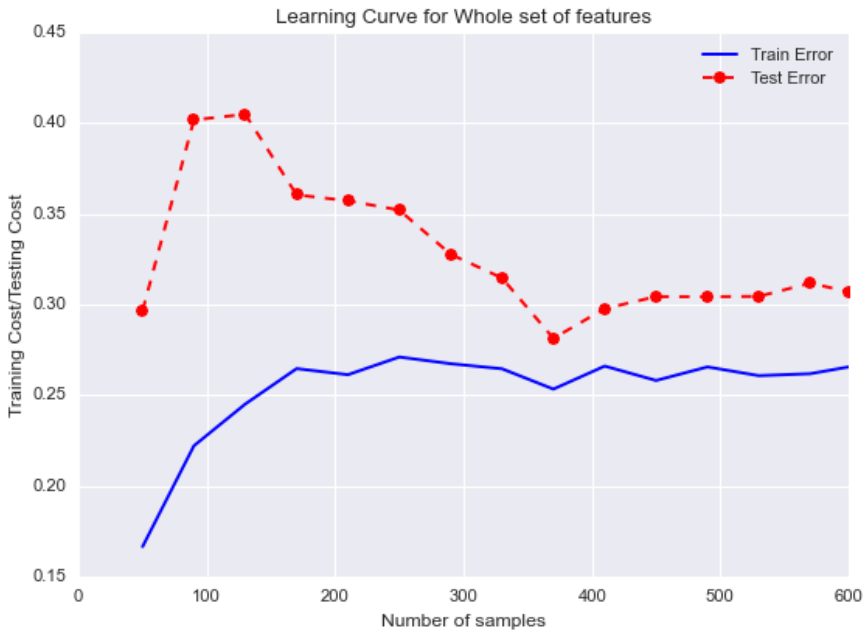


Figure 13. Learning curve obtained using complete set of features

In the learning curve, if both the test error curve and training error curve have high values and are close to one another, then the addition of more training data will not help to improve the model. This is because the model suffers from a high bias problem. If the training error curve and test error curve have a large gap between them for a maximum number of training examples, adding more training examples will cause both curves to converge which ultimately increases the generalization of the model. For a complete 24 set of features (Figure 13), our training and test error was around 25-30% and the train/test error curve appeared to be converging. This suggests that addition of more data will likely cause the curve to converge. For selected important features (Figure 12), train and test errors appeared to be around 29%. In addition, the curve appeared to converge, suggesting that model had learned enough from the features. Learning curves for linguistic features appear to be around 25-40% and the model was overfitting. While the learning curves for homophilic features suggested that model was underfitting and the

addition of more features was not going to help. We used regularized logistic regression as a model here with 2-class classification. The two classes were obtained by combining good and close user friend pairs as one class, and casual and acquaintance friend pairs as another class. This has been discussed in the subsequent data subset selection section.

Data Subset Selection

Figure 14 represents the two-dimensional scatter plot of all 24 features representing the four classes. This plot was rendered by performing Principle Component Analysis[14] on the data set against two principle components. Precise observation of the plot tells us that different stances of close, good and causal intersect each other. In addition, the 24 selected features for the friend classification produces 2 clusters. Since we did not have a very big sample of the dataset, we tried different combinations of the classes of friends. For instance, we combined “good and close” into one and “casual and acquaintance” into other (Figure 15).

In our dataset, many close and good friends had too little interaction. Similarly, there were casual friends who had similar interactions with the close and good friends. In addition, “close and good friends”, “casual, and acquaintances” had similar interactions. We were sure these would bring more anomalies in the classification.

[14] https://en.wikipedia.org/wiki/Principal_component_analysis

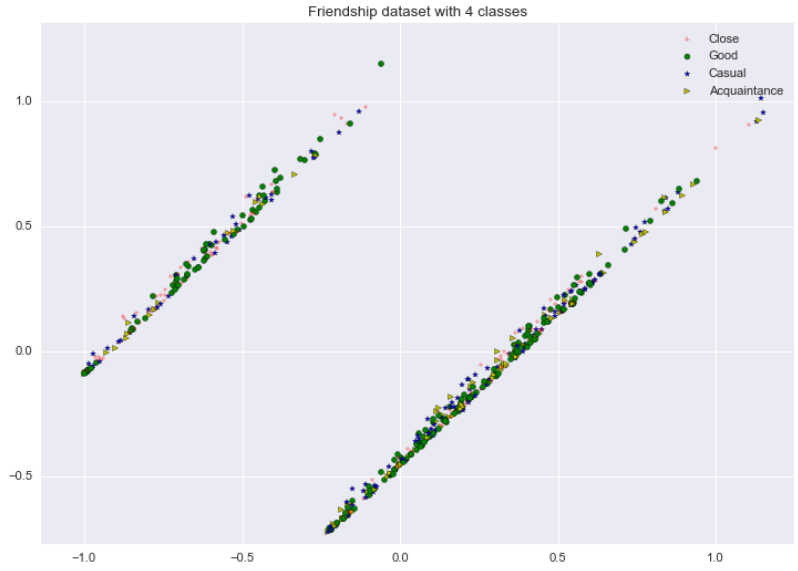


Figure 14. 2D scatter plot of 24 features showing all friend class pairs

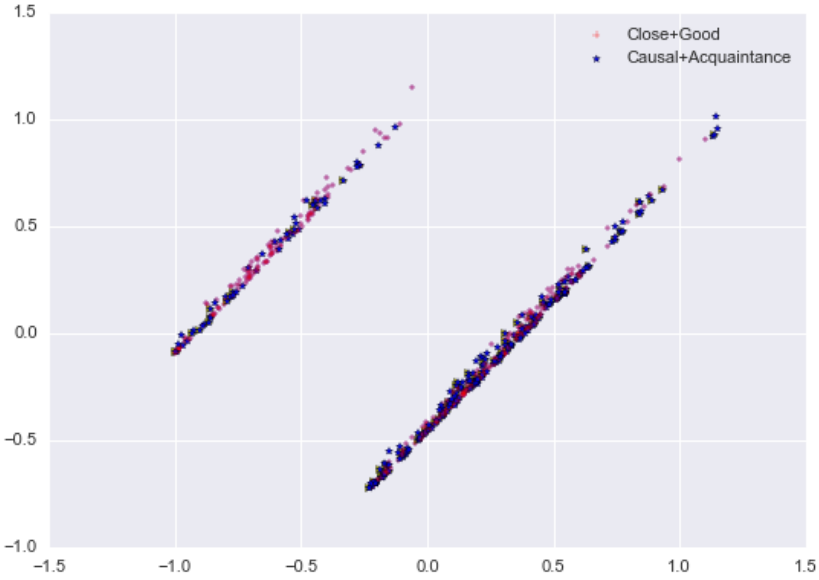


Figure 15. 2D scatter plot of 24 features showing combination of close and good friends pair vs combination of casual and acquaintance friend pair

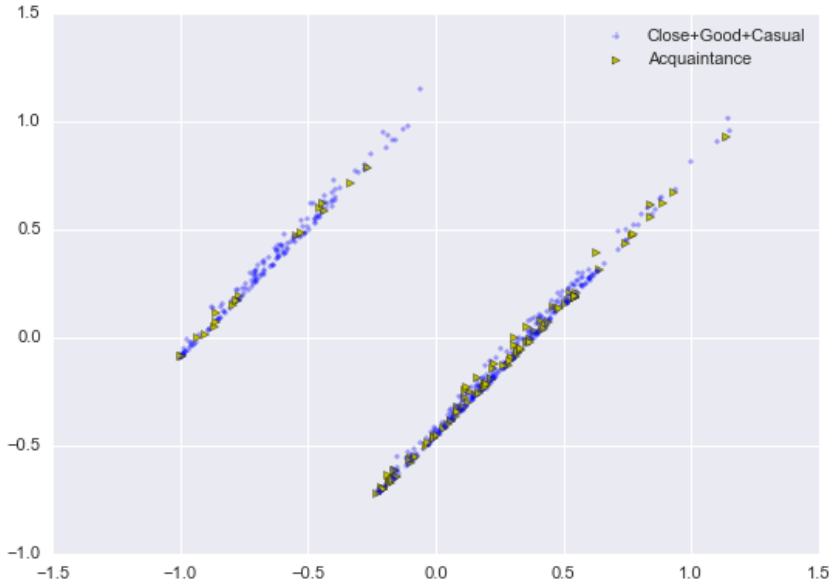


Figure 16. 2D scatter plot of 24 features showing combination of close, good and casual friend pair vs acquaintance friend pair

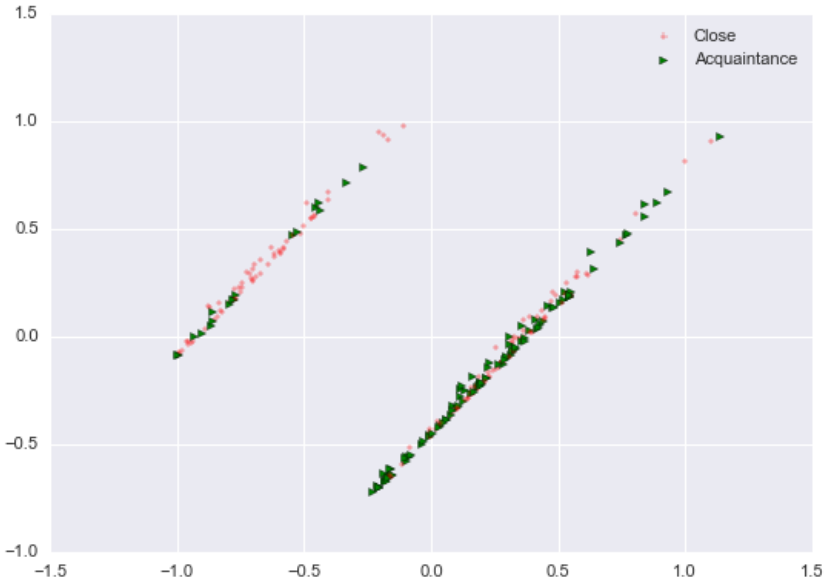


Figure 17. 2D scatter plot of 24 features showing only close and acquaintance friend pair

Therefore, we decided to go for a combination of these datasets to see if the dataset fit our model. Our combination included a different subset of data like “close and good vs acquaintances and casual” (Figure 15),”close good and acquaintances vs causal”

(Figure 16) and “close vs acquaintance” (Figure 17). In all of these combinations, our features divided the dataset into two distinct classes. We did not use PCA for preprocessing as it destroys too much information therefore increasing the misclassification.

The analysis of the false positives between the close/good vs acquaintances/casual showed that there were many close and good friend pairs incorrectly classified as acquaintances. The reason behind this was zero reactions and zero mutual friends. Although it is highly likely that close friends do not interact much with social media, but since we were considering interactions for most of our features, we selected the ones having likes and mutual friends greater than zero. In addition, it is a general observation that close friends privately chat in social media. Since we did not have private chat messages, we removed these zero interactions.

Therefore, based upon the learning curve visualization, and above-mentioned three feature selection techniques, we filtered our different feature sets. Feature subset 1 to feature subset 4 as discussed in the Models to be Evaluated section includes different subsets of the important features we extracted from the Pearson correlation coefficient, Recursive Feature Elimination (RFE) and ensemble of decision trees for feature importance. We then evaluated the nine different models.

Models to be Evaluated

To answer our research questions, we used different subsets of features and studied behavior across all four categories of friends. Following were the models used for evaluation:

All Features Only

In this model, we took all of our 24 sets of features from 680 user-friend pairs.

Bag of Words Model (BOWM) on Comments

Upon evaluating the comments shared among close and good friends in our datasets, we found that words like *hehe, hate, happy, handsome, hurt, hug, honest, forever, babe, xoxo, darling, honey, lmfao, bullshit* etc. were among the top 50 common words. In addition, we saw many words common in a particular category of friends as seen in Closeness Variable section. Therefore, we concluded that the Bag of Words Model could be used in prediction of friendship strength and subsequently used it as a feature. We selected the bag of words from the whole comments shared between a friend and a user. Following were the subtasks carried out to perform BOWM procedure.

i. Tokenization

Since we were using the inverse document frequency matrix for evaluating bag of words, our comments had to be preprocessed. Here, we replaced all the segments of the comment containing html links to “http://addr”. We then performed tokenization. Tokenization involves converting the sentence into lower case and splitting them into individual words. Then we removed the stop words from the comments. Stop words are those words, which have no meaning, but they occur frequently in a sentence. For example *a, is, the* are stop words. We developed our own set of stop words in combination with the NLTK (Natural Language Toolkit) standard dictionary of stop words.

ii. Stemming and lemmatizing

After removing stop words, we stemmed and lemmatized the comments. Stemming is the process of removing different forms of words in a sentence into common base form[15]. For example *organize*, *organizes* and *organizing* can be all converted into the common base word *organize*. Stemming chops off the last part of the word, which sometimes might not make sense. So, in that case, lemmatizing helps. Lemmatizing is a process of converting a word to its base form by the use of vocabulary and morphological analysis of words[24].

iii. Selecting n-grams

In *Scikit* learn *CountVectorizer* object performs the bag of words analysis. Here we can specify the tokenizer, dictionary for stop words, ngram range (unigram, bigram or ngram) and a number of features we want to use. We did an analysis of our Bag of Words Model using unigram, bigram and trigram features.

All Features Combined with BOWM

We combined all the 24 features as explained in the first model containing all features with BOWM on comments here to see if it improved the classification accuracy. Table 7 shows this combination.

[15] http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Table 7. Combined feature set of bag of words along with additional features

		t_1	t_2	...	t_m	feature 1	…	feature 24
d_1		$f_{1,1}$	$f_{1,2}$	...	$f_{1,m}$	feature 1, d_1	…	feature 24, d_1
d_2		$f_{2,1}$	$f_{2,2}$	...	$f_{2,m}$	:	:	:
:		:	:	...	:	:	:	:
:		:	:	...	:	:	:	:
d_n		$f_{n,1}$	$f_{n,2}$	...	$f_{n,m}$	feature 1, d_n	:	feature 24, d_n

Feature Subset 1

We evaluated model on feature subset 1. We obtained this subset after feature selection. The feature subset 1 features included *average comment length, likes, loves, friend posts, mutual friends* and *closeness variable.*

Feature Subset 1 + BOWM

In this model, we combined the BOWM on comments with the feature subset 1 to see if the performance of the combined model increases.

Linguistic Features

Linguistic features include *net comment ratio, agreement rank, contradiction rank* and *closeness variable*. We combined these features to evaluate this model.

Feature Subset 2

This subset included *average comment length, contradiction Rank, likes, friend posts, mutual friends* and *interests similarity.*

Feature Subset 3

This feature subset included *average comment length, contradiction rank, likes, sad, friend posts, mutual friends,* and *interests similarity*.

Feature Subset 4

This feature subset included *average comment length, contradiction rank, likes, loves, friend posts, mutual friends, interest similarity* and *closeness variable*.

With the above-mentioned models, we were able to answer our research questions **Q1** and **Q2.**

Train and Test Sets

We performed 10-fold cross-validation on the 680 user-friend pairs using a SVM classifier, logistic regression classifier, and random forest classifier. To remove the errors due to unbalanced combinations of classes, we used stratified random sampling[16] to select the train-test pair. Stratified random sampling in cross-validation generates stratified folds where each set contains approximately same percentage of each target class as the complete dataset. We ran our experiment on 64-bit Intel(R) Core(TM) 2 Duo CPU with 3.17 GHz CPU and 4GB RAM.

We used Precision, Recall, F-Measure, Accuracy, and AUROC (Area under Receiving Operating Characteristics Curve) as the evaluation metrics.

Results

We started our experiments by using a logistic regression classifier with regularization parameter C as a baseline and performed classification in nine different models. Our primary objective was to find the optimal value of configuration parameters.

[16] http://scikit-learn.org/stable/modules/cross_validation.html

These configuration parameters were regularization parameter (C) in logistic regression classifier and SVM classifier and a number of trees in a random forest classifier. We performed 10-fold cross-validations on different subset of the datasets with different variations of C.

Close and Good Vs Acquaintances and Causal

Our initial experiments in 4-class classification did not produce desirable results. 4-class classification with a random forest containing 100 trees brought a classification accuracy of 0.46 (+/- 0.04). In addition, manual evaluation of the classes suggested that close and good were similar to one another while acquaintances and casual had similarities. In addition, Data Subset Selection studies suggested that our features divided our classes into 2 clusters. Hence, in this section, we performed binary classification of the combination. As explained in the data subset selection section, we did not consider user-friend pairs with zero reactions and zero mutual friends. In addition, since we were using comments as our major source of the feature, we removed the user-friend pairs whose total comment length was less than 50. We, however, kept the ones with the closeness signifying words as discussed in BOWM.

Table 8 shows the results of cross-validation using a logistic regression classifier with a value of regularization parameter (C) 1,100 and 1000 respectively. Here we see that the precision, recall, f-measure and accuracy increases as the value of C increases from 1 to 100 for all 9 different models. Upon increasing the value of C from 100 to 1000, we did not see many changes in the performance of the model. So we chose C=100 as our baseline parameter. For models containing a bag of words, we tried unigram, bigrams and trigrams with 1000, 5000 and 8000 features. After checking the

performance, we settled with unigrams and 8000 features. Among all the models, the model containing feature subset 3 (*average comment length, contradiction rank, likes, sad, friend posts, mutual friends* and *interests similarity*) performed best with a precision of 0.69, recall of 0.83, f-measure of 0.75 and accuracy of 0.71 (see row 8 Table 8) at C=100. Figure 18 shows the area under the ROC curve for this model is 0.69. Feature subset 3 contains most of the features from the feature selection process using RFE. It was interesting to see that model with all training features (see row 1 Table 8) performed weaker than the model with feature subset 3 (see row 8 Table 8). This was mainly because some of the features of 24 features did not correlate with the output variable causing misclassification. In addition, we see that BOWM of comments was not able to provide a clear separation of friend classes. Neither did "linguistic features only", producing an accuracy of 0.61 only. However, feature subset 2,3 and 4 which contained linguistic feature in one form or another contributed to the accuracy from 0.69 (feature subset 4) to 0.70 (feature subset 2).

Table 8. Performance comparison of different regularization parameters for 9 different feature sets in "CL+GO Vs AC+CA" classification using logistic regression

Evaluated Models	Average Precision			Average Recall			Average F-Measure			Average Accuracy		
	C			C			C			C		
	1	100	1000	1	100	1000	1	100	1000	1	100	1000
All features only	0.62	0.68	0.68	0.76	0.81	0.77	0.68	0.74	0.72	0.62	0.69	0.68
BOWM of comments	0.60	0.60	0.59	0.66	0.60	0.61	0.63	0.60	0.60	0.58	0.57	0.57
All features + BOWM	0.60	0.60	0.60	0.66	0.62	0.63	0.63	0.62	0.61	0.58	0.57	0.57
Feature Subset 1	0.61	0.68	0.68	0.79	0.83	0.83	0.69	0.74	0.75	0.62	0.69	0.70
Feature Subset 1 + BOWM	0.60	0.59	0.61	0.67	0.61	0.64	0.64	0.60	0.62	0.59	0.57	0.58
Linguistic features only	0.56	0.60	0.60	0.79	0.70	0.72	0.66	0.64	0.65	0.59	0.61	0.62
Feature Subset 2	0.60	0.68	0.68	0.93	0.83	0.82	0.73	0.75	0.74	0.63	0.70	0.70
Feature Subset 3	**0.60**	**0.69**	**0.69**	**0.93**	**0.83**	**0.82**	**0.73**	**0.75**	**0.74**	**0.63**	**0.71**	**.70**
Feature Subset 4	0.62	0.67	0.69	0.79	0.82	0.82	0.69	0.74	0.74	0.62	0.69	0.70

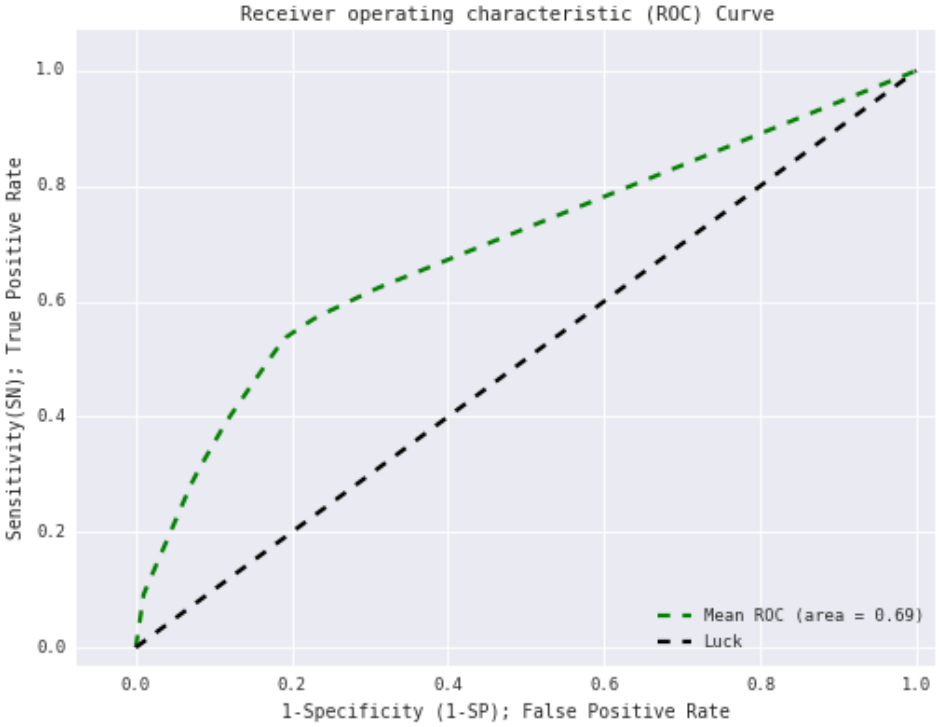


Figure 18. ROC curve of "feature Subset 3" in "CL+GO Vs AC+CA" classification using logistic regression.

We wanted to see if another classifier could bring better accuracy in classifying feature subset 3. Hence, we tried to fit in our features using SVM classifier and random forest classifier. Using the SVM classifier with C=100, 10-fold cross validation produced precision of 0.68, recall of 0.83, f-measure of 0.75 and accuracy of 0.70. Although SVM produced less accuracy than logistic regression, random forest proved to be best one among three. Random forest with 200 trees produced 75% accuracy. Table 9 shows the comparison between three classifiers in fitting our feature subset 3. Figure 19 shows the area under the curve is 0.75 using random forest on feature subset 3.

Table 9. Comparison between classifiers for "feature subset 3" in "CL+GO Vs AC+CA" classification

Evaluated Models	Classifiers	Average Precision	Average Recall	Average F-Measure	Average Accuracy
Feature Subset 3	Log Regression	0.69	0.83	0.75	0.71
	SVM	0.68	0.83	0.75	0.70
	Random Forest	0.78	0.75	0.76	0.75

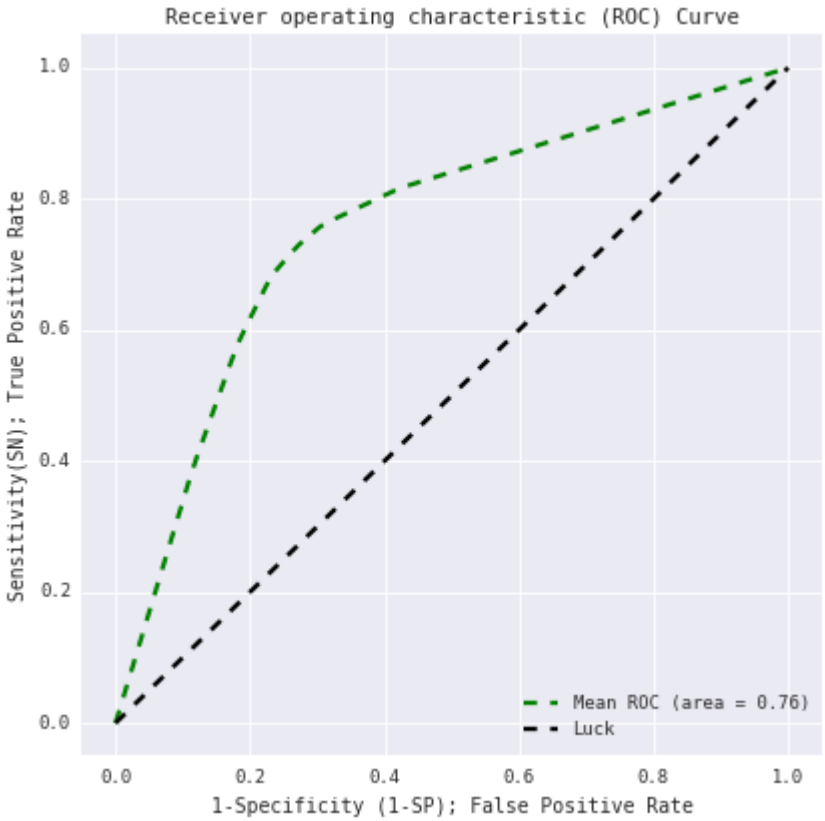


Figure 19. ROC curve of "feature Subset 3" in "CL+GO Vs AC+CA" classification using the random forest.

Close, Good and Causal Vs Acquaintances

In this experiment, we combined close, good and causal user-friend pairs and performed classification against the acquaintances to see how well the model fits our features. Like before, we performed 10 fold cross validations on nine different features and evaluated the performance of our model. We used a logistic regression classifier with C=1 as a baseline and tried to find the optimal value of C. Table 10 shows the result of this cross-validation with C as 1,100 and 1000 respectively. For models using BOWM, we chose unigram with 8000 features as a parameter. Like before, we obtained these parameters by observing the classifier performance on repetitive experiments.

Out of different models evaluated, models containing feature subset 1(*average comment length, likes, loves, friend posts, mutual friends,* and *closeness variable*) and models containing feature subset 4 (*average comment length, contradiction rank, likes, loves, friend posts, mutual friends, interest similarity* and *closeness variable)* performed better. Models containing features subset 1 performed best with the precision of 0.72, recall of 0.83, f-measure of 0.75 and accuracy of 0.74 (see row 4 Table 10) at C=100. Figure 20 shows the area under the ROC curve for model containing feature subset 1 is 0.74. Models containing features subset 4 performed equally well with the precision of 0.71, recall of 0.81, f-measure of 0.75 and accuracy of 0.74 (see row 9 Table 10) at C=1000. Here, linguistic features performed slightly better than "close and good vs causal and acquaintance" classification with an accuracy of 0.65. This might be because acquaintances use less proximity related words (as explained in BOWM) than the good, casual and close friends do. Our linguistic feature contained closeness variable that considers these words. Feature subset 2 and 3, which contained linguistic feature in one

form or the other, contributed to the accuracy up to 0.73. In addition, we see that BOWM of comments features did not work. This might be due to the smaller number of comments we had in the dataset.

Table 10. Performance comparison of different regularization parameters for 9 different feature sets in "CL+GO+CA Vs AC" classification using logistic regression

Evaluated Models	Average Precision			Average Recall			Average F-Measure			Average Accuracy		
	C			C			C			C		
	1	100	1000	1	100	1000	1	100	1000	1	100	1000
All features only	0.62	0.72	0.71	0.66	0.78	0.76	0.63	0.74	0.73	0.63	0.73	0.73
BOWM of comments	0.61	0.59	0.60	0.64	0.61	0.65	0.62	0.59	0.62	0.61	0.58	0.59
All features + BOWM	0.61	0.60	0.59	0.64	0.63	0.65	0.62	0.61	0.61	0.61	0.59	0.58
Feature Subset 1	**0.58**	**0.72**	**0.70**	**0.83**	**0.83**	**0.82**	**0.68**	**0.75**	**0.75**	**0.61**	**0.74**	**0.73**
Feature Subset 1 + BOWM	0.60	0.62	0.61	0.63	0.63	0.65	0.61	0.62	0.61	0.60	0.61	0.59
Linguistic features only	0.60	0.70	0.66	0.68	0.61	0.60	0.63	0.63	0.62	0.61	0.65	0.64
Feature Subset 2	0.66	0.69	0.69	0.74	0.85	0.85	0.68	0.75	0.76	0.66	0.72	0.73
Feature Subset 3	0.67	0.70	0.69	0.75	0.85	0.85	0.69	0.76	0.76	0.67	0.73	0.73
Feature Subset 4	**0.59**	**0.70**	**0.71**	**0.82**	**0.8**	**0.81**	**0.68**	**0.73**	**0.75**	**0.61**	**0.71**	**0.74**

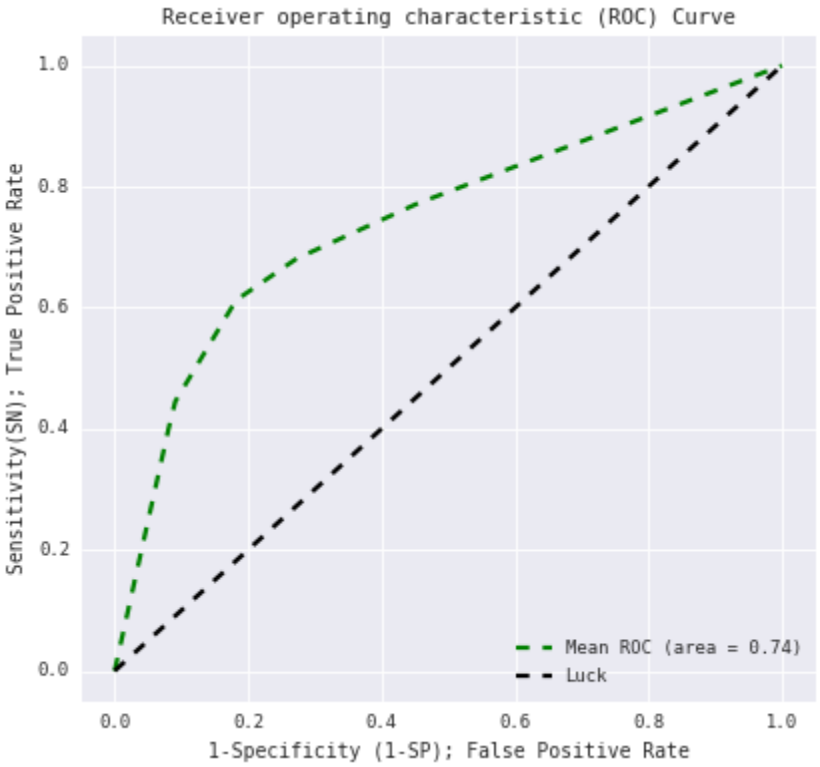


Figure 20. ROC curve of "feature Subset 1" in "CL+GO+CA Vs AC" classification using logistic regression

As stated earlier, we wanted to see if another classifier could bring better accuracy in classifying our feature subset 1. Using the SVM classifier with C=100, 10-fold cross validation produced average precision of 0.69, recall of 0.88, f-measure of 0.77 and accuracy of 0.73. SVM surpassed logistic regression with an accuracy of 0.74. However, random forest proved to be best one among three. Random forest with 100 trees produced an accuracy of 81%. Table 11 shows the comparison between three classifiers in fitting our "feature subset 1". Figure 21 shows the area under the curve is 0.82 using random forest on "feature subset 1".

Table 11. Comparison between classifiers for "feature subset 1" in "CL+GO+CA Vs AC" classification

Evaluated Models	Classifiers	Average Precision	Average Recall	Average F-Measure	Average Accuracy
Feature Subset 1	Logistic Regression	0.72	0.83	0.75	0.74
	SVM	0.69	0.88	0.77	0.73
	Random Forest	0.85	0.76	0.80	0.81

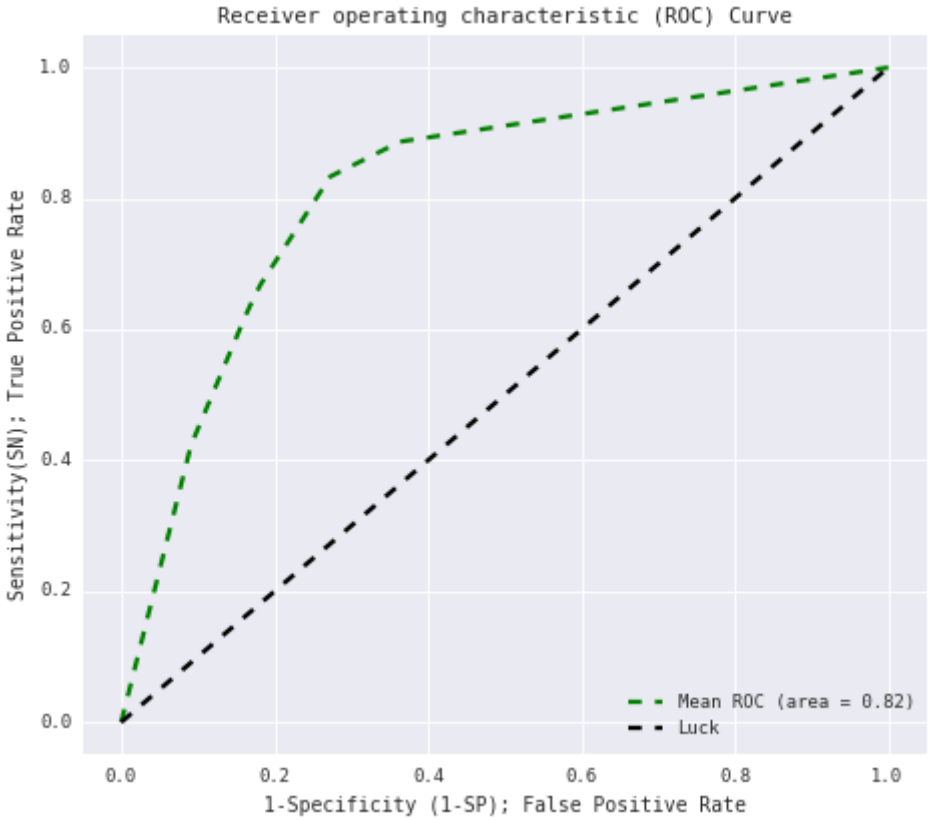


Figure 21. ROC curve of "feature Subset 1" in "CL+GO+CA Vs AC" classification using the random forest.

Close Vs Acquaintance

There is a significant difference in interactions, structural and homophilic features between casual and acquaintance user-friend pairs in the social network. In this experiment, we tried to encapsulate this idea and observed how classifiers work across these distant class of friends. Table 12 shows the performance comparison of different regularization parameters for 9 different feature sets using logistic regression. Like before we started by using C=1 as our baseline and increased to 100 and 1000 to find the optimum value of C. We chose 1000 features and unigrams as our parameters in models containing bag of words.

Table 12 shows that the model containing feature subset 1(*average comment length, likes, loves, friend posts, mutual friends* and *closeness variable*) outperforms all other models here. With C=100, the classifiers achieved the accuracy up to 82% while classifying close and acquaintance user-friend pair using this model. Figure 22 shows the high area under the ROC curve for this model of about 0.81. Surprisingly, models containing BOWM did not work here as well. A model containing only linguistic features did not work either. However linguistic features like closeness variable contributed to our best models containing feature subset 1.

Table 12. Performance comparison of different regularization parameters for 9 different feature sets in "CL Vs AC" classification using logistic regression

Evaluated Models	Average Precision			Average Recall			Average F-Measure			Average Accuracy		
	C			C			C			C		
	1	100	1000	1	1	100	1000	1	1	100	1000	1
All features only	0.68	0.73	0.77	0.81	0.87	0.9	0.73	0.79	0.82	0.69	0.75	0.79
BOWM of comments	0.66	0.63	0.61	0.76	0.75	0.73	0.70	0.68	0.66	0.65	0.63	0.60
All features + BOWM	0.66	0.64	0.62	0.76	0.76	0.73	0.70	0.69	0.67	0.65	0.64	0.61
Feature Subset 1	**0.64**	**0.79**	**0.79**	**0.83**	**0.93**	**0.92**	**0.72**	**0.85**	**0.84**	**0.66**	**0.82**	**0.81**
Feature Subset 1 + BOWM	0.67	0.63	0.62	0.73	0.76	0.75	0.69	0.69	0.68	0.65	0.63	0.61
Linguistic features only	0.63	0.63	0.63	0.80	0.77	0.77	0.70	0.69	0.69	0.63	0.62	0.63
Feature Subset 2	0.60	0.73	0.78	0.99	0.91	0.89	0.75	0.81	0.82	0.64	0.77	0.79
Feature Subset 3	0.61	0.74	0.78	0.99	0.92	0.89	0.75	0.82	0.82	0.65	0.78	0.79
Feature Subset 4	0.68	0.73	0.77	0.81	0.87	0.9	0.73	0.79	0.82	0.69	0.75	0.79

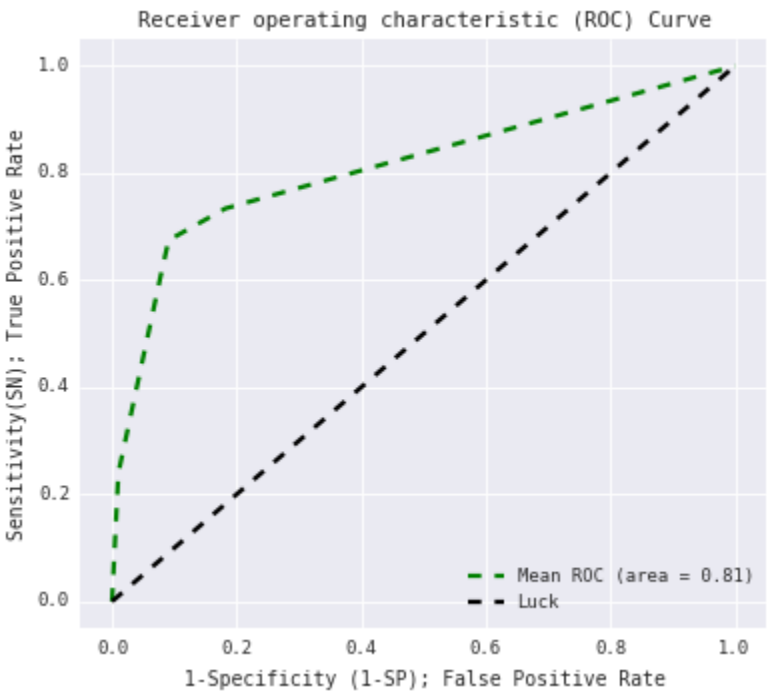


Figure 22. ROC curve of "feature Subset 1" in "CL Vs AC" classification using logistic regression

We compared our accuracy obtained from logistic regression with SVM and random forest classifiers. Random forest again produced a better performance against the other 2 classifiers with the precision of 0.87, recall of 0.86, f-measure of 0.86, accuracy of 0.85 and good area under the curve of 0.85. For SVM, optimum value of regularization parameter was 100 while for random forest we used 100 trees for optimum performance. Table 13 shows the comparison between these three classifiers. Figure 23 shows the area under the curve on feature subset 1using the random forest.

Table 13. Comparison between classifiers for "feature subset 1" in "CL Vs AC" classification

Evaluated Models	Classifiers	Average Precision	Average Recall	Average F-Measure	Average Accuracy
Feature Subset 1	Logistic Regression	0.79	0.93	0.85	0.82
	SVM	0.80	0.92	0.86	0.83
	Random Forest	0.87	0.86	0.86	0.85

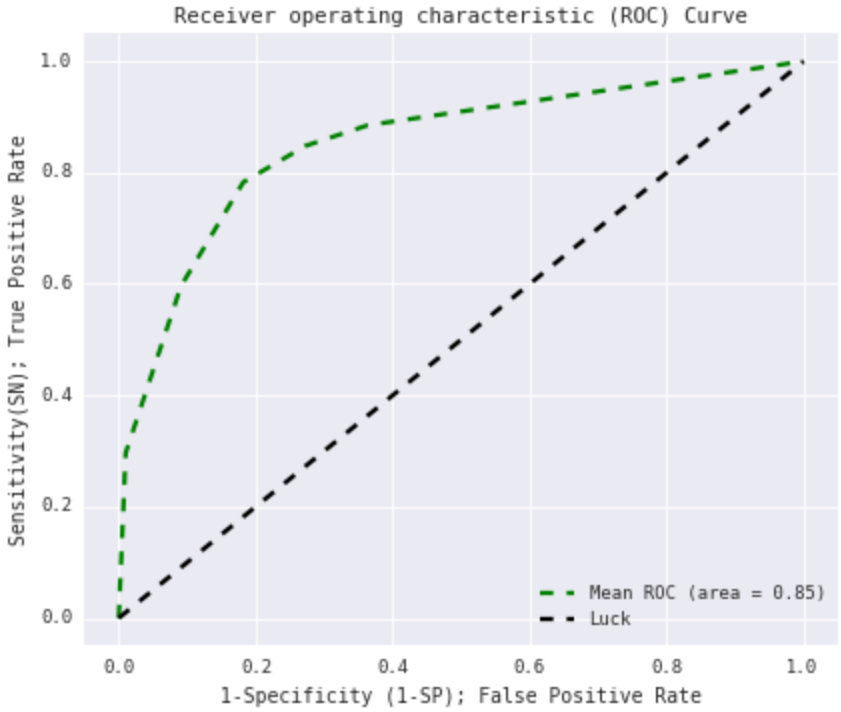


Figure 23. ROC curve of "feature Subset 1" in "CL Vs AC" classification using random forest

Close Vs Acquaintance and Casual

Our third experiment on close and acquaintance friend classification produced good classification accuracy of 85%. In this experiment, we wanted to see if the classifier was able to distinguish close versus a combination of casual and acquaintance. Like the previous experiments, we obtained the subset of the dataset using stratified random sampling and performed the 10 fold cross validation using nine different models. We started with logistic regression classifier with regularization parameter (C) 1.

Models containing feature subset 1(*average comment length, likes, loves, friend posts, mutual friends* and *closeness variable*) worked better here as well. With C=100, the classifiers achieved the accuracy up to 75% and ROC of 0.74. Table 14 shows the performance comparison between different models and Figure 24 shows the area under the ROC curve for this model. The classification performance obtained in this experiment does not match up with the accuracy of 82% obtained between close and acquaintance user-friend pair. There were many false positives showing a casual friend as a close friend. Models containing bag of words did not work here and neither did the model containing only linguistic feature. However linguistic features like closeness variable contributed to our best feature subset 1.

Table 14. Performance comparison of different regularization parameters for nine different feature sets in "CL Vs AC+CA" classification using logistic regression

Evaluated Models	Average Precision			Average Recall			Average F-Measure			Average Accuracy		
	C			C			C			C		
	1	100	1000	1	1	100	1000	1	1	100	1000	1
All features only	0.60	0.68	0.68	0.70	0.75	0.73	0.63	0.70	0.69	0.62	0.69	0.69
BOWM of comments	0.62	0.63	0.55	0.62	0.69	0.59	0.61	0.65	0.56	0.61	0.64	0.56
All features + BOWM	0.62	0.63	0.55	0.62	0.69	0.59	0.61	0.65	0.56	0.61	0.64	0.55
Feature Subset 1	**0.62**	**0.75**	**0.72**	**0.79**	**0.80**	**0.80**	**0.69**	**0.75**	**0.75**	**0.65**	**0.75**	**0.74**
Feature Subset 1 + BOWM	0.61	0.63	0.54	0.58	0.65	0.55	0.58	0.63	0.54	0.60	0.63	0.54
Linguistic features only	0.56	0.64	0.63	0.71	0.66	0.65	0.62	0.64	0.62	0.57	0.65	0.64
Feature Subset 2	0.71	0.74	0.71	0.8	0.80	0.8	0.75	0.75	0.74	0.74	0.74	0.73
Feature Subset 3	0.69	0.71	0.72	0.78	0.81	0.79	0.73	0.75	0.75	0.72	0.74	0.73
Feature Subset 4	0.60	0.71	0.71	0.78	0.79	0.77	0.67	0.73	0.74	0.63	0.73	0.72

Figure 24. ROC curve of "feature Subset 1" in "CL Vs AC+CA" classification using logistic regression

We compared the result of logistic regression with the random forest and SVM. Like before, random forest performed better than logistic regression and SVM. Table 15 shows this comparison. With random forest, we obtained precision of 0.78, which was higher than our baseline logistic regression. Although accuracy with random forest dropped by 1 point but the area under the ROC curve was higher than logistic regression. Figure 25 shows the area under the curve on feature subset 1 using the random forest.

Table 15. Comparison between classifiers for "feature subset 1" in "CL Vs AC+CA"classification

Evaluated Models	Classifiers	Average Precision	Average Recall	Average F-Measure	Average Accuracy
Feature Subset 1	Logistic Regression	0.75	0.80	0.75	0.75
	SVM	0.70	0.80	0.74	0.73
	Random Forest	0.78	0.71	0.71	0.74

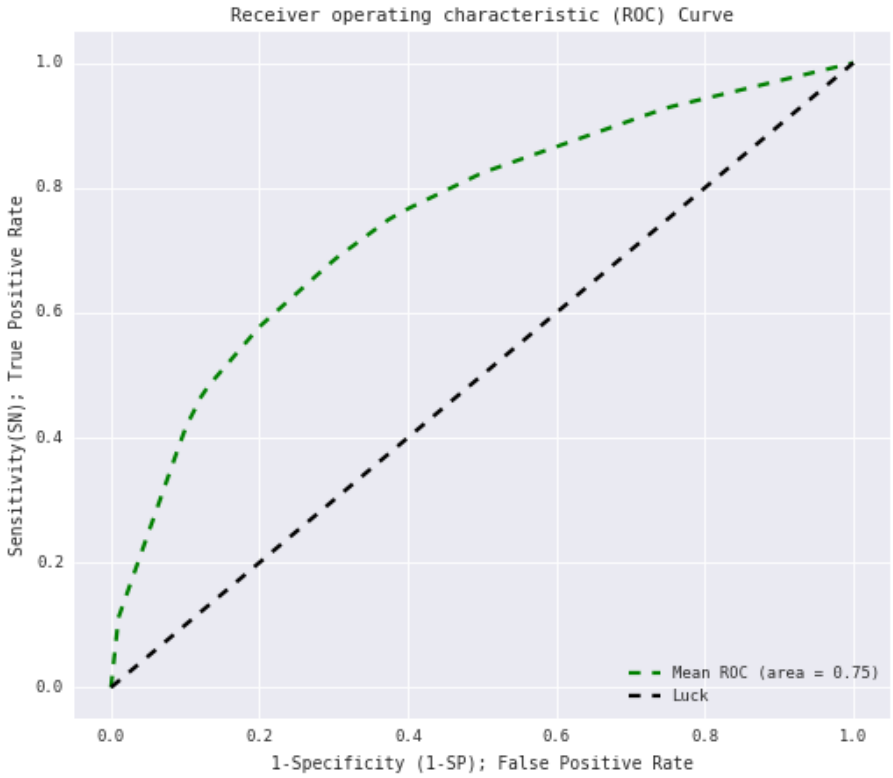


Figure 25. ROC curve of "feature Subset 1" in "CL Vs AC+CA" classification using random forest

Only Interaction Features

In this experiment, we tried to see the importance of interaction features in predicting friendship strength. Here we took the results of the best performing models in the four-dataset combinations. For combinations, “close, good and casual vs acquaintance”, “close vs acquaintance” and “close vs acquaintance and casual”, we took results of random forest on feature subset 1 (see Figure 21, Figure 23 and Figure 25), since it performed better. Similarly, for dataset containing “close and good vs casual and acquaintance”, we took results of random forest on feature subset 3 (see Figure 19) as it performed better. Feature subset 1 (*average comment length, likes, loves, friend posts, mutual friends* and *closeness variable)* consisted of *average comment length, likes, loves* and *friend posts* as interaction features. Feature subset 3 *(average comment length, contradiction rank, likes, sad, friend posts, mutual friends* and *interests similarity)* consisted of *likes, sad, friend posts* as interaction features.

For our best performing models in datasets, “close, good and casual vs acquaintance” and “close vs acquaintance”, we found that interaction features alone has a very good capability in determining friendship strength. The high area under the ROC curve of 0.80 of “close, good and casual vs acquaintance” (Figure 27) proves this. Similarly, for “close vs acquaintance” we get area under the ROC curve of about 0.75.

Interaction feature alone was not the deciding feature in classifying the dataset combinations, “close and good vs acquaintances and casual” (Figure 26) and “close vs acquaintance and casual” (Figure 27). However, obtaining an area under the ROC curve of more than 0.62 signifies that interaction features played its positive part in determining friendship strength across these datasets as well. Since close, good and casual have

similar interactions in the dataset, prediction by separating these classes with interaction features may not have helped much.

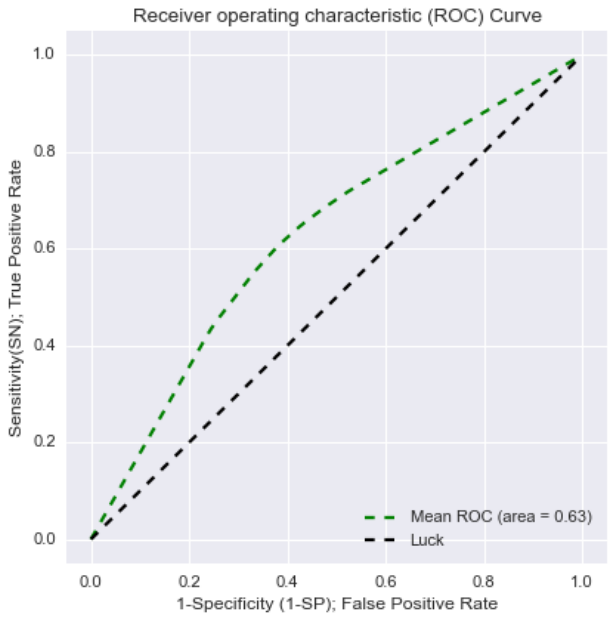


Figure 26. ROC curve in "CL+GO Vs AC+CA" classification with interaction features only

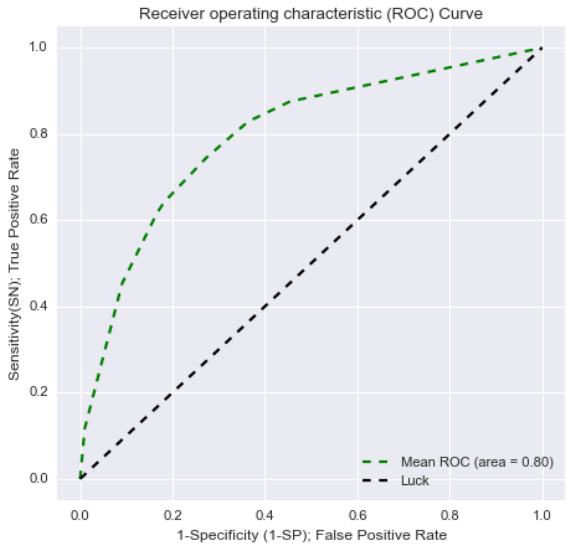


Figure 27. ROC curve in "CL+GO+CA Vs AC" classification with interaction features only

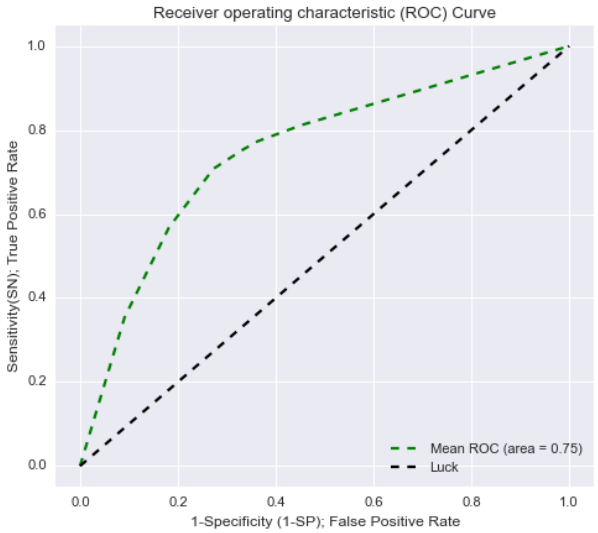


Figure 28. ROC curve in "CL Vs AC" classification with interaction features only

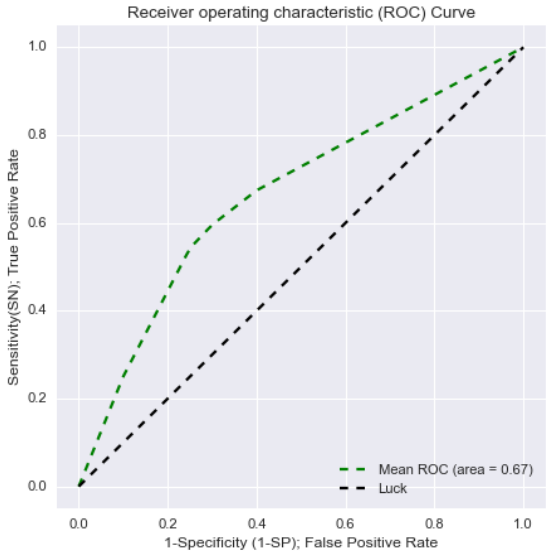


Figure 29. ROC curve in "CL Vs AC +CA" classification with interaction features only

Linguistic Vs Other Features

In this experiment, we compared the classification result with and without linguistic features to identify the strength of linguistic features. Like the experiment on interaction features only, we took best results obtained from all 4 combination of datasets. For dataset combinations, "close, good and acquaintance and casual", "close vs

acquaintance" and "close vs acquaintance and casual", we took results of random forest on feature subset 1 (see Figure 21, Figure 23 and Figure 25), as it performed better. Similarly, for the dataset containing "close and good vs casual and acquaintance", we took results of random forest on feature subset 3 (see Figure 19), as it performed better. Feature subset 1 (*average comment length, likes, loves, friend posts, mutual friends* and *closeness variable)* consisted of *closeness variable* as the language feature. Feature subset 3 *(average comment length, contradiction rank, likes, sad, friend posts, mutual friends* and *interests similarity)* consisted of *contradiction rank* as language feature. We checked the classifier performance with and without this linguistic feature on four of these dataset combinations.

We observed that the performance of the classifiers decreases as we remove the linguistic features from the model. The ROC curve obtained without the addition of linguistic features suggested this decrease in performance level (see Figure 30, Figure 31, and Figure 32). Although there is a decrease in area under the ROC curve in 3 of our datasets, area under the ROC curve of dataset "close vs acquaintance and casual" does not show any change in the performance after removal of linguistic features. Here other features played important role in predictability.

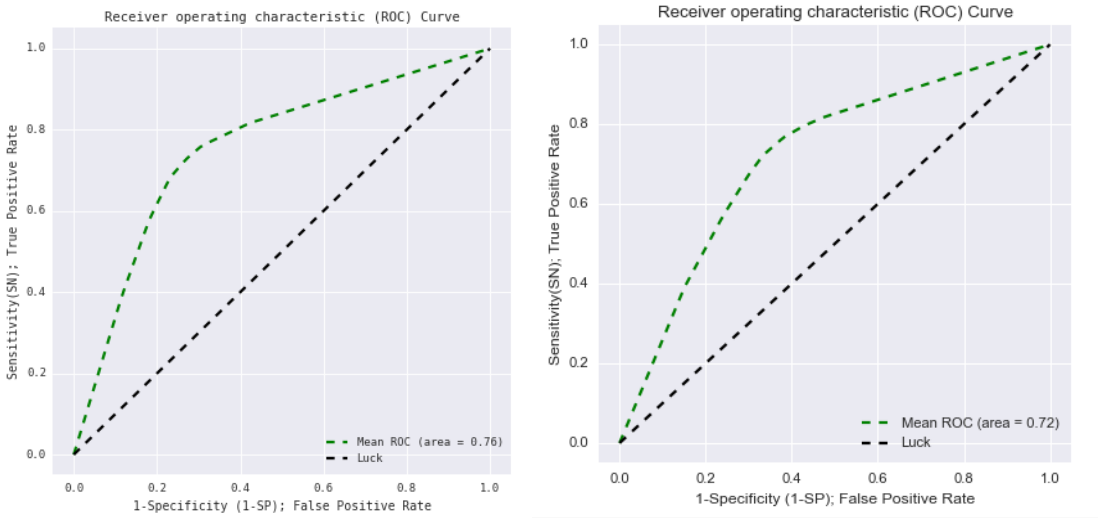


Figure 30. Comparison between ROC curves across "CL+GO Vs AC+CA" classification with and without linguistic features

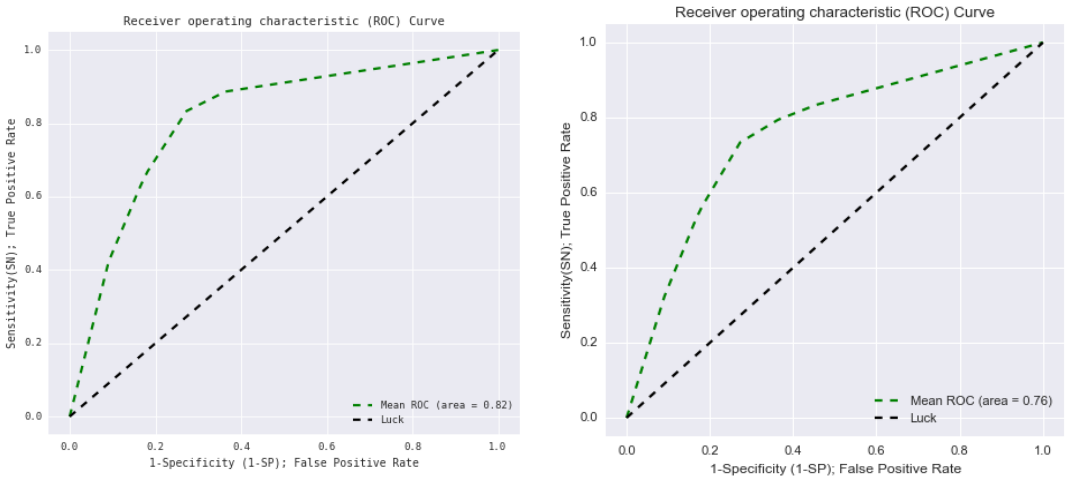


Figure 31. Comparison between ROC curves across "CL+GO+CA Vs AC" classification with and without linguistic features

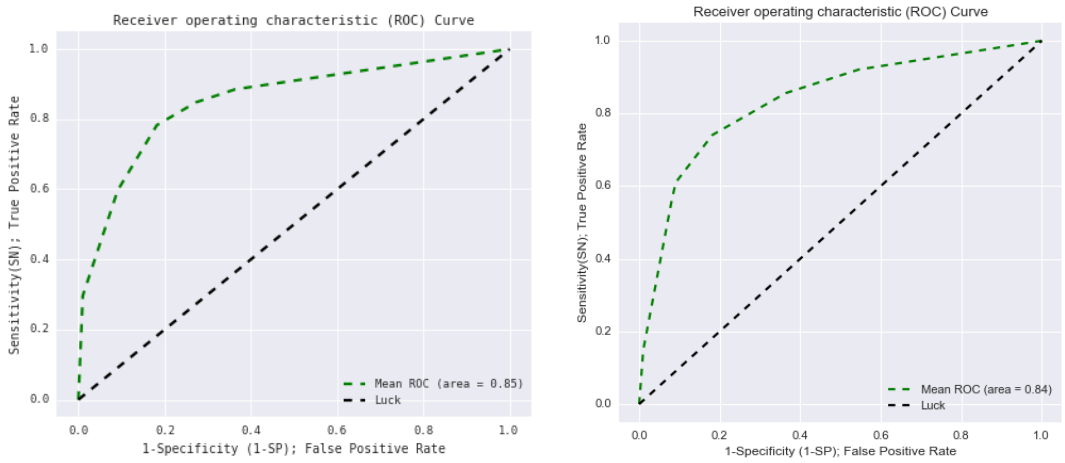


Figure 32. Comparison between ROC curves across "CL Vs AC" classification with and without linguistic features

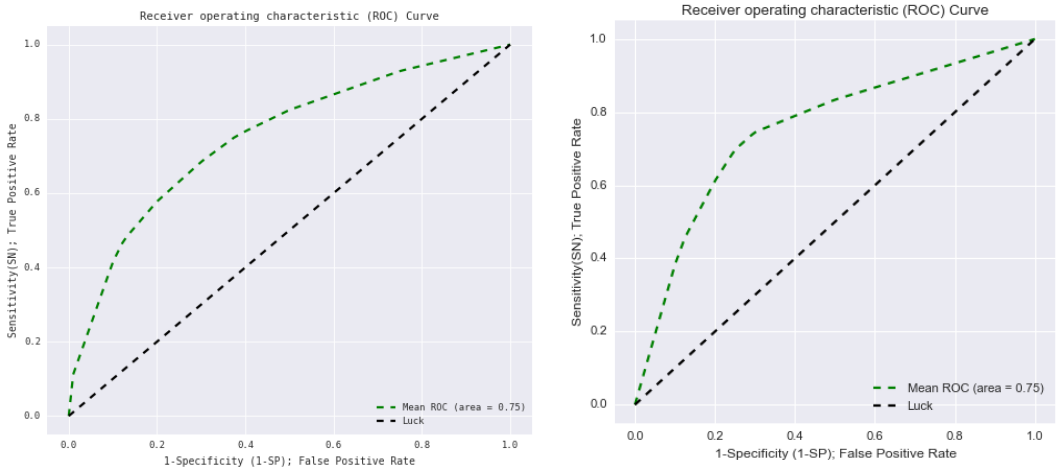


Figure 33. Comparison between ROC curves across "CL Vs AC+CA" classification with and without linguistic features

Analysis

Among the nine different subsets of features used for friend classification, models containing feature subset 1 (*average comment length, likes, love, friend posts, mutual friends* and *closeness variable*) and models containing feature subset 3 (*average comment length, contradiction rank, likes, sad, friend posts, mutual friends* and *interests similarity*) performed well. These models contained structural features like *mutual friends*, linguistic features like *closeness variable, contradiction rank,* interaction features like *likes, love*

and *friend posts* and homophilic features like *interests similarity*. We saw that interaction variables alone has the greater predictability in finding friendship strength across all combinations of datasets (see Only Interaction Features). This answers our research question **Q1**, that interaction variables are good estimators of friendship strength across different friend categories. We also wanted to see if linguistic feature plays an important role in determining the friendship strength to answer our research question **Q2**. Our assumption was not successful in the sense that the model containing only linguistic features did not predict well. However, comparison of the models with and without linguistic features (see Linguistic Vs Other Features) showed that addition of linguistic features along with other features contributed in classifying strong and weak friends across the majority of the datasets. Linguistic features like *contradiction rank* and *closeness variable* contributed to our best performing models (model containing feature subset 1 and model containing feature subset 3). Our assumption of good friends contradicting less than acquaintances worked here. In addition, our assumption that good friends used more closeness related words worked as well.

Although close friends were seen to be somewhat more positively interacting with social media than others were, the general trend of sentiment polarity was seen to be positive for all four classes of friends. Figure 34 shows the average sentiment intensity of comments across different categories of friends. The figure shows similar average sentiment polarity across all user-friend pair. This might be the reason behind linguistic features like *net comment ratio* and *agreement rank* not producing good accuracy in our model.

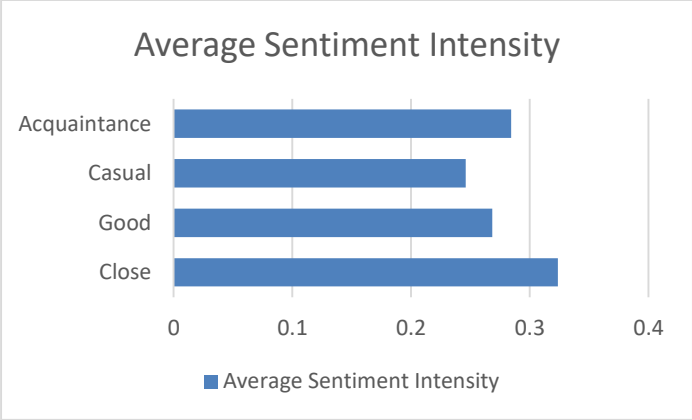


Figure 34. Average sentiment intensity among different categories of friend

It is noted that we did not put forward all different combinations of friend classes in the results section. Our dataset showed that close, good and casual user-friend pairs have similar interactions while acquaintance user-friend pairs did not show any such interactions. Hence, there was a poor performance in classifying close vs good, casual vs acquaintance etc. Although the interaction variable works better in classifying strong and weak friend combination as we did in the results section, it does not seem to work best with the 4-class classification.

Our assumption that BOWM should perform better failed in all the dataset combinations. Since there were very few user-friend pairs with the closeness related comments exchanged between them, the classifiers failed to learn much from bag of words as a feature. We also saw that the model containing all training features, performed more weakly than other models resulting in many false positives. This was mainly because most of the features did not correlate to the friendship strength.

Out of the different classifiers used in our experiment, random forest surpassed the performance of both SVM and logistic regression in almost all cases. Random forest produced best classification results while classifying close and acquaintance friends with an area under ROC curve of 0.85 and accuracy of 0.85.

Jones et.al [9] research of inferring tie strength from online-directed behavior achieved the area under ROC of 0.92 between strong and weak friends. They pointed *media multiplexity* as the primary reason for their success. In their research, online Facebook interactions were easily measurable with offline Facebook interactions. In addition, they have used private Facebook messages sent between the user-friend pairs as one of the features. Our dataset of a 680-user friend pairs were obtained within a very small group of former students of Boise State University. We did not see much *media multiple*xity in our dataset. There were close friends in real life without any interactions in Facebook while there were casual and acquaintances friends having a large number of interactions. Although we removed close friends having zero interactions in the due course of experiments, we believed the acquaintance and casual friends with many interactions were the primary reason for false positives. In addition, features like *average comment length, likes, love, friend posts, mutual friends* and *closeness variable* produced a good predictive power while other features including homophilic ones did not produce much predictability in our dataset.

Limitation

The biggest limitation of our research is the dataset. Sampling was conducted in Boise State University with exclusive inclusion of native English speakers in the sample. This compromises the variability of the study. As the sample group does not represent the true population, caution must be taken while generalizing the results of this experiment.

CHAPTER FIVE: CONCLUSION AND FUTURE WORK

In this paper, we developed a friendship strength model to evaluate the strength of friendship in social media by utilizing interaction, structural, homophily and linguistic based features. Work done by Gilbert et.al [4] in Facebook has utilized many feature variables for predicting tie strength between friends and obtained accuracy of 80%. Jones et.al [9] research of inferring tie strength have obtained accuracy of 86% utilizing different features. Their work is dependent upon the intimate feature variables, which they extracted between user friend pairs. With the time constraint of researchers, and current advancements in privacy settings in Facebook, those variables could not be extracted in a limited period. However, we developed a new set of features that were not used before in predicting tie strength. In addition, none of the previous studies have used linguistic features to predict the tie strength.

We have obtained accuracy of 85% in successfully identifying the close and acquaintance group of friends by using interaction, structural and linguistic features. Our experiments clearly depict that features like *average comment length, likes, love, friend posts, mutual friends* and *closeness variable* had a positive correlation with friendship strength. Our experiment also shows that linguistic features help to improve the performance of the model. Therefore, from this paper, we suggest that combining these language features along with other features would produce more accurate and trustworthy model to evaluate the friendship strength.

In addition, our model only worked for two classes of friends. The current set of features were not able to classify all four classes of friends with proper accuracy. We believe that more-complicated language based features in an even larger data set could produce identifiable results. Since average sentiment polarity across user friend pair tends to be same (see Figure 34), we would like to understand how this sentiment varies across different topics for different friend categories. In our future work, we would like to conduct a similar experiment using diverse dataset to determine whether topic modeling can actually predict friendship strength. Brief detail in topic modeling is defined in Appendix B. In addition, since we did not classify the posts written by users, classification of user posts into different categories would also be in our future work.

REFERENCES

[1] Akbas, M. I., Avula, R. N., Bassiouni, M. A., & Turgut, D. (2013). Social network generation and friend ranking based on mobile phone data. In *Communications (ICC), 2013 IEEE International Conference on* (pp. 1444–1448). IEEE.

[2] Dickerson, J. P., Kagan, V., & Subrahmanian, V. S. (2014). Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 620–627).

[3] Dindia, K., & Canary, D. J. (1993). Definitions and theoretical perspectives on maintaining relationships. *Journal of Social and Personal Relationships*, *10*(2), 163–173.

[4] Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 211–220). ACM

[5] Hogenboom, A., Bal, M., Frasincar, F., Bal, D., Kaymak, U., & Jong, F. De. (2014). Lexicon-based Sentiment Analysis by Mapping Conveyed Sentiment to Intended Sentiment. *Int. J. Web Eng. Technol.*, *9*(2), 125–147.

[6] Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[7] Tanbeer, S. K., Leung, C. K.-S., & Cameron, J. J. (2012). DIFSoN: discovering influential friends from social networks. In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on* (pp. 120–125). IEEE.

[8] West, R., Paskov, H. S., Leskovec, J., & Potts, C. (2014). Exploiting social network structure for person-to-person sentiment analysis. *arXiv Preprint arXiv:1409.2450*.

[9] Xiang, R., Neville, J., & Rogati, M. (2010). Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web* (pp. 981–990). ACM.

[10] Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C., & Fowler, J. H. (2013). Inferring Tie Strength from Online Directed Behavior. *PLoS ONE*, *8*(1), 1–6.

[11] Arnaboldi, V., Guazzini, A., & Passarella, A. (2013). Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook. *Computer Communications, 36*(10–11), 1130–1144.

[12] Le, Q. V, & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv Preprint arXiv:1405.4053*.

[13] Alsmadi, I., Xu, D., & Cho, J.-H. (2016). Interaction-based Reputation Model in Online Social Networks. In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy* (pp. 265–272).

[14] Sherchan, W., Nepal, S., & Paris, C. (2013). A Survey of Trust in Social Networks. *ACM Comput. Surv., 45*(4), 47:1–47:33.

[15] Singh, S., & Bawa, S. (n.d.). A Privacy, Trust and Policy based Authorization Framework for Services in Distributed Environments.

[16] Golbeck, J. A. (2005). *Computing and Applying Trust in Web-based Social Networks*. University of Maryland at College Park, College Park, MD, USA.

[17] Golbeck, J. (2006). Generating Predictive Movie Recommendations from Trust in Social Networks. In *Proceedings of the 4th International Conference on Trust Management* (pp. 93–104). Berlin, Heidelberg: Springer-Verlag.

[18] Cameron, J. J., Leung, C. K. S., & Tanbeer, S. K. (2011). Finding Strong Groups of Friends among Friends in Social Networks. In *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing* (pp. 824–831).

[19] Jiang, F., Leung, C. K. S., & Tanbeer, S. K. (2012). Finding Popular Friends in Social Networks. In *2012 Second International Conference on Cloud and Green Computing* (pp. 501–508).

[20] Tanbeer, S. K., Jiang, F., Leung, C. K. S., MacKinnon, R. K., & Medina, I. J. M. (2013). Finding groups of friends who are significant across multiple domains in social networks. In *2013 Fifth International Conference on Computational Aspects of Social Networks* (pp. 21–26).

[21] Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 74–77). New York, NY, USA: ACM.

[22] Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics.

[23] Onnela J, Saramaki J, Hyvonen J, Szabo G, Lazer D, et al.. (2007) Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of the United States of America, 104(18): 7332–7336.

[24] Zhang, H., & Dantu, R. (2010). Predicting social ties in mobile phone networks. In *2010 IEEE International Conference on Intelligence and Security Informatics* (pp. 25–30).

[25] Baym, N.K., Ledbetter, A (2009) Tunes that Bind? Information, Communication & Society, Vol 12(3), 408-427.

[26] C.K.-S. Leung & S.K. Tanbeer, "Mining social networks for significant friend groups," in DASFAA 2012 Workshops, pp. 180-192.

[27] M. Granovetter. The strength of weak ties: A network theory revisited. Sociological Theory, 1:201–233, 1983.

[28] H. Kashima and N. Abe. A parameterized probabilistic model of network evolution for supervised link prediction. In ICDM ’06, 2006.

[29] S.G. Roberts, Constraints on Social Networks, in: Social Brain, Distributed Mind, Proceedings of the British Academy, 2010, pp. 115–134.

[30] R.A. Hill, R.I.M. Dunbar, Social network size in humans, Human Nature 14 (2003) 53–72.

[31] Mehto, A., & Indras, K. (2016). Data mining through sentiment analysis: Lexicon based sentiment analysis model using aspect catalogue. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* (pp. 1–7).

[32] Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). SentiFul: A Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, *2*(1), 22–36.

[33] Avanco, L. V, & d. G. V. Nunes, M. (2014). Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese. In *2014 Brazilian Conference on Intelligent Systems* (pp. 277–281).

[34] Bhoir, P., & Kolte, S. (2015). Sentiment analysis of movie reviews using lexicon approach. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1–6).

[35] David M. Blei , Andrew Y. Ng , Michael I. Jordan, Latent dirichlet al.location, The Journal of Machine Learning Research, 3, p.993-1022, 3/1/2003

APPENDIX A

Evaluation Metrics

ROC Curve

ROC or receiver operating characteristic curve is a two-dimensional curve, which shows the performance of binary classification under varying discrimination threshold[17]. It is the plot of true positive rate (TPR) against the false positive rate (FPR) at different threshold values. In a classifier, accuracy is sensitive to the imbalance in classes. For example, in 100 data samples, if 80 of them are positive and 20 are negative, the classifier results in an accuracy of 80 percent at least. This means that the accuracy is only showing the distribution of classes in the dataset. With 80 percent data as positive, the probability of getting a positive sample is already 80 percent. The ROC curve is insensitive to this class imbalance.

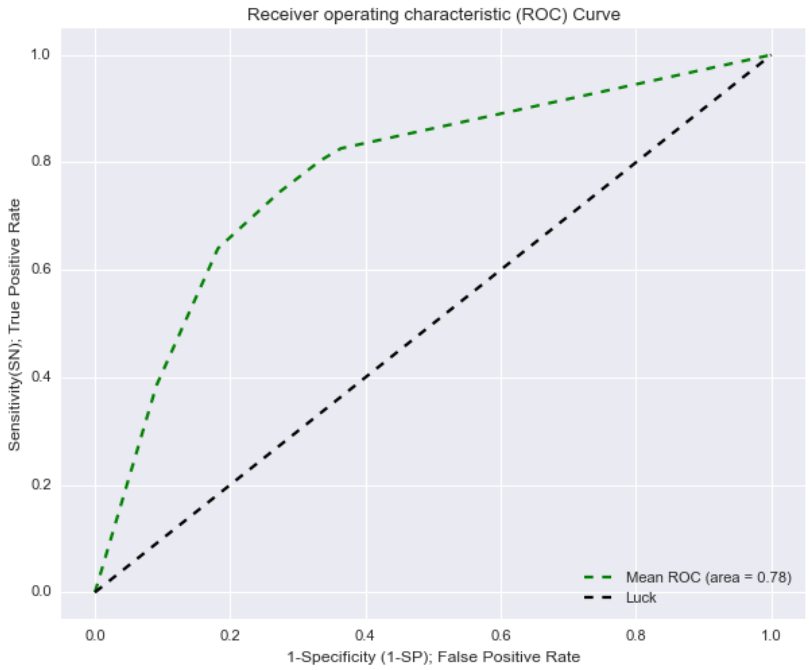


Figure 35. Sample ROC Curve

[17] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Figure 35 represents the sample ROC curve with an area of 0.78. The luck line shown in the diagram represents any classifier with random performance level. This is a baseline representing the performance level of the classifier. ROC curve in the upper left corner represents a good classification, while the ROC curve in the lower right corner represents poor classification.

AUROC (Area under the ROC curve)

The area under the ROC curve represents the probability that the classifier ranks a randomly chosen positive sample higher than the randomly chosen negative sample[18]. The AUROC of the excellent classifier is 1. The dotted blue line in Figure 35 has an area of 0.5. Therefore, any random predictor has AUROC of 0.5, which is used as a baseline in identifying the usefulness of the model.

Confusion Matrix

Binary classification produces four possible outcomes.

i. True Positive (TP): Positive samples predicted as positive
ii. False Positive (FP): Negative samples predicted as positive
iii. True Negative (TN): Negative samples predicted as negative
iv. False Negative (FN): Negative samples predicted as positive

A 2 by 2 table showing these four outcomes of the binary classification is called confusion matrix. Table 16 shows the confusion matrix with four possible outcomes.

[18] https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve

Table 16. Confusion matrix

		Predicted samples	
		Positive	Negative
Ground truth	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

True Positive Rate (TPR) or Sensitivity or Recall

The true positive rate is the ratio of a number of correctly predicted positive samples to the total number of positive samples. The maximum value of TPR is 1 when FN is 0 while the minimum value is 0 when TP is 0.

i.e $TPR = \frac{TP}{TP+FN}$

False Positive Rate (FPR)

False positive rate is the ratio of a number of incorrect negatively predicted positive samples to the total number of negative samples. The maximum value of FPR is 1 while the minimum is 0.

i.e. $FPR = \frac{FP}{TN+FP}$

False positive rate is also defined as 1-specificity.

Specificity or True Negative Rate (TNR)

It is the ratio of a total number of correctly identified negative samples to the total number of negative samples in the dataset. Specificity is defined as

$$TNR = \frac{TN}{TN+FP}$$

Precision or Positive Predictive Value (PPV)

Precision is the ratio of a number of correctly predicted positive samples to the total number of positively predicted samples. This measures the probability about how relevant is the retrieved document.

i.e. $PPV = \frac{TP}{TP+FP}$

F-Measure

This measure is the harmonic mean of the precision and recall and is given by

F-Measure = $\frac{2.Precision.Recall}{Precision+ Recall}$

APPENDIX B

Topic Modelling

Introduction

In Natural Language Processing, Topic modeling is a clustering approach to automatically extract topic from the documents. It is useful for search or browsing. Topic modeling uses a most common method called LDA (Latent Dirichlet al.location) [35] that is a probabilistic model. The LDA model assumes that documents are just the collection of topics and each topic has some probability of generating a particular word.

For example, in social media people comment on varieties of topics. With the help of LDA, we could generate random N number of topics. Since topic modeling itself looks upon the probability distribution of words to extract topics, we can look at the distribution of words across the topics to infer topic as political, technological, sports etc. It is not necessary for each topic generated to be of particular meaning. Having said that topic generation is an iterative and evaluative process.

Topic modeling can be dealt with 2 techniques. First, non-negative matrix factorization[19] as available in python scikit library[20] while second is genism LDA model. The visualizing of the topics model could be done using pyLDAvis[21] library in python, which provides interactive D3[22] based web visualization. This visualization provides the global view of the topics, through the similarities and prevalence with each other in a limited space [20] [22] .

---

[19] https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

[20] http://scikit-learn.org/stable/auto_examples/applications/topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-topics-extraction-with-nmf-lda-py

[21] https://github.com/bmabey/pyLDAvis

[22] https://d3js.org/

While performing topic modeling, we would gather all the comments from our user friend pair and discover topics that occur in the collection of comments. We would try to generate different subsets of topics (10, 20, 30 and 40) and evaluate sentiments of the different friends with respect to the topics. For example, if t1, t2, t3, t4 and t5 are 5 topics then we would model as follows.

Table 17. Example of Topic Modelling of comments across different friend categories

Friend category	**Topics**				
	t1	**t2**	**t3**	**t4**	**t5**
Close	positive			Extremely positive	negative
Good		negative	negative		
Casual					
Acquaintance			Extremely negative		positive

Table 17 shows four different friend categories and the sentiment of the topic they are talking about. For example, close friends are talking positively about topic 1 (see blue box in Table 17) and acquaintances are talking extremely negative about topic 3 (see blue box in Table 17). It is likely that close friends may not talk about topic 2 and topic 3 so the spaces are left blank. The purpose of doing this would be twofold. First, to find which topics the dataset population is talking about and second, to find the anomalies in the sentiment patterns of friends across different topics.

Topic Modeling in Current Dataset

Topic modeling is a clustering approach to separate the documents into a number of topics. We have generated 50 different topics from the comments of all 680-user friend pairs. Each topic can mean something. Figure 36 represents topic cloud representing 50 different topics from our dataset. The distance between the circles represents the inter-topic distance. If the circles are close to each other, it means the topic share similar words. While if the circles are far from each other it means the topic are quite distinct. Topic modeling is an iterative approach and hence the entire topic generated may not mean something.

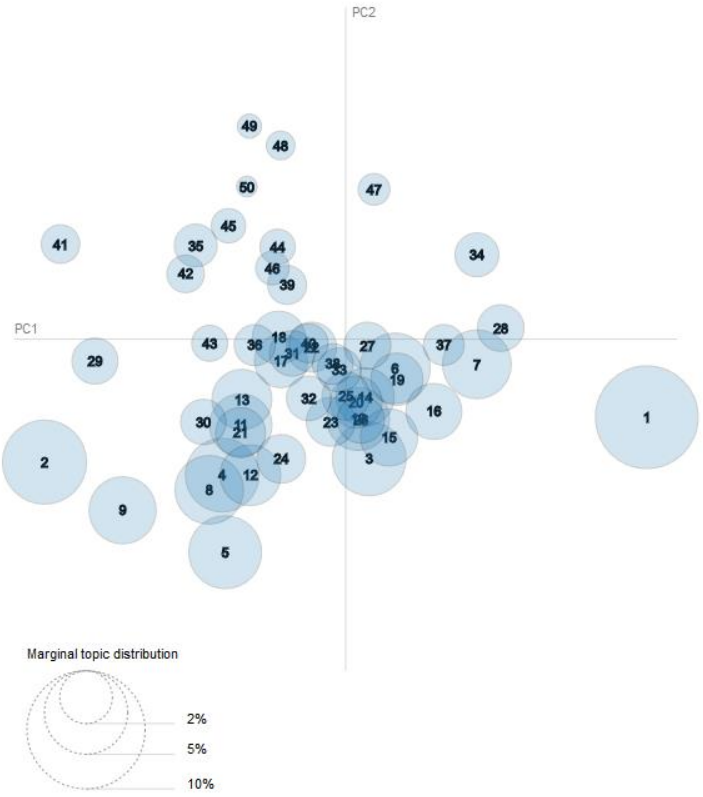


Figure 36. Inter topic distance map showing distribution of 50 different topics

Each topic generated during topic modeling contains probability distribution of relevant words/tokens/terms. Figure 37 represents the frequent words in topic 1. By

observing the tokens we can infer that topic 1 may be related to proximity. In our dataset, the tokens represented in topic 1 represents 7.5 percent of the total tokens generated from the comments. Different categories of friends might have positive or negative sentiments towards this topic, which can be the matter of further investigation.

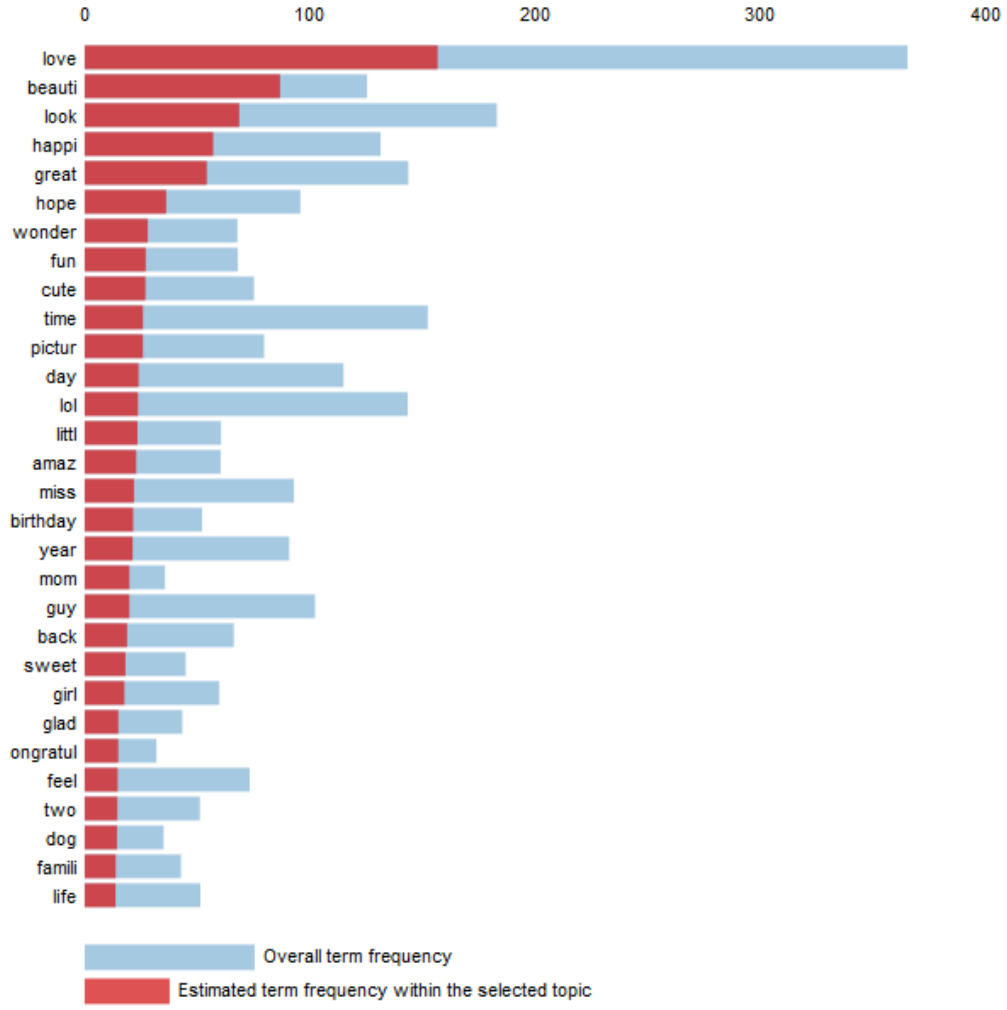


Figure 37. Top-30 relevant terms for topic 1 identified as close topic

Similarly, Figure 38 represents the top 30 terms in topic 2. By observing the tokens in the following figure we can infer that the topic relates to politics and economics. These are the topics that casual and acquaintances discuss the most. Evaluating the sentiments of different categories of friends across these topics could

bring further insights. These relevant terms in topic 2 represent 5.1% of the whole comments.

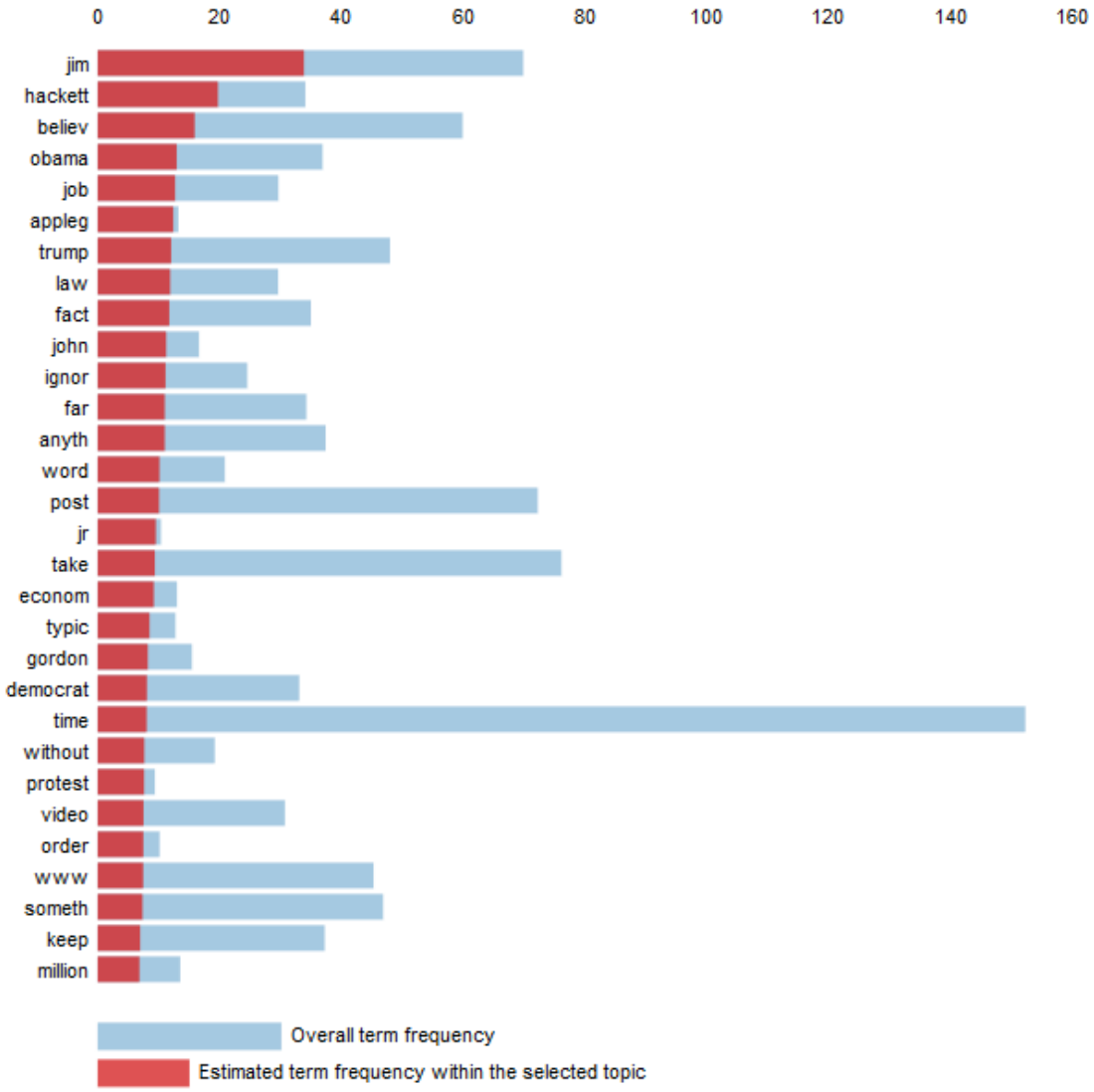


Figure 38. Top 30 relevant terms for topic 2 identified as a political topic.

APPENDIX C

Institutional Review Board

This research was conducted under the approval of the Institutional Review Board at Boise State University, protocol # 131-SB16-112.