WHY DON'T YOU ACT YOUR AGE?

RECOGNIZING THE STEREOTYPICAL 8-12 YEAR OLD SEARCHER BY THEIR SEARCH BEHAVIOR

by

Michael Green

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

August 2021

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Michael Green

Thesis Title: Why Don't You Act Your Age? Recognizing Stereotypical 8-12 Year Old Searchers By Their Search Behavior

Date of Final Oral Examination: 14th August 2021

The following individuals read and discussed the thesis submitted by student Michael Green, and they evaluated the presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Maria Soledad Pera, Ph.D.	Chair, Supervisory Committee
Michael Ekstrand, Ph.D.	Member, Supervisory Committee
Casey Kennington, Ph.D.	Member, Supervisory Committee

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

Dedicated to Keith Baderman and Marino Francioni.

ACKNOWLEDGMENTS

The author wishes to express gratitude to Dr. Maria Soledad Pera, Shane Panter, and Luke Hindman. Your passion and encouragement have continually inspired. The author further wishes to thank Ashlee Milton and Garrett Allen, for their support. Making the process easier by sharing the weight, emotionally and scientifically. Furthermore, the author wishes to express gratitude to Boise State University's Computer Science department, which has funded this research. Thank you.

ABSTRACT

Online search engines for children are known to filter retrieved resources based on page complexity, and offer specialized functionality meant to address gaps in search literacy according to a user's age or grade. However, not every searcher grouped by these identifiers displays the same level of text comprehension, or requires the same aid with search. Furthermore, these search engines typically rely on direct feedback to ascertain these identifiers. This reliance on self identification may cause users to accidentally misrepresent themselves. We therefore seek to recognize users from skill based signals rather than utilizing age or grade identifiers, as skill dictates appropriate aid and resources. Therefore, in this thesis we propose a strategy that automatically recognizes users on the fly by analyzing search behavior found in search sessions. In particular, our efforts focus on recognizing the stereotypical 8 to 12 year old searcher, who we posit exhibits skills defined by developmental stages that have a strong impact on language development (Piaget's concrete operational stage) and search literacy (digital competency's first level). This strategy analyzes user-generated text extracted from queries and patterns of search interactions in order to infer features that are leveraged by a random forest classifier in order to determine whether or not a user is a part of this specific segment of searchers. The outcomes from this thesis lay the groundwork for enabling search engines to recognize users based on their search skills and provides further insight into the search behavior of youths.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Children in grades K-8[1] regularly use Search Engines (**SE**) [28, 63, 64] specifically designed for them in order to complete school assignments as well as satisfy their curiosity [6, 27, 59]. However, these young searchers can encounter difficulties completing successful searches [30], in part due to their trouble formulating queries [21], comprehending retrieved results [10], or navigating SE results pages [34]. The root of these struggles often correspond to a user's skill, or lack thereof, in reading, writing, and search literacy [34]. Users that have the same set of skills typically encounter similar problems [7]. SE for children that offer algorithmic support to address these struggles [32] typically target users based on age or grade [21, 27, 32]. This includes strategies that ease query formulation [21], provide query suggestions that match children's search topics [49], collect and offer relevant and comprehensible resources via content curation [13], i.e., manually-collected websites for a selected audience; and employ adaptive interfaces [32], i.e., interfaces that allow retrieved content to be adjusted for a user based on their explicit feedback. However, targeting users by broad ranges of ages [8, 21] and grades [4] works under the assumption that these groups of users share an uniform set of skills. However, we know that users in the same age or grade have varying abilities in terms of reading, writing, and search literacy [28]. Resources and aid deemed fit for a user recognized by age/grade can

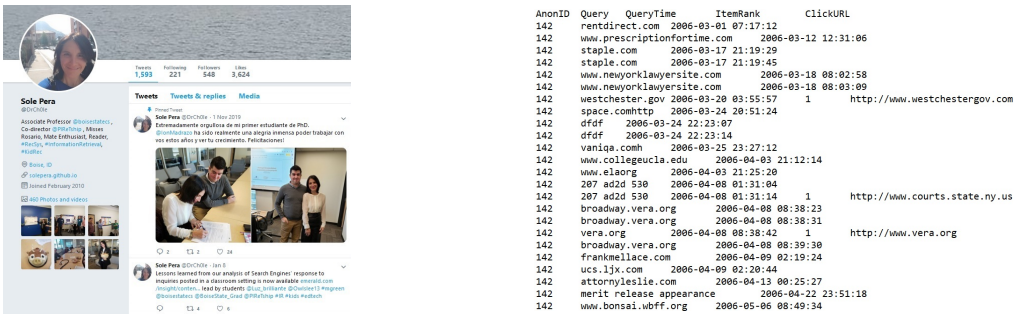
[1] In the US school system, children in K-8 are usually between 5 to 14 years of age.

fail to meet that user's true capabilities, providing content that could be well below or above their comprehension level or offer tools that can cause frustration rather than enable ease. Providing SE with a way to recognize their users capabilities could enable these platforms a way to ensure that the right resources are getting to the right users. We therefore see it as pivotal to determine how young searchers can be recognized by their potential skills, as there currently exists no strategy for doing so.

One way to recognize a user by their set of skills is identifying that user's type [8, 32]–a method of determining and leveraging key characteristics of users in order to differentiate one from another and place them into groups [46]. Grouping is often based on users that share common characteristics, such as age and gender [50, 67], personality [65], and level of expertise [69]. Research in domains outside of SE has explored approaches that are successful in the detection of children based on their interactions in: chat rooms [46, 67], social media websites [3], and online questionnaires [55]. Unfortunately, these user-system interactions differ greatly from user-SE interactions (as seen in Figure 1.1), typically containing paragraphs of text compared to text-sparse queries, preventing cross-domain application. Furthermore, these strategies utilize domain-specific historical data to make these distinctions (blogs, social media sites, chat rooms). Due to stringent measures by COPPA (Children's Online Privacy Protection Act), there is a lack of historical data documenting young searchers interactions with SE [26].

SE often address this lack of historical data by relying on the user to define their user type via direct feedback. Typically, this feedback is drawn from user profiles that contain the relevant user information [65, 66, 67]. However, SE for children address this lack of profile information by requiring a user to define their type upon entry [32, 4]. Unfortunately, this type of direct feedback can be unreliable as users

may lie about their identity [25], can lack the search literacy to properly self identify i.e. accidentally clicking the wrong age or grade and being unsure of how to undo that selection [8], or the current user may change after login. Additionally, direct feedback is typically used to recognize a user's age or grade, as these are correlated to skill. Even if a user is able to accurately self-identify their age or grade using direct feedback, we are still left with our initial problem. There is no guarantee that the user will "act their age" [43](or the age their grade correlates to) as development, instead of age, dictates how children behave, as well as defines their potential skills [56].



(a) Twitter (b) Search Engine

Figure 1.1: Visual example of the differences between data generated by Twitter usage (seen in Fig 1.1a), which contains numerous textual interactions, profile data, and images; versus data generated by interactions with SE (seen in Fig 1.1b), which contains queries, clicks, and timestamps.

We suspect that the differences in data available between domains disallows the one to one application of previously mentioned strategies for identifying user type and question the reliability of direct feedback. Furthermore, considering the strong limitations in regards to having access to the historical data of young searchers, we must consider alternative strategies in regards to determining their skills. One of these alternatives is leveraging inferred information about a user based on their interactions

with a system [69, 70]. For SE, these interactions are either implicit (such as what time the session started) or explicit (such as text from the query itself); and together comprise a user's search behavior. Typically, these interactions are archived in query logs; each entry usually containing a user id, a query, and a timestamp (an example can be seen in Figure 1.1b). Entries may then be grouped into sessions[2], providing a body of text to be analyzed and enabling the inference of search behavior (like number of queries in a session, or time between each query) that cannot be surmised from a single query [28].

This inferred information can be used by author profiling strategies [46, 67] or expert identification strategies [69, 71, 72] to identify a user's type. The former is the process of analyzing a *text* with express purpose of identifying attributes of an author. Researchers have used author profiling strategies to group users by types like: male or female [55], extroverted or introverted [65], and even teen or adult [46]. These strategies depend on large amounts of text, such as Facebook posts [65] or blog entries [50]. Unfortunately, search sessions rarely meet the lower limit of 100 words needed to effectively identify user type. Expert identification strategies focus on analyzing search behavior in order to determine a user's expertise in a particular domain [69, 71, 72]. Expertise can be established by inspecting features relating both to implicit search behavior, such as session length and average number of queries per session, as well as explicit search behavior, such as average query length or use of technical terms. Since children struggle in the domain of SE use [6, 30], their developing search literacy could provide an unique opportunity to identify their lack of expertise. However, these previous examples of expert identification are designed

[2]A search session can be defined as a set of queries generated by a user and grouped by the cumulative goal of fulfilling an information need or meeting a certain threshold of time.

to recognize experts in particular domains, not to establish search literacy. Due to the aforementioned limitations of author profiling and expert identification strategies, we suspect that neither type of strategy can be directly applied to recognizing the skills of SE automatically. However, both will inform our feature selection once we clearly define a user type that demonstrates a lack of search literacy and also displaying textual behavior that correlates to literary skills displayed in their queries.

One common method to recognize users who lack historical data is known as stereotyping. This methodology seeks to group users based on common characteristics that they all share and are expected of them. In order for us to determine a stereotype contained with in the umbrella term of "children" to recognize, we need to understand characteristics that define sub-groups with in this population, particularly attributes related to skill. Children's proficiency in certain skills are correlated to stages of development, as articulated by modern theories of psychology. Some of these stages are aligned to curriculum and correspond with what is taught in school, such as reading, writing, and typing [41]. Others deal with the cognitive development of their brain [56], changing how they think, allowing for abstract thought and lateral thinking. These stages may be dependent upon one another. For example: formulating keywords for search requires abstract thought (reducing a search to its component parts), but also requires search literacy (knowing that a keyword query will retrieve results). Theories of development are comprised of different stages which have clearly defined skill sets described in each stage.

Since children's ability to search is dependent on their ability to craft queries, both in the terms of search literacy [28] as well as language development [31], in this thesis we propose a classification strategy called **RYSe** (Recognizing Young Searchers), intended to identify the average 8-12 year old searcher based on their

search behavior. This stereotype is selected as it encompasses stages of development that dictate language development (Piaget's third stage: the concrete operational stage [56]), as well as search literacy (the first level of digital competency [15]). To clarify, we seek to recognize users who display the traits defined by these stages but as these stages of development are complex and nuanced we do not seek to explicitly label users as belonging to these stages. Hence the use of the stereotype. Furthermore, we set our scope to recognize users based on their display of American English.

As Piaget's third and the first level of digital competency define skills that correlate to language development and search literacy, we draw inspiration from author profiling and expert identification strategies when selecting features that recognize these skills. Children in the concrete operational stage display a fluency in writing that grants them the ability to craft queries [18] but will also struggle with abstract thought. As such, we investigate features related to crafting queries such as lexicon and vocabulary, while also investigating features that deal with the differences between natural language and keyword queries such as syntax. We also consider characteristics unique to children's lack of expertise in the domain of SE usage. The first level of digital competency states that a user can, with assistance: identify an information need, find and access data through simple searches, and navigate between these sources of data. Any difficulty in accomplishing any of these tasks will reflect in both their implicit and explicit search behavior, allowing us to recognize users that struggle with search literacy [6]. Therefore we consider search operators, query reformulation, and temporal features to determine if a user is at the first level of digital competency. All of these features are then utilized in a Random Forest Classifier [45]. Since the Random Forest Classifier randomly sub-samples features for each tree, it can be populated with trees that special in recognizing recognizing users based on particulars subsets

of the features we investigate [48].

By identifying the average 8-12 year old search from their search behavior, we showcase that children can be identified by their display of skills defined by specific stages of development. Through the recognition of one stereotype of young searcher based on their skills, the door opens to identifying other non-traditional stereotypes based on their search behavior such as: searchers under the age of 8, teenagers in general, as well as the elderly. RYSe also provides SE for children and strategies designed to aid young searchers with the ability to recognize the skill of users who need aid. This can enable SE to provide skill appropriate resources to these users. Moreover, outcomes from this thesis provide insight into the search behavior of children.

1.1 Thesis Statement

We hypothesize that the average 8-12 year old searcher can be recognized by examining search behavior inferred from user-to-search-engine interactions for displays of search literacy and language development which correlate to the stages of development we expect this stereotype to be in: Piaget's third stage and at the first level of digital competency.

CHAPTER 2

BACKGROUND AND RELATED WORK

While identifying users over the age of 13 from online text found in blogs and chat rooms has been a focus of research since the mid-2000s [46, 67, 69], recognizing the stereotypical 8-12 year old searcher from their interactions with SE remains an unrecognized possibility that RYSe seeks to explore. In this section, we discuss key concepts that offer background and inform the design of RYSe.

2.1 Query Logs and Search Sessions

Search interactions are defined as an explicit interaction a user has with a SE, such as submitting a query or clicking on a result. These interactions can be stored in an itemized list of entries known as query logs [23, 39, 61, 69, 70]. These logs not only contain queries (a string containing input text), but can also contain a time stamp (reflecting date and time), user id (or session ID, delineating one user/session from another), query (containing the text used to perform the search), URL clicked (if any selected, as not all queries yield clicked results), and ordinal value of URL clicked (if any are selected). Entries can then be grouped into search sessions, defined as a "sequence of pages visited by a single user at a single web site for a specified length of time" [58]. While there are many approaches for determining what constitutes a

search session [29], we focus on time based strategies as research demonstrates that time based strategies accurately match user's patterns of activity [33].

2.2 Theories of Development

It has been speculated that user's SE interactions can be indicative of their current stage of education and development [23, 25, 69]. Numerous theories exist with the express purpose of identifying and categorizing these stages: Piaget's 4 stages of cognitive development [56], which focuses on cognitive skills that enable language acquisition; and Cooper and Kiger's 5 stages of literacy [18] as well Kroll's 4 stages of writing development [41], which both describe skill sets used in comprehending and expressing language. We see the 5 stages of literacy and the 4 stages of writing as encompassed within the 4 stages of cognitive development, as cognitive development informs language development. The 4 stages of writing development dictate an individuals ability to express themselves in written language, while the 5 stages of literacy determine their ability to read describe the skills necessary to express language. Furthermore, there exists the 8 levels of digital competency, which define a person's search literacy, as established by Carretero et al. [15]. Digital competency is unique among these aforementioned stages as the skills dictated by this theory can be acquired at any age.

All of these stages can have an impact on SE use, whether it is abstraction for keyword formulation (defined by cognitive development), reading and writing abilities (defined by literacy and writing development), or search literacy (defined by digital competency). Furthermore, stages of development may not only overlap with each other, but also have curriculum that lines up with these developing skills

sets. For example, Common Core State Standards Initiative (an American education curriculum that prescribes vocabulary by grade level) [38], and Age of Acquisition (a pyscholinguistic theory that ascertains the common age at which words are learned) [35] correlate to stages of writing and literacy development.

Piaget's (and subsequently Kroll's and Cooper and Kiger's) stages of development, as well as correlated curriculum, emphasize how queries generated by users can provide distinct clues to the stages of development as user is likely to be in. The same can be said of digital competency, as the skills defined can be seen in displays of search literacy. However, the stages in these previously mentioned theories of development are are soft-bounded. In order to determine if a user is displaying search behavior associated with certain stages of development, a clearly defined user type must be established. This enables us to determine what stages of development that user is expected to display characteristics of being in.

2.3 Identifying User Type

One method for addressing the identification of an ambiguous user type is known as stereotyping [16]. Originating from the domain of recommendation, the basis of this approach is grounded in grouping a user with others based on common characteristics they all share. These characteristics can be related to "demographic, geographic, or psychographic information" [16]. Doing so allows users to be recognized based on shared characteristics. Therefore, segmenting users based on their stereotype enables the identification of that user type. Since "children" is a broad user type, establishing a stereotype contained within this spectrum allows us to clearly state the skills defined stages of development that we intend to recognize.

The field of author profiling, also known as stylometrics, is interested in ascribing authorship to texts of unknown origin. By analyzing user generated text for defining characteristics, it is possible to determine attributes of an author such as their age and gender [46, 50, 67] and personality [65]. Researchers have gone so far as to identify online predators [12, 37, 54] and bullies [40]. Existing author profiling strategies usually rely on binary classifiers trained on textual features inferred from user generated text found in sources such as blogs, chat logs, and social media websites [46, 67, 68]. Yet, to our knowledge, none of these author profiling strategies identify the user type of searchers based on search queries. Given that users' search sessions captured in query logs have only a fraction of the user generated text that the aforementioned strategies require, we posit that one to one application of these these strategies would not be effective in the domain of SE. This limitation inspires us to consider a wider array of textual features than those found in the aforementioned examples, while also prompting the consideration of non-textual features to be used by RYSe.

There do exist strategies that leverage more than just textual features in order to recognize user type. These strategies demonstrate methods of identification that rely on supplementing textual features with information unique to the SE environment, such as sites visited or time spent searching; in order to identify users by their age and gender [23], domain expertise [69] or knowledge gain [70]. Of particular note is expert identification, which is shown to identify experts in a wide variety of domains, which we hypothesize can be leveraged to recognize expertise in skill sets directly correlating to stages of development [69]. While expert identification strategies show that it is possible to recognize expertise, there does not currently exist a strategy to determine the expertise in the domain of search literacy, specifically amongst children.

CHAPTER 3

METHOD

RYSe recognizes the stereotypical 8-12 year old searcher from their search behavior by analyzing different aspects of user-SE interactions. We focus on recognizing skills expected to be displayed by searchers in Piaget's third stage (P3), which correlates to language development, as well as at the first level of digital competency, which correlates to search literacy (DC1). These skills are inferred from features found in users' generated text (described in section 3.1); as well their search interactions (described in section 3.2).

3.1 Text Based Features

As previously mentioned, numerous strategies for author profiling demonstrate that user generated text can be used to recognize user type from chat logs, blogs, and social media sites [46, 67, 68]. We draw inspiration from these strategies, analyzing user generated text, i.e. search queries, in order to recognize behavior common to our user type. This analysis focuses on four groups of features: Lexical, Syntactical, Spelling Errors, and Vocabulary. These features are computed on a query by query basis; for sessions that include more than a single query, we calculate the mean value for each feature across each query in the respective session.

3.1.1 Lexical Features

Lexical features allow us to assess a user's sophistication and diversity of vocabulary when they express themselves with in written American English. Determining these abilities requires us to examine the lexical richness of text generated by individuals [47].

Lexical richness is comprised of three primary parts: lexical density, lexical sophistication, and lexical variation. Lexical density examines the ratio of "nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, "be," and "have"), and adverbs with an adjectival base, including those that can function as both an adjective and adverb (e.g., "fast") and those formed by attaching the -ly suffix to an adjectival root (e.g., "particularly")" [47] to words in a sample of text. In our case, this ratio of lexemes to total words per query reveals a capacity for articulation which is limited in an individual still learning a language. Lexical sophistication is defined as "the proportion of relatively unusual or advanced words in the learner's text" [52], demonstrate a user's familiarity with a language. Lexical variation "refers to the range of a learner's vocabulary as displayed in his or her language use" [47] and not only demonstrates a grasp of language but also helps to highlight the differences between keyword and natural language queries.

Beyond lexical richness, we also consider lexical characteristics of queries which have been used to determine domain expertise [69] and further highlight differences between natural language and keyword queries. These features are important, as not all lexical differences between those two types of queries may be captured in lexical richness. A detailed explanation on how to compute each of the lexical-related features can be found in Table 3.1.

Feature Name	Feature Computation	Feature Type	P3	DC1
Lexical density	N_{lex}/N	Density	✓	
Lexical sophistication-I	S_{lex}/N_{lex}	Sophistication	✓	
Lexical sophistication-II	SW_{types}/N_{types}	Sophistication	✓	
Verb sophistication-I	SV_{types}/V	Sophistication	✓	
Corrected VS1	$SV_{types}/\sqrt{2V}$	Sophistication	✓	
Verb sophistication-II	SV^2_{types}/V	Sophistication	✓	
Number of different words	T	Variety	✓	✓
Type-Token Ratio (TTR)	T/N	Variety	✓	✓
Corrected TTR	$T/\sqrt{(2N)}$	Variety	✓	✓
Root TTR	$T/\sqrt{(N)}$	Variety	✓	✓
Lexical word variation	T_{lex}/N_{lex}	Variety	✓	✓
Verb variation-I	T_{verb}/N_{verb}	Variety	✓	✓
Squared VV-I	T^2_{verb}/N_{verb}	Variety	✓	✓
Corrected VV-I	$T_{verb}/\sqrt{(N_{verb}}$	Variety	✓	✓
Verb variation-II	T_{verb}/N_{lex}	Variety	✓	✓
Noun variation	T_{noun}/N_{lex}	Variety	✓	✓
Adjective variation	T_{adj}/N_{lex}	Variety	✓	✓
Total number of words	N	Characteristic		✓
Number of characters	N_{char}	Characteristic		✓
Average word length	N_{char}/N_{words}	Characteristic		✓
Number of syllables	N_{syl}	Characteristic		✓
Average syllable per word	N_{syl}/N_{words}	Characteristic		✓
Easy words	N_{sim}	Characteristic		✓
Complex words	N_{com}	Characteristic		✓
Max syllables per word	$ArgMax(N_{syl})$	Characteristic		✓
Min syllables per word	$ArgMin(N_{syl})$	Characteristic		✓

Table 3.1: Lexical features, where sophisticated word (SW_{types}), lexical word (S_{lex}) and verb types (SV_{types}), which are defined as words or types not found in the "2,000 most frequent words of the British National Corpus" [44], N_{sim} and N_{com} are defined as words less than 3 syllables and words 3 syllables or more, respectively.

3.1.2 Vocabulary

The vocabulary searchers use when formulating their queries provides us with insights into the developmental state of users, both in terms of their ability to leverage language and craft queries. We therefore examine vocabulary found in queries from multiple perspectives in order to determine what features to inspect. At each grade level students are expected to know certain vocabulary. This leads us to consider features that count the occurrence of terms defined by Common Core State Standards Initiative [38], which can be found in Common Core Vocabulary lists [1]. These standards dictate vocabulary that should be taught and learned by a specific grade. Furthermore, users are also expected to know certain words by certain ages as defined by the Age of Acquisition (AoA) [35], a pyscholinguistic variable that dictates the average age that individuals are expected to learn certain words. AoA ratings are also used to determine the complexity of a query, which is correlated to search expertise [70]. As such we consider features that use the 50K word AoA dataset[2]. We also know that children tend to use different vocabulary than adults when searching [49], therefore we count the occurrence of words per query found in the most common words found in children's websites from the "Looking for the Movie Seven or Sven from the Movie Frozen?" dataset[3]. To further examine vocabulary used by our stereotype in their searches, we extract the top 250 words found in queries generated by the stereotypical 8-12 year old searcher, as well as the top 250 words used by users who are not our stereotype; and compute features related to these two lists. The threshold of 250 is established after locating where the Zipf's distribution curve flattens on the list of words in our stereotype's query. We further extract the top 50 word

[1]Vocabulary lists found at https://www.flocabulary.com/wordlists/.
[2]This dataset can be found at http://crr.ugent.be/papers/AoA_51715_words.zip.
[3]This dataset can be found at https://scholarworks.boisestate.edu/cs_scripts/5/

bi-grams for users who are not our stereotype; based on the same premise and using Zipf's distribution to determine threshold. Expanding on the premise of examining the vocabulary of queries for vocabulary commonly found in queries, we perform a more fine-grained approach to recognizing the vocabulary used, and not used, by our stereotype. This involves calculating TF-IDF values of words used by all users, words used by our the stereotypical 8-12 year old searcher, as well as words used by users that are not our stereotype.

Domain specific word usage has been used to identify domain expertise on selected domains such as medicine and computer science [69]. As such we examine vocabulary that contains search operators and url prefixes/suffixes which serve as indicators of digital competency [14]. Any user employing these prefixes/suffixes in their queries demonstrate a level of search expertise beyond DC1. To further distinguish between natural language queries and keyword queries, but also highlight search expertise, we use the NLTK [11] stop word list to count the occurrences of stopwords per query. Users proficient in the use of SE will avoid the use of stop words. Finally, we inspect queries to determine if they contain an interrogative word as these words are commonly found in keyword queries. A detailed explanation on how to compute each of the vocabulary-related features can be found in Table 3.2

3.1.3 Spelling and Punctuation

The stereotypical 8-12 year old user makes spelling errors. This is due not only to their limited but growing skill in writing and typing [31], but also can emerge from their struggle to use the "are you searching for" and auto-correct functionality of some SE [34]. This motivates us to consider features which explore different types of misspellings found in queries.

Feature Name	Formula	P3	DC1
Ratio of words found in core vocabulary list	N_{core}/N	✓	
Ratio of words not found in core vocabulary list	$N_{notCore}/N$	✓	
Lowest AoA rating	AoA_{min}	✓	✓
Highest AoA rating	AoA_{max}	✓	✓
Query complexity (average AoA rating)	AoA_{avg}	✓	✓
Ratio of words with AoA rating less than 13	$(N_{AoA} < 13.00)/N$	✓	
Ratio of words found in "Seven or Sven" dataset	N_{sven}/N	✓	
Number of words found in top 250 stereotype words	N_{250}	✓	
Ratio of number of words found in top 250 stereotype words	N_{250}/N	✓	
Ratio of number of words not found in top 250 stereotype words	$1 - (N_{250}/N)$	✓	
Number of words found in top 250 non-stereotype words	$N_{250_{ns}}$	✓	
Ratio of number of words found in top 250 non-stereotype words	$N_{250_{ns}}/N$	✓	
Ratio of number of words not found in top 250 non-stereotype words	$1 - (N_{250_{ns}}/N)$	✓	
Number of words found in top 50 stereotype word bi-grams	N_{250bi}	✓	
Ratio of number of words found in top 50 stereotype word bi-grams	N_{50bi}/N	✓	
Ratio of number of words not found in top 50 word bi-grams	$1 - (N_{50bi}/N)$	✓	
Number of words found in top 50 non-stereotype word bi-grams	$N_{50bi_{ns}}$	✓	
Ratio of number of words found in top 50 non-stereotype word bi-grams	$N_{50bi_{ns}}/N$	✓	
Ratio of number of words not found in top 50 non-stereotype word bi-grams	$1 - (N_{50bi_{ns}}/N)$	✓	
TF-IDF of query based on all user's vocabulary	$TFIDF_{all}$	✓	
TF-IDF of query based on stereotype's vocabulary	$TFIDF_s$	✓	
TF-IDF of query based on non-stereotype's vocabulary	$TFIDF_{ns}$	✓	
Number of stop words	N_{stop}		✓
Contains www	$www \in Q$		✓
Contains .com	$com \in Q$		✓
Contains .net	$net \in Q$		✓
Contains .org	$org \in Q$		✓
Contains .edu	$edu \in Q$		✓
Contains .gov	$gov \in Q$		✓
Contains http	$http \in Q$		✓
Contains AND	$AND \in Q$		✓
Contains OR	$OR \in Q$		✓
Contains ""	$"" \in Q$		✓
Contains interrogative word	$inter \in Q$		✓

Table 3.2: Vocabulary features, with *inter* being the following list of interrogative words: "who", "what", "when", "where", "why", "how", "is", "are", "can", "could", "should", and "would", and all variables defined by being contained with in the query as boolean values.

Feature	Formula	P3	DC1
Number of spelling errors	N_s		✓
Number of spelling errors found in KidSpell dataset	$\sum_{i=1}^{N_s} N_{s_i} \epsilon KidSpellTypos$		✓
Number of off by one spelling errors	$\sum_{i=1}^{N_s} \sum_{j=1}^{N_c N_{s_i}} N_{c_j} \epsilon LD(Ns_i, N_{c_j}) = 1)$		✓
Contains upper-cased word	$U_{word} \in Q$		✓
Contains punctuation	Q_{punct}		✓

Table 3.3: Spelling and punctuation features, where N_c represents a list of suggested corrections for a typo, LD stands for Levenhstein Distance (measured on a character level), and Q_{punct} is calculated by first ensuring that the query contains no search prefixes/suffixes (as defined in Table 3.2), then calculates a boolean value based on whether the query contains punctuation in the following list: .!?, .

We explore the total number of typos per query. We also compare each misspelled word found with a list of words compiled from the KidSpell dataset [20] which include typos made by children. Doing so allows us to recognize queries that contain typos that have been made by users who are similar to the stereotypical 8-12 year old searcher. Finally, we compare all misspelled words with a list of suggested correction, checking for any with a Levenhstein distance[4] of one (known as off by one typos). This kind of typo encompasses a wide variety mispelling errors made by children [20].

Furthermore, as the stereotypical 8-12 year old searcher are prone to treating queries as sentences. We suspect this includes errors that are typically corrected by query-suggestion, "did you mean" corrections, and auto-complete; such as the removal of punctuation and casting the query to lower case [34]. As such, we see these as displays of lacking search literacy. Therefore, we inspect queries for the sentence based punctuation as well as upper cased words. A detailed explanation on how to compute each of the spelling and punctuation features can be found in Table 3.3

3.1.4 Syntax

There are a number of ways in which natural language and keyword queries can differ. One of those ways is in their syntax. Keyword queries tend to be strings of nouns, while natural language queries more closely resemble sentences, containing articles, prepositional phrases; even adjectives and adverbs. In order to investigate this difference we calculate the ratio of number of part of speech to number of words in query for each part of speech tag shown in Table 3.4. However, there are several

[4]Defined as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other

hundred permutations of part of speech bi-grams, and several thousand for part of speech tri-grams. In order to determine which bi-grams and tri-grams to count in a query, we first calculate the top 10 occurring bi-grams found in queries generated by the stereotypical 8-12 year old searcher, as well as the top-5 tri-grams found in the same queries. We view all of these parts of speech features as being correlated to DC1, as we are using these features to elucidate the difference between natural language and keyword queries, not highlight a user's language development capabilities. A detailed explanation on how to compute each of the part of speech-related features can be found in Table 3.4

D-Level analysis is a process of establishing how developmentally complex the syntax in sentence is based on a sentence's syntax tree. This analysis is performed by first tagging the parts of speech for each word in the sentence, then using a probabilistic context-free grammar to parse those tags into syntax trees. The output of this parsing is used by the D-Level analyzer to determine the D-Level of a sentence. Each sentence is given a 0-7 rating, with 0 being syntactically simple, and 7 being a syntactically complex sentence [19]. Natural language queries will have higher D-Levels than keyword queries, as keyword queries are syntactically simpler than natural language queries or may even contain unparseable syntax trees, providing a clear distinction between these two kinds of queries. A detailed explanation on how these features are calculated can be found in Table 3.5

3.2 Session Based Features

Search sessions contain information beyond textual features found on a query by query basis. As these search interactions can provide insight into a user's search literacy

Feature Name	Formula	PoS Type	P3	DC1
Ratio of coordinating conjunctions	N_{cc}/N	Unigram		✓
Ratio of cardinal digits	N_{cd}/N	Unigram		✓
Ratio of determiners	N_{dt}/N	Unigram		✓
Ratio of existential theres	N_{ex}/N	Unigram		✓
Ratio of foreign words	N_{fw}/N	Unigram		✓
Ratio of preposition/subordinating conjunctions	N_{in}/N	Unigram		✓
Ratio of adjectives	N_{jj}/N	Unigram		✓
Ratio of compartive adjectives	N_{jjr}/N	Unigram		✓
Ratio of superlative adjectives	N_{jjs}/N	Unigram		✓
Ratio of list markers	N_{ls}/N	Unigram		✓
Ratio of modals	N_{md}/N	Unigram		✓
Ratio of nouns	N_{nn}/N	Unigram		✓
Ratio of plural nouns	N_{nnp}/N	Unigram		✓
Ratio of proper nouns	N_{nnps}/N	Unigram		✓
Ratio of plural proper nouns	N_{nns}/N	Unigram		✓
Ratio of predeterminers	N_{pdt}/N	Unigram		✓
Ratio of possessive endings	N_{pos}/N	Unigram		✓
Ratio of personal pronouns	N_{prp}/N	Unigram		✓
Ratio of possessive pronouns	N_{rb}/N	Unigram		✓
Ratio of adverbs	N_{rbr}/N	Unigram		✓
Ratio of comparative adverbs	N_{rbs}/N	Unigram		✓
Ratio of superlative adverbs	N_{rp}/N	Unigram		✓
Ratio of particles	N_{sym}/N	Unigram		✓
Ratio of to	N_{to}/N	Unigram		✓
Ratio of interjections	N_{uh}/N	Unigram		✓
Ratio of base Verbs	N_{vb}/N	Unigram		✓
Ratio of past tense verbs	N_{vbd}/N	Unigram		✓
Ratio of present particple verbs	N_{vbg}/N	Unigram		✓
Ratio of past participle verbs	N_{vbn}/N	Unigram		✓
Ratio of singular present verbs	N_{vbp}/N	Unigram		✓
Ratio of third person singular verbs	N_{vbz}/N	Unigram		✓
Ratio of determiners	N_{wdt}/N	Unigram		✓
Ratio of pronouns	N_{wp}/N	Unigram		✓
Ratio of wh-abverbs	N_{wrb}/N	Unigram		✓
Ratio of noun noun phrases	N_{nn_nn}/N	Bigram		✓
Ratio of comparative adjective noun phrases	N_{jj_nn}/N	Bigram		✓
Ratio of noun plural noun phrases	N_{nn_nns}/N	Bigram		✓
Ratio of to base verb phrases	N_{to_vb}/N	Bigram		✓
Ratio of comparative adjective plural noun phrases	N_{jj_nns}/N	Bigram		✓
Ratio of comparative adjective to phrases	N_{jj_to}/N	Bigram		✓
Ratio of noun preposition phrases	N_{nn_in}/N	Bigram		✓
Ratio of plural noun preposition phrases	N_{nns_in}/N	Bigram		✓
Ratio of preposition noun phrases	N_{in_nn}/N	Bigram		✓
Ratio of determiner noun phrases	N_{dt_nn}/N	Bigram		✓
Ratio of comparitive adjective noun noun phrases	$N_{jj\_nn\_nn}/N$	Trigram		✓
Ratio of noun noun noun phrases	$N_{nn\_nn\_nn}/N$	Trigram		✓
Ratio of comparitive adjective to verb Phrases	$N_{jj\_to\_vb}/N$	Trigram		✓
Ratio of noun noun plural noun phrases	$N_{nn\_nn\_nns}/N$	Trigram		✓
Ratio of to verb noun phrases	$N_{to\_vb\_nn}/N$	Trigram		✓

Table 3.4: Parts of speech features which are defined by the NLTK part of speech tagger.

Feature Name	Formula	P3	DC1
Number of D-Level 1 sentences in query	N_{dl1}		✓
Number of D-Level 2 sentences in query	N_{dl2}		✓
Number of D-Level 3 sentences in query	N_{dl3}		✓
Number of D-Level 4 sentences in query	N_{dl4}		✓
Number of D-Level 5 sentences in query	N_{dl5}		✓
Number of D-Level 6 sentences in query	N_{dl6}		✓
Number of D-Level 7 sentences in query	N_{dl7}		✓
Mean D-Level of query	$\frac{\sum_{i=1}^{7} N_(dl_i)}{N_{sent}}$		✓

Table 3.5: D-Level features, which are computed using the D-Level analyzer which can be found at [5].

[31], we draw inspiration from examples of expert identification that examine session information and recognize domain experts in fields like Medicine and Computer Science [69, 70] in order to shape our approach in recognizing expertise in the domain of search literacy. These features are calculated at a session level. However, we calculate Levenhstein distance for features at a query level, and for all sessions longer than this value is averaged over the session.

3.2.1 Query Based Interactions

There are several common characteristics of the stereotypical 8-12 year old searcher that separate their query based interactions from other users. Their lack of knowledge in how search engines work can cause them to repeat the same query multiple times in hopes for different results, or even press the search button repeatedly [9]. Furthermore, we know that search experts display refinement in their query generation, generally adding or removing a single word to clarify their results [70], while non experts may struggle with query reformulation. We also see inexperienced users can type slower faster and take longer to craft queries [60], whereas experienced users may spend less time perusing results, and even reformulate a query multiple times before clicking on a link. Informed by this knowledge we investigate the features found in

Table 3.6.

Feature Name	Formula	P3	DC1
Number of queries in a session	N_q		✓
Number of unique queries	N_{set_q}		✓
If all queries are the same query	$N_{set\_q} = 1$		✓
Number of repeat queries	N_{repQ}		✓
Ratio of queries to clicks	N_q/N_c		✓
Levenhstein distance between queries	$\frac{\sum_{i=1}^{N_q-1} LevenhsteinDistance(Q_i,Q_{i+1})}{N_q-1}$		✓
Time between queries	$\frac{\sum_{i=1}^{N_q-1} TimeStamp_{(Q_i+1)}-TimeStamp_{(Q_i)}}{N_q-1}$		✓

Table 3.6: Query based interaction features where Set_q is the set of all queries, Q_i is the i-th query in a session, and $TimeStamp_{Q_i}$ is the time stamp for i-th query. Levenhstein distance between queries is measured on a character level.

3.2.2 Click Based Interactions

Lack of search expertise can also manifest in how searchers interact with the search results retrieved by SE in response to their queries.the stereotypical 8-12 year old searcher are known to repeatedly click on the same result when the page does not come up instantly, tend to favor the clicking on the first result that shows up, and revisit web sites that they have already clicked on [9]. As such, we investigate the following click based features found in Table 3.7.

3.3 Classification

We simultaneously consider each of the features described in Section 3.1 and Section 3.2.1 when analyzing a search session. These features serve as evidence in determining whether or not a user is displaying skills expected from the stereotypical 8-12 year old searcher in their search behavior the stereotypical 8-12 year old searcher. We

Feature Name	Formula	P3	DC1
Number of clicks	N_c		✓
Number of unique clicks	N_{set_c}		✓
If all clicks are the same website	$N_{set\_c} = 1$		✓
Number of repeat clicks	N_{repC}		✓
Average click position	$\frac{\sum_{i=1}^{N_c} ClickPosition(C_i)}{N_c}$		✓
Time between clicks	$\frac{\sum_{i=1}^{N_c-1} TimeStamp_{(C_i+1)} - TimeStamp_{(C_i)}}{N_c-1}$		✓

Table 3.7: Click features where C_i is the ith click in the session, and $TimeStamp_{C_i}$ is the timestamp of the ith click in the session.

treat the task of recognizing this stereotype as a classification problem, and thus use these features as input to a Random Forest Classifier [45]. We have chosen this classifier for numerous reasons. The first reason is the way in which this classifier performs its feature selection. As seen in [48], the random forest classifier is noted for its potential ability to build trees that correlate to specific subsets of features. This feature sub-sampling is seen as a way to build specialized trees that recognize users based off these subsets of features. Furthermore, the spaced represented by our feature set is non-linear. Drawing a clear line down the middle of these numbers is a poor approach considering the potential variance of user skill within the stereotypical 8-12 year old searcher. As such, we have chosen a non-linear classifier.

CHAPTER 4

EXPERIMENTAL SETUP

In this chapter we describe the key components necessary for performing the experiments which allow us to determine the efficacy of RYSe.

4.1 Datasets

Due to concerns regarding the privacy and protection of children [26], there are no publicly available search sessions generated by the stereotypical 8-12 year old searcher. With this in mind, we take advantage of existing data sources (details of which are summarized in Table 4.1) in order to build datasets that enable the development and assessment of RYSe: Sessions With Clicks (as described in Section 4.1.1) and Single Query Sessions (as described in Section 4.1.2).

4.1.1 The Sessions With Clicks dataset

We use two data sources in order to build Sessions With Clicks. The first is the TREC Session Track query logs from 2011-2014 [2], which contain search sessions generated by adults. The second is the AOL query log [36], which has not been labeled, i.e., search sessions belonging to different user segments have not been identified. In order to label this data source so that in turn we can set the ground truth required for recognizing the stereotypical 8-12 year old searcher, we rely on several rules. We

Name	Source	Description
AOL Query Logs	[53]	36 million queries. Each entry contains a time stamp, user id, query, URL clicked, and position of URL clicked.
Queries From "Looking for the Movie Seven or Sven from the Movie Frozen? A Multi-perspective Strategy for Recommending Queries for Children"	[49]	602 queries individual queries. No session data. 301 queries generated by users between the grades of K to 9 (ages 6-13), 301 generated by adults.
TREC 2011-2014 Session Track	[2]	Approximately 1800 search sessions of predefined lengths with varying amounts of session data generated by adults.

Table 4.1: Data sources used to create datasets.

start by identifying user search sessions with a threshold of up to an hour [33]. We then examine every session that contains a click to a web site designated as "for users between the ages of 8-12" [23][1]. If a session exclusively retrieves websites designed for 8-12 year old's, then it is labeled as a session generated by the stereotypical 8-12 year old searcher. If a session does not exclusively retrieve sites designed for these users, we then consider that session's duration. The average session duration for children is approximately 3.75 minutes, while adults' sessions are typically 8.35 minutes [24]. We consider a search session containing a click on a website designed for children and a session duration closer to 3.75 minutes rather than 8.35 minutes as indicative of a search session generated by the stereotypical 8-12 year old searcher.

To build Sessions With Clicks, we use all of the TREC sessions with at least one click, and then sample the labeled AOL query logs maintaining an 80/20 (non-stereotype/stereotype) ratio, as this matches the distribution of SE users [1]. Fur-

[1]This designation is determined using a list of websites classified by their content.

	Stereotype Sessions With Clicks	Non-Stereotype Sessions With Clicks	Stereotype Single Query Sessions	Non-Stereotype Single Query Sessions
# Unique Sessions	7,980	31,920	301	1,204
Avg # Queries Per Session	1.87	2.95	1	1
Avg # Clicks Per Session	2.31	2.86	-	-
Avg # Words Per query	2.36	2.33	-	-
Avg Duration of Session (Minutes)	.95	12.15	-	-
% Unique Queries	41.1%	64.8%	100%	100%

Table 4.2: Description of the datasets used in our experiments.

thermore, we only consider AOL query log sessions that contain at least one-click, as sessions belonging to the stereotypical 8-12 year old searcher are labeled by this behavior and not doing so would skew the data. Given that we use several rules for session labeling, as well as 2 different data sources, the resulting Sessions With Clicks captures a variety of search behaviour indicative of our stereotype, while also drawing on sessions we can be certain are *not* generated by our target group.

4.1.2 Single Query Sessions

We also create a dataset of single queries with no clicks which are intended to emulate the start of a search session. To build this dataset, we first extract all queries generated by users between the grades of K-8 (ages 6-13) from the training data found in the "Looking for the Movie Seven or Sven from the Movie Frozen? A Multi-perspective Strategy for Recommending Queries for Children" data source (as described in Table 4.1), labeling these users as the stereotypical 8-12 year old searcher as these are queries formulated by the stereotypical of 8-12 year old. We then sample single queries drawn from the TREC Session Track query logs from 2011-2014 [2], as we can be certain that these queries are generated by users who are not our stereotype. Note that in building Single Query Sessions, we maintain the 80/20 (non-stereotype/stereotype) ratio mentioned in Section 4.1.1, so as to accurately represent typical SE use distributions. This dataset enable us to assess how well RYSe can recognize our stereotype at the start of a session.

4.2 Baselines

In order to contextualize the effectiveness of our strategy, we compare and contrast the performance of RYSe with that of suitable baselines. As there currently does not exist an established method for identifying the stereotypical 8-12 year old searcher based on their SE interactions, we must adapt user identification methods that have been used to recognize or label users similar to the stereotypical 8-12 year old searcher in different domains, such as chat-rooms and social media websites. We describe these baselines below.

Majority Classifier We start with the Majority Classifier, a naive baseline that

labels every session as being to the majority class of the user base. This strategy is meant to emulate the paradigm of major SE which assume users without a user profile, or not logged in, to be a traditional user.

Rule Based Classifier This baseline establishes a set of rules that are used to discern one user type from another. We adapt the user identification method found in [23], which labels sessions as belonging to different user types based on the websites they click, into a Rule Based classifier. We do so by identifying all sessions with a duration less than half a hour, and tagging all sessions that generate a click on website designated as for users between the ages of 8-12 as belonging to our user type. This classifier serves as a suitable baseline because it employs a strategy that is a step up from the Majority Classifier, but does not leverage more sophisticated methods of identification such as machine learning.

Text Based Classifier As an example of machine learning classification, we consider the user identification method proposed by Tam and Martell [67], that identifies the age of chat room users based on chat logs. Users are categorized as being either teens (13-19), in their 20s (20-29), 30s (30-39), 40s (40-49), 50s (50-59), or grouped together as Adults (20-59). Tam and Martell [67] determine that an Support Vector Machine (SVM) trained on tri-grams grams is the most effective classifier in recognizing teens. We therefore adapt this user identification method by utilizing a tri-gram bag of words SVM model trained on the text of each session's queries concatenated as input. This classifier serves as an suitable baseline as it represents one of the first attempts to identify online user's age based on their generated text using machine learning.

Multi-Feature Classifier This classifier leverages more than just word and charac-

ter tokens, as it also uses parts of speech and content features to recognize users by age. One user identification method which employs this type of classifier is presented by Santosh et al. [62], who recognize users in the following age ranges based on blogs written in English and Spanish: (10s: 13-17, 20s: 23-27 and 30s: 33-47). By utilizing a decision tree of classifiers, SVMs trained on bag of word and parts of speech n-grams as well logistic regression classifiers trained on Latent Dirichlet allocation topic models; this user identification method is able to identify a user's age. We adapt this method by utilizing the same models trained on the same features, using all the queries in a session concatenated together as the input. We have chosen this as a baseline as it represents a step forward from leveraging only word and character n-grams.

Multi-Model Classifier This classifier differs from the Multi-Feature Classifier by shifting the focus from being feature-based to model-based. Instead of using a wide variety of features on similar models, this classifier instead uses a wide variety of models on similar features. Nemati [51] provides us with a user detection method that uses this kind of classifier, recognizing user's age (15-19, 20-24, 25+) and gender based on a selection of documents gathered from a variety of social media websites (Twitter, Facebook, and others). Nemati's ensemble model is composed of 4 unique classifiers trained on user generated text: Logistic Regression, Naive Bayes, Multilayer Perceptron Network, and Gradient Boosting. We adapt this approach by utilizing the same models on the same features, but instead of training on social media documents we use queries concatenated by session as the input text. While the ages Nemati [51] recognizes are slightly older than our user type, it is the only example we found

using neural networks to recognize users under the age of 18 from their online generated text and utilize it due to this distinction.

4.3 Metrics

To quantify the performance of our strategy, as well as compare and contrast with respect to the performance of other baseline approaches, we employ several metrics. The first is *Accuracy* (computed as seen in Equation 4.1), as this is the go-to metric for assessing performance of binary classifiers in general [57]. However, Accuracy can misrepresent results as "it provides an overoptimistic estimation of the classifier ability on the majority class" [17]. We therefore examine *True Positive Rate* (**TPR**) (as seen in Equation 4.2) and *True Negative Rate* (**TNR**) (as seen in Equation 4.3). TPR enables us to evaluate a strategy's ability to recognize the stereotypical 8-12 year old searcher interacting with SE, while TNR enables us to evaluate the opposite; how effective a strategy is at correctly recognizing users who are *not* our stereotype. Both are important, as SE would want to provide appropriate aid and resources to users who belong to our target group, while ensuring that users who do not belong to that segment are also provided with services that align to their skills. By utilizing TPR, TNR, and Accuracy, we are granted a multi-faceted understanding of the effectiveness RYSe, as well as our baselines.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

where TP (True Positive) is the number of users that are correctly identified as being the stereotypical 8-12 year old searcher, TN (True Negative) is the number of users that are correctly identified as not the stereotypical 8-12 year old searcher, FP (False

Positive) is the number of users who are identified as being the stereotypical 8-12 year old searcher, but in actuality are not, and FN (False Negative) is the number of users who are identified as not being the stereotypical 8-12 year old searcher, but in actuality are.

$$TPR = \frac{TP}{TP + FN} \tag{4.2}$$

where TP and FN are as defined in Equation 4.1.

$$TNR = \frac{TN}{TN + FP} \tag{4.3}$$

where TN and FP are as defined in Equation 4.1.

4.4 Experiment Preparation

In order to perform our experiments, we must first prepare our data for tuning, training, and testing.

4.4.1 Hyper Parameter Tuning

Before executing our experiments, we tune the hyper parameters of our model. This process first involves splitting our Sessions With Clicks dataset into a disjoint training/tuning/test subset of 65/15/20 ratio (as seen in Figure 4.1). We then perform a grid search using the training and tuning splits from this subset across the following parameters, with the tuning data discarded afterwards:

- Number of estimators
- Class Weight (none or balanced)

- Criterion (Gini impurity or information gain)
- Bootstrap samples (uses bootstrap samples or the entire data to build trees)

We select the set of hyper-parameters that yields the highest Accuracy and TPR on the tuning split. The results of our tuning is a Random Forest Classifier with 450 estimators, using an unbalanced weight class, information gain as our criterion, with no bootstrap samples.

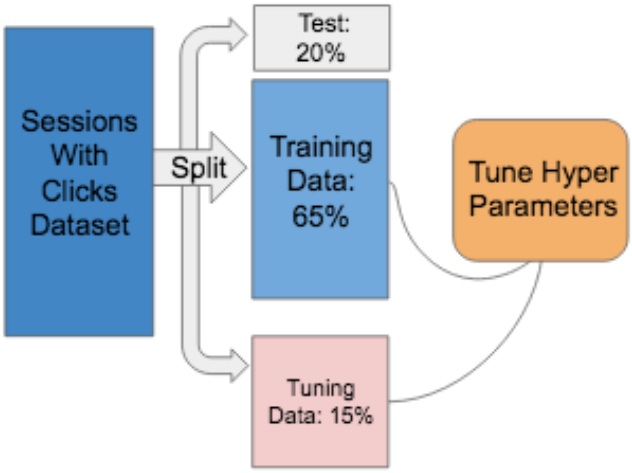


Figure 4.1: Figure highlighting how we create our disjointed training/tuning/testing subset for hyper parameter tuning from Sessions With Clicks.

4.4.2 Splitting Our Data For Testing And Training

All our experiments detailed in Chapter 5 will require us to train and test both RYSe as well as our baselines. The generation of our test and train splits requires a thorough description given the unique nature of our data. Sessions With Clicks is split 80/20

(training/testing), and Single Query Sessionsis split 20/80 (training/testing). RYSe and the Baselines then train on the training data from both datasets, and then test separately on both sets of testing data. The reason behind combining the training data is due to the nature of sessions that Single Query Sessions emulates. These sessions differ dramatically from those found in Sessions With Clicks, so RYSe and corresponding baselines (as seen in Section 4.2) require some training samples of Single Query Sessions in order to properly test on this dataset. This process we have just described can be seen further illuminated in Figure 4.2.

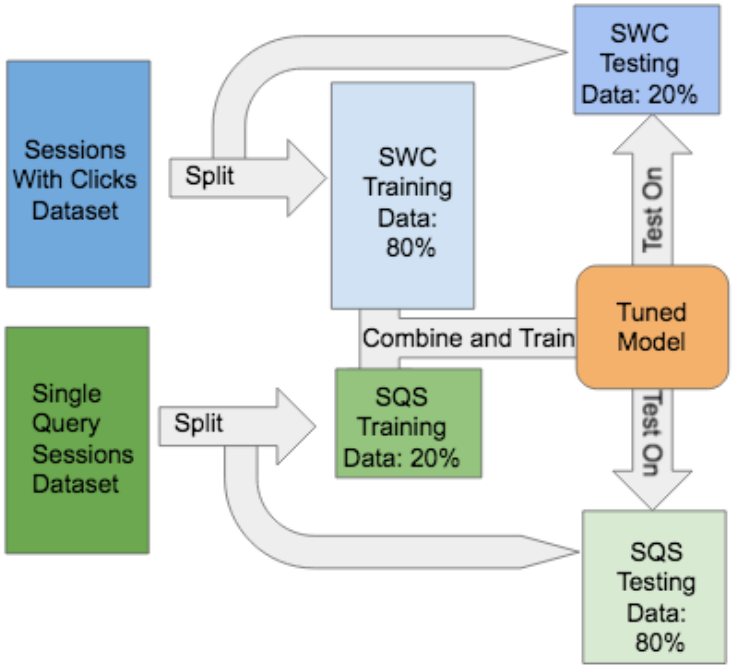


Figure 4.2: Figure highlighting how we create our training/testing data splits, containing information on how the training splits are combined and what portions of each dataset is tested on.

4.5 Validating Our Results

All of the experimental results reported in Chapter 5 are the result of 5-fold cross validation over the test and train splits of Sessions With Clicks and Single Query Sessions. Since we are reporting on dichotomous values as results rather than continuous values, the significance of these results is established using the McNemar test [42] using a p-value threshold of .05.

CHAPTER 5

RESULTS AND ANALYSIS

We conduct a series of experiments in order to determine how effective RYSe is at recognizing the stereotypical 8-12 year old searcher from their search behavior. Doing so allows us to establish the validity of our premise, as well as highlight the strengths and weaknesses of RYSe.

5.1 Feature Effectiveness

In this section, we describe the ablation study we conducted in order to showcase the validity of the design for RYSe. In particular, this study enables us to determine the effectiveness for each set of features when recognizing our stereotype and lets us understand the role that each feature set plays in this recognition. By analyzing these results we are then granted a multi-faceted view of our strategy's strengths, while also shown areas that we can improve upon. As described in Chapter 3, we group our features based on the type of data considered: text based features or session based features. These features are then described as indicative of skills related to either P3 or DC1 (with some features being indicators for both). We perform this ablation study on each feature set (textual features, session based features, features related to P3, and features related to DC1) on both Sessions With Clicks and Single Query Sessions separately. By conducting this ablation study on 2 different datasets, we can

portray multiple perspectives that further justify the applicability and effectiveness of our defined feature sets, as well as RYSe overall.

Feature Set	Accuracy	TNR	TPR
RYSe	0.947	0.988	0.783
Textual Features	0.942	0.985	0.768
Session Based Features*	0.788	0.940	0.178
Features Related to P3*	0.946	0.982	0.802
Features Related to DC1*	0.879	0.976	0.493

Table 5.1: Results from ablation study on Sessions With Clicks. * indicates statistical significance of a given feature set with respect to RYSe ($p \leq .05$).

When looking at the results of the ablation study performed on Sessions With Clicks (as seen in Table 5.1), we notice that textual features recognizes a majority of users who are, and are not, our stereotype. This feature set approximately recognizes 3 out of 4 of the users who belong to our target group, and 99 out of 100 users who do not. We then notice the lack of statistical significance between RYSe and the textual features which causes us to reason that textual features have a strong signal when it comes to recognizing our stereotype. As seen in Chapter 3, a majority of our features are textual, which may contribute to the strength of this signal. The observations made regarding our textual features feature also cause us to wonder if our target group can be recognized from solely textual features.

When examining session based features, we see a TPR that indicates about 1 in 5 users who are our stereotype are successfully recognized, which is observably less than the TPR of textual features. We also see that sessions based features recognizes 94% of the users who do not belong to our target group. The relatively high TNR of session based features demonstrates that users displaying stronger search literacy can be identified. These results also seem to indicate that while textual features

may have a stronger signal than that of session based features, session based features account for a number of users not recognized from text alone (based on comparing session based features and text based features to the results to RYSe). Session based features also leverage information which textual features can not, such as click data and the relationships between queries.

When we examine features related to P3, we see this feature set achieving a higher TPR than RYSe. This demonstrates that the stereotype we seek to recognize does display search behavior that allows us to identify a display of skills we expect them to possess based on the stages of development they are in, at least on this dataset. We do see a slight dip in TNR in features related to P3 when compared to RYSe. Solely utilizing features related to P3 to recognize our stereotype may be worth the trade off of misidentifying users who are not our stereotype. When looking at the results of features related to DC1, we see a TNR that is comparable to the other feature sets, and a TPR which demonstrates that every other user belonging to our target group can be recognized by their display of skills associated with DC1. As this TPR is lower than that of the features related to P3, we question what role features related to DC1 play in recognizing our stereotype. In order to more fully understand these feature set's roles effectiveness, it is important to consider the results from performing our ablation study on Single Query Sessions.

When looking at Table 5.2, we see features related to DC1 displaying the highest TPR of all feature sets. The lack of statistical significance between RYSe and the DC1 feature set highlights the strength of the signal this feature set provides. While the results of the related to P3 on Sessions With Clicks demonstrates that our target group can be recognized by their displays of skill in sessions with more data, the result of the DC1 feature set on Single Query Sessions demonstrates similar ideas.

Feature Set	Accuracy	TNR	TPR
RYSe	0.853	0.978	0.463
Textual Features*	0.817	0.991	0.156
Session Based Features*	0.800	1.000	0.000
Features Related to P3*	0.816	0.987	0.156
Features Related to DC1	0.868	0.978	0.464

Table 5.2: Results from ablation study on Single Query Sessions. * indicates statistical significance of a given feature set with respect to RYSe ($p \leq .05$).

We surmise that textual features, as well as features related to P3, play a strong role in performing classification on Sessions With Clicks due to the supposition that sessions with more than one query have a stronger textual signal. However, when only given one query, the textual features and features related to P3 seem to be a weaker signal. We also notice lack of statistical significance between features related to P3 and textual features when testing on Single Query Sessions, which demonstrates how strong a signal that features related to P3 have with regards to textual features in this context. Furthermore, in this context, the textual features and features related to P3 are outperformed by features relating to DC1. As a side note, the session based feature set performs as we expected, turning into a majority classifier due to the lack of session information for users in Single Query Sessions.

The varied performance of different feature sets on the two datasets shows all are necessary, as they each recognize a users' behavior from complementary perspectives. Features relating to DC1 aid in early session detection, where-as P3 features aid in detecting users further into their sessions. However, while the session based feature set has a low TPR we can surmise from overall results that combining this feature set with the textual feature set yields a higher overall result.

Referring to Table 5.1, we see that RYSe achieves an Accuracy of approximately

95%, successfully able to recognize approximately 80% of the users who belong to our target group and approximately 99% of the users who do not. This shows that a majority of the stereotypical 8-12 year old searcher from our Sessions With Clicks can be recognized by the display of skills we look for in their search behavior. Furthermore, with RYSe's TNR being approximately 99%, we see that users who do not belong to our target group can be reliably detected by features that indicate skills. However, these same that users that do not belong to our target group are known to employ natural language queries and may use language similar to our target group (simple words), which may account for our slightly lower TPR (with our model erring on the side of TN). We also investigate the results of RYSe on Single Query Sessions (Table 5.4), a dataset meant to emulate the start of a search session. We see RYSe with an Accuracy of 87.5%, recognizing 99% of users who are not our stereotype, and 40% that are. This TPR show that even with one query and no session information, RYSe is still capable of recognizing 2 out of 5 users who are the stereotypical 8-12 year old searcher. The TNR of testing on Single Query Sessions lines up with our results of testing Sessions With Clicks; that it may be possible to recognize users who do not belong to our target group from a single query. To gain a deeper understanding of what these results may mean, and verify our observations, we need to place them in the proper context.

5.2 Comparison to Baselines

In order to contextualize RYSe's overall performance, we compare RYSe alongside the baselines adapted to perform in the domain of recognizing our stereotype. This experiment allow us to not only determine where RYSe excels, but ascertain areas

for potential improvement. For this experiment, we train all of our adapted baselines on the concatenated queries of the sessions found in our composite training data (as detailed in Section 4.4.2), and then test separately on Sessions With Clicks and Single Query Sessions. Furthermore, we also perform McNemar tests on the results of our baselines, determining the statistical significance of their results not only in regards to RYSe but also to each other. Unless stated otherwise, reported results are statistically significant ($p \leq .05$).

When we first look at Table 5.3, we notice that RYSe significantly outperforms all our baselines in terms of Accuracy. The Majority classifier performs as expected, recognizing the majority perfectly while failing to recognize any users who are our stereotype. The Rule-Based classifier achieves an Accuracy of approximately 90%, with a TNR of 100%, and a TPR of 52%. This baseline recognizes almost as many users that are not our stereotype as the Majority classifier, while also being able to recognize users who are belong to our target group. However, when we refer to Table 5.4, the limitations of the Rule-Based classifier become apparent as this baseline recognizes the exact same users as the Majority classifier. This failure of the Rule-Based classifier to recognize any the stereotypical 8-12 year old searcher from Single Query Sessions highlights the shortcomings of relying on this method of classification to identify our stereotype. The Text-Based classifier is more effective than the Majority classifier, yet less effective than the Rule-Based classifier, on Sessions With Clicks. However the Text-Based classifier outperforms both of them on Single Query Sessions. The results of the Text-Based classifier on both these datasets demonstrates that our stereotype may be recognizable from just their text, albeit significantly less users are recognized when comparing this classifier to RYSe. However, there are two more classifiers that leverage text to identify the stereotypical 8-12 year old searcher,

the Multi-Feature classifier as well as the Multi-Model classifier. The Multi-Feature classifier has a higher TPR when compared to the Multi-Model classifier, but the Multi-Model classifier has a higher TNR. This trend continues when examining both these models effectiveness when testing on Single Query Sessions. The Multi-Feature classifier has a higher TPR while the Multi-Model classifier has a higher TNR. The performance from both of these models demonstrates that there are trade-offs when recognizing users who are, or are not, our stereotype from their text. The Multi-Feature classifier trades a higher TPR for a lower TNR, while the Multi-Model classifier does the reverse. RYSe however, has a TPR comparable to the Multi-Feature classifier, and a TNR comparable to the Multi-Model classifier. Whether textual recognition is performed using features such as content or parts of speech tags (like Multi-Feature classifier) or from an aggregate of models trained using bag of words and tf-idf features (like the Multi-Model classifier), neither approach can compare to RYSe's overall performance detecting users based on features designed to recognize displays of skill.

Feature Set	**Accuracy**	**TNR**	**TPR**
RYSe	0.948	0.988	0.783
Majority Classifier*	0.801	1.000	0.000
Rule-Based Classifier*	0.904	0.999	0.523
Text-Based Classifier*	0.856	0.995	0.299
Multi-Feature Classifier*	0.915	0.943	0.807
Multi-Model Classifier*	0.935	0.990	0.712

Table 5.3: Results of our performance evaluation when comparing RYSe to the baselines on Sessions With Clicks. * indicates statistical significance of a given baseline with respect to RYSe ($p \leq .05$).

As mentioned in Section 4.3, TPR reflects our primary goal of recognizing the stereotypical 8-12 year old searcher. We therefore consider the Multi-Feature classifier

Feature Set	Accuracy	TNR	TPR
RYSe	0.875	0.978	0.463
Majority Classifier*	0.800	1.000	0.000
Rule-Based Classifier*	0.800	1.000	0.000
Text-Based Classifier*	0.808	1.000	0.041
Multi-Feature Classifier*	0.833	0.987	0.221
Multi-Model Classifier*	0.834	0.996	0.179

Table 5.4: Results of our performance evaluation when comparing RYSe to the baselines on Single Query Sessions. * indicates statistical significance of a given baseline with respect to RYSe ($p \leq .05$).

to be our strongest baseline, as it has the highest TPR among all baselines on both datasets. However, when we directly compare the Multi-Feature classifier to RYSe we observe that RYSe has a higher TNR than the Multi-Feature classifier when testing on Sessions With Clicks. We acknowledge that a trade-off of a lower TNR for a higher TPR would be worthwhile, as long as a classifier's results remains effective across all datasets. When looking the Multi-Feature classifier's TPR on Single Query Sessions, while significantly stronger than all the baselines, it is half that of RYSe's. Furthermore, the Multi-Feature classifier's higher TPR on Sessions With Clicks comes with the trade-off of incorrectly recognizing approximately 5% of our non-stereotype users. This trade-off could have unforeseen consequences when applied in a real world setting, such as requiring users who are not our stereotype to confirm that in fact they are not the stereotypical 8-12 year old searcher and therefore once again rely on direct feedback. The comparable TPR and lower TNR of the Multi-Feature classifier on Sessions With Clicks, alongside the significantly lower TPR Multi-Feature classifier has on Single Query Sessions emphasizes how RYSe, overall; outperforms the Multi-Feature classifier.

5.3 Impact of Session Length on Effectiveness

As RYSe is intended to automatically recognize users from their search behavior, we emulate the context of automatically detecting a user during a search session. This accomplished by examining the results of RYSe based on session length as defined by number of queries. We investigate how RYSe and the Multi-Feature classifier (chosen due to its distinction in being the strongest baseline) perform in identifying our stereotype by examining the results of testing on Sessions With Clicks segmented by session length (number of queries) as well as the results of testing on Single Query Sessions. Doing so enables us to understand how much information we require from a user in order to successfully recognize them and provides us with an in-depth look at the performance of RYSe compared to the Multi-Feature classifier.

When looking at Table 5.5, we see that RYSe is better able to recognize our stereotype from shorter sessions. When looking at the RYSe's TNR, we see the inverse; the longer the session the more likely RYSe is able to recognize the user generating that session as not belonging to our stereotype. Even though the TPR steadily decreases, and the TNR steadily increases, when looking at the overall results we see these trends balanced out. This is likely due to the fact that approximately two thirds of sessions we test on have a session length of two or less queries, which means the results from those sessions carry more weight overall. Furthermore, we also consider Table 5.4, which as mentioned before, shows RYSe is capable of recognizing 4 out of every 10 users who are our stereotype. In order to contextualize these results, we will compare them to the top-performing baseline.

When examining the results found in Table 5.4 and Table 5.5, we observe that RYSe has a higher TPR than the Multi-Feature classifier on sessions with 1 query

(without and without clicks), comparable TPR on sessions with 2-3 queries with clicks, and a significantly lower TPR on any sessions longer than that. However, when looking at the confusion matrix behind the metrics, this translates to RYSe recognizing approximately 30 less users (out of 1500) who belong to our target group than the Multi-Feature classifier, making their results comparable with their differences being emphasized by the normalization of TPR. We also see that RYSe's TNR is significantly higher than Multi-Feature classifier's on sessions of any length, which hearkens back to the discussion of friction being added to the identification process mentioned in the previous subsection. The upward trend of TNR and downward trend of TPR on RYSe's results also causes us to wonder what the impact of aggregating features over a session has on recognizing the stereotypical 8-12 year old searcher. It is possible that users who are not the stereotypical 8-12 year old searcher may have relatively uniform search behaviour, especially in longer search sessions, while users who are our stereotype may have less uniform behaviour that is harder to capture when aggregated over longer sessions. Examining these sessions query by query rather, than aggregating over sessions, could yield stronger results. The methodology of considering query by query features may be a contributing factor in enabling the Multi-Feature baseline to maintain a TPR over 70%. Furthermore, the Multi-Feature baseline highlights how our stereotype can be recognized from their text, as this classifier's results are comparable to RYSe when looking at sessions with only one query and clicks (the most common length of a session in Sessions With Clicks as well as online [23]).

When looking at Table 5.4 however, RYSe recognizes approximately 46% of users who are our stereotype, compared to the the Multi-Model Classifier which recognizes approximately 22% of these users. These differing results of the Multi-feature classifier

demonstrate that textual classification is not the most effective choice when attempting to recognize our target users at the start of their session. It bears to reason that the earlier in a session our stereotype is detected, the faster the right resources and aid can be provided by the platform they are using. This reasoning sets RYSe apart from Multi-Model classifier, as it significantly outperforms this baseline on Single Query Sessions, a dataset intended to emulate new sessions with one query and no clicks.

RYSe				***Multi-Feature Classifier***			
# Queries	**Acc**	**TNR**	**TPR**	**# Queries**	**Acc.**	**TNR**	**TPR**
1	0.946	0.982	0.844	**1**	0.936	0.979	0.817
2	0.935	0.986	0.764	**2***	0.923	0.962	0.794
3	0.946	0.992	0.731	**3***	0.914	0.937	0.805
4	0.941	0.994	0.613	**4***	0.901	0.929	0.727
5	0.954	0.996	0.616	**5***	0.902	0.913	0.811
6+	0.973	0.997	0.464	**6+***	0.860	0.867	0.716
Overall	0.948	0.988	0.783	**Overall***	0.915	0.943	0.807

Table 5.5: Performance evaluation of RYSe (results on the left) compared to Multi-Feature Classifier (results on the right) on sessions of varying length. * indicates statistical significance ($p \leq .05$) in regards to RYSe results on sessions of the same length.

5.4 Discussion

In this chapter, we have presented a number of experiments conducted to determine the efficacy of RYSe detecting the stereotypical 8-12 year old searcher from displays of skill inferred from their search behaviour. These experiments were performed over two different datasets; one that simulates a variety of user styles and interactions by primarily using rule-based labeling on an unlabeled data source (Sessions With Clicks), and the other which draws on queries generated by users who are known

to belong to our target group (Single Query Sessions). Given the nature of these datasets, it is important to draw distinctions in performance based on the outcomes of each experiment performed on them. Experiments performed in Section 5.1 on Sessions With Clicks show that our stereotype can be recognized from displays of skill indicative of users in P3, while experiments on Single Query Sessions shows that our stereotype can be recognized from displays of skill indicative of users in DC1. These differing datasets also highlight the importance of considering features that point to both stages of development, showing that they are crucial in recognizing the stereotypical 8-12 year old searcher throughout various parts of these user's sessions.

Even though RYSe is comparable to the strongest baseline when recognizing our stereotype on simulated data, RYSe does significantly outperforms that baseline when detecting the stereotypical 8-12 year old searcher from a single query. With a TPR of 46% when testing on Single Query Sessions, compared to that of the Multi-Feature classifier at 22%, demonstrates that RYSe can recognize our stereotype from displays of skill found in a single query twice as effectively as the nearest baseline on this dataset. The importance of RYSe's efficacy when performing recognition on this dataset is further amplified by the premise that detecting our stereotype early in their session is important. This importance is two-fold. First, the earlier a user who may require specific tools and resources can be recognized in their session, the earlier a SE can respond to those needs. Second, if a user who needs these resources is only detected after a few queries, there is no guarantee they will receive assistance, and these users may terminate their search session due to frustration or confusion before this recognition can occur.

Despite RYSe being more effective overall than the baselines, it is important to highlight what can be learned from these adapted baselines. Search sessions generated

by the stereotypical 8-12 year old searcher can provide enough text to recognize our stereotype, whether through bag of word models or style and content based features. As we mentioned previously, the Multi-Feature classifiers shows particular promise when recognizing the stereotypical 8-12 year old searcher. What is notable about this classifier is its use of topic-modeling, which leads us to wonder how this sort of feature can be used to recognize indicators of skills displayed by our stereotype. Another difference is how this baseline considers features based on concatenated queries, rather than aggregating features over sessions. Both warrant further investigation.

Ultimately the experiments we have conducted show that our stereotype can be recognized by displays of skill that indicate what stages of development these users may be in. Furthermore, these users can be recognized early in their sessions, significantly outperforming all the baselines on queries that emulate the start of a search session (Single Query Sessions). While there is room for improvement, as a proof of concept, these experiments demonstrate that RYSe can recognize the stereotypical 8-12 year old searcher by displays of skill indicative of the stages of development we expect these searchers to be in and that this recognition can occur automatically.

CHAPTER 6

CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this thesis we introduce RYSe, a strategy designed to automatically detect the stereotypical 8-12 year old searcher from their search behavior. RYSe does so by examining text based features and session based features, as both are indicators of skills that we expect the stereotypical 8-12 year old searcher to display based on the stages of development that they are likely to be in. These features are then utilized by a random forest classifier to recognize our stereotype. When looking at the performance of RYSe, we see that our stereotype can in fact be identified by displays of skills expected from someone in P3 and DC1. This recognition can occur from a session in progress (as demonstrated by experiments performed on Sessions With Clicks) or even at the start of a session (as demonstrated by experiments performed on Single Query Sessions). Furthermore, the results of experiments performed on these two datasets imply that RYSe can provide SE with a strategy that recognizes a user's on based on displays skill (rather than relying on age or grade), automatically.

While designing and developing RYSe, we did encounter some limitations. Though our datasets are useful for performing experiments that enable us to establish the efficacy of RYSe, due to the nature of labeling Sessions With Clicks, there are some features we could not use when performing this recognition. Since session duration is used for labeling, we are unable to use it when delineating our stereotype from other

users. Furthermore, the AOL query log is already pre-processed, with punctuation stripped out of most queries and text cast to lower case. We still consider these features, but their sparsity in our largest dataset may affect the strength of the signal that punctuation and casing carry. Finally, the Sessions With Clicks dataset is limited to sessions with at least one click, meaning that only certain types of sessions are captured by our labeling method. We manage to mitigate these limitations by also using Single Query Sessions for assessment purposes, which contains the particular type of session missing from Sessions With Clicks. Furthermore, Single Query Sessions contains queries we are certain to be generated by users who are our stereotype. However, we also acknowledge that the Single Query Sessions dataset displays limitations. As described in Section 4.1.2, Single Query Sessions is comprised of query samples from users who belong to our target group, as well as users who are over the age of 18. It is worth noting, however, that SE interactions from an important segment of users are missing not only from Single Query Sessions but also Sessions With Clicks: the stereotypical 13-17 year old. We expect these users to have the skill set closest to our stereotype and as such could be the most challenging to distinguish from our target group. The fact that both datasets lack this segment of users may impact the results of testing on these datasets, as there currently exists this notable gap in expected displays of skill. Furthermore, both datasets lack granularity beyond recognizing users who either are, or are not, our stereotype which can be perceived as another limitation. While utilizing both datasets in our experiments helps us to somewhat mitigate the limitations that we would otherwise encounter when only using one or the other, we intend to further address these limitations through a series of users studies. The first user study we propose will be designed to generate a new dataset. We intend to capture user-SE interactions from a more

granulated range of users: the stereotypical 8-12 year old searcher, the stereotypical 13-17 year old searcher, and users who are neither of those stereotypes. This will address aforementioned limitations and at the same time open the door to future research in recognizing new stereotypes. Building this new dataset allow us to apply our own pre-processing, ensuring that the valuable information missing from, or used to label, Sessions With Clicks can be explored in future experiments. The second user study we intend to perform will be an online assessment of RYSe. This assessment will allow us to perform a more in-depth analysis and evaluation than the offline assessment presented in this work while also enabling us to analyze the efficacy of RYSe recognizing the stereotypical 8-12 year old searcher in a real time environment.

Successful recognition of the stereotypical 8-12 year old searcher also lays the foundation for exploring how SE can adjust their algorithmic support based on recognizing users by their potential skills. SE for children can use RYSe to provide resources and aid based on establishing the skill sets of their recognized users. This can entail curating results comprehensible to our target group, adjusting query suggestions and snippets based on a user's expected literacy, and even using spell checkers designed to recognize spelling errors common to the stereotypical 8-12 year old searcher. Commercial SE can also leverage RYSe to accomplish similar results, while also potentially screening inappropriate content to users recognized as our stereotype as well as enforcing COPPA standards. Finally, other researchers can use RYSe as a foundational tool to recognize a specific segment of searchers when researching and designing tools that address these user's struggles, as the ability to recognize our stereotype goes hand in hand with providing them with the aid that they need.

While the scope of RYSe is the stereotypical 8-12 year old searcher, children are

a wide group of individuals, with unique talents and abilities that extend past our stereotype and onto children that exhibit traits from different stages of development. Since RYSe demonstrates that our stereotype can be recognized by their displaying skills associated with the stages of development we expect these users to be in, it lays the foundation for recognizing other stereotypes that can also struggle with SE use. This includes teenagers (13-17), as mentioned earlier; users younger than our stereotype (8 and under), and users over the age of 65. All of these stereotypes embody stages of development that will have an impact on search behavior.

While performing the experiment detailed in Section 5.2, we realized that beyond aggregating feature values over sessions there also exists the potential for using query by query based input to recognize the stereotypical 8-12 year old searcher as well. While RYSe currently offers the groundwork for performing classification in order to recognize the stereotypical 8-12 year old searcher, future research directions involve considering strategies that can perform classification based on query by query input, such as time series classification. There are a number of models that consider click-through data in order to provide personalization which can potentially be adapted to recognize our stereotype. As personalization relies on recognizing identifying characteristics of users, there is the possibility that these strategies can provide a starting point for potentially expanding our model to utilize time-series data to recognize indicators of skill and improve RYSe's overall efficacy. Furthermore there are other features we could also consider that displayed promise when recognizing the stereotypical 8-12 year old searcher, such as topic-modeling, and ones that remain unexplored, like contextual features.

Even though it has been acknowledged that stages of development influence children's search behavior [22, 31], recognizing these users based on displaying skills be-

longing to these stages remained unexplored. With RYSe, we demonstrate that when categorized by a stereotype, young searchers can be recognized by displaying skills dictated by the stages of development we expect them to be in. This discovery opens the doors of research mentioned in the previous paragraphs such as: recognizing other stereotypes, designing tools that more directly focus on aiding the stereotypical 8-12 year old searcher, and has implications that SE for children can tailor their response to these users not based on age or grade but instead by automatically recognizing a young searcher's skill. Furthermore, RYSe stands as a proof of concept, harboring the distinct possibility of further improving this strategy's ability to recognize the stereotypical 8-12 year old searcher by considering time-series classification as well as content-based features.

REFERENCES

[1] Computer and Internet Access in the United States: 2012. https://www.census.gov/data/tables/2012/demo/computer-internet/computer-use-2012.html, 2012.

[2] Trec session track 2010-2014. https://trec.nist.gov/data/session.html, 2014. (accessed 2020-18-09).

[3] Twitter. http://www.twitter.com/, 2019.

[4] Time for kids. https://www.timeforkids.com/, 2020.

[5] D-level analyzer. http://www.personal.psu.edu/xxl13/downloads/d-level.html, 2021.

[6] Oghenemaro Anuyah, Ashlee Milton, Michael Green, and Maria Soledad Pera. An empirical analysis of search engines' response to web search queries associated with the classroom setting. *Aslib Journal of Information Management*, 72(1):88–111, 2019.

[7] Ion Madrazo Azpiazu, Nevena Dragovic, and Maria Soledad Pera. Finding, understanding and learning: Making information discovery tasks useful for children and teachers. In *Proceedings of the 2nd International Workshop on Search as Learning, Co-Located with the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, 2016.

[8] Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera, and Jerry Alan Fails. Online searching and learning: Yum and other search tools for children and teachers. *Information Retrieval Journal*, 20(5):524–545, 2017.

[9] Dania Bilal. Children's use of the yahooligans! web search engine. iii. cognitive and physical behaviors on fully self-generated search tasks. *Journal of the American Society for information science and technology*, 53(13):1170–1183, 2002.

[10] Dania Bilal and Li-Min Huang. Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing. *Aslib Journal of Information Management*, 71(2):241–259, 2019.

[11] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[12] Pamela J. Black, Melissa Wollis, Michael Woodworth, and Jeffrey T. Hancock. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse and Neglect*, 44:140–149, 2015.

[13] Elana Broch. Children's search engines from an information search process perspective. *School Library Media Research*, 3, 2000.

[14] Elana Broch. Children's search engines from an information search process perspective. *School Library Media Research*, 3, 2000.

[15] Stephanie Carretero, Riina Vuorikari, and Yves Punie. DigComp 2.1: The Digital Competence Framework for Citizens with eight proficiency levels and examples of use. May 2017.

[16] Òscar Celma. *The Recommendation Problem*, pages 15–41. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[17] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[18] J.D. Cooper and N.D. Kiger. *Literacy Assessment: Helping Teachers Plan Instruction.* Houghton Mifflin Company, 2001.

[19] Michael A Covington, Congzhou He, Cati Brown, Lorina Naci, and John Brown. How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale. 2006.

[20] Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Sole Pera, Katherine Wright, and Casey Kennington. Kidspell: A child-oriented, rule-based, phonetic spellchecker. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6937–6946, 2020.

[21] Nevena Dragovic, Ion Madrazo Azpiazu, and Maria Soledad Pera. "is sven seven?": A search intent module for children. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 885–888, New York, NY, USA, 2016. ACM.

[22] Huizhong Duan and Bo-June Hsu. Online spelling correction for query completion. In *Proceedings of the 20th international conference on World wide web*, pages 117–126, 2011.

[23] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. An analysis of queries intended to search information for children. In *Proceedings of the third symposium on Information interaction in context*, pages 235–244. ACM, 2010.

[24] Sergio Duarte Torres and Ingmar Weber. What and how children search on the web. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 393–402, 2011.

[25] Sergio Duarte Torres, Ingmar Weber, and Djoerd Hiemstra. Analysis of search and browsing behavior of young users on the web. *ACM Transactions on the Web (TWEB)*, 8(2):7, 2014.

[26] Michael D Ekstrand. Challenges in evaluating recommendations for children. In *International Workshop on Children & Recommender Systems*, 2017.

[27] Jerry Alan Fails, Maria Soledad Pera, Oghenemaro Anuyah, Casey Kennington, Katherine Landau Wright, and William Bigirimana. Query formulation assistance for kids: What is available, when to help —& what kids want. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 109–120, New York, NY, USA, 2019. Association for Computing Machinery.

[28] Elizabeth Foss, Allison Druin, Robin Brewer, Phillip Lo, Luis Sanchez, Evan Golub, and Hilary Hutchinson. Children's search roles at home: Implications for designers, researchers, educators, and parents. *Journal of the American Society for Information Science and Technology*, 63(3):558–573, 2012.

[29] Daniel Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822–1843, 2009.

[30] Tatiana Gossen. *Search engines for children: search user interfaces and information-seeking behaviour*. Springer, 2016.

[31] Tatiana Gossen, Julia Hempel, and Andreas Nürnberger. Find it if you can: usability case study of search engines for young users. *Personal and Ubiquitous Computing*, 17(8):1593–1603, 2013.

[32] Tatiana Gossen, Michael Kotzyba, and Andreas Nürnberger. Knowledge journey exhibit: Towards age-adaptive search user interfaces. In *Advances in Information Retrieval: 37th European Conference on IR Research*, pages 781–784, 03 2015.

[33] Aaron Halfaker, Os Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. User session identification based on strong regularities in inter-activity time. In *Proceedings of the 24th International Conference on World Wide Web*, pages 410–418, 2015.

[34] Hyejung Han. Childrens help-seeking behaviors and effects of domain knowledge in using google and kids.gov: Query formulation and results evaluation stages. *Library & Information Science Research*, 40(3-4):208–218, 2018.

[35] Arturo E Hernandez and Ping Li. Age of acquisition: its neural and computational mechanisms. *Psychological bulletin*, 133(4):638, 2007.

[36] Jeff Huang. Aol query logs. `https://jeffhuang.com/search_query_logs.html`, 2006.

[37] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, volume 30, 2012.

[38] Common Core State Standards Initiative. Common core state standards initiative. `http://www.corestandards.org/`, 2021.

[39] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. "I know what you did last summer" - Query logs and user privacy. In *International Conference on Information and Knowledge Management, Proceedings*, pages 909–913, 2007.

[40] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 195–204, New York, NY, USA, 2013. ACM.

[41] Barry M Kroll. Developmental relationships between speaking and writing. *Exploring speaking-writing relationships: Connections and contrasts*, pages 32–54, 1981.

[42] Peter A Lachenbruch. Mcnemar test. *Wiley StatsRef: Statistics Reference Online*, 2014.

[43] Cheryl Laz. Act your age. In *Sociological forum*, volume 13, pages 85–113. Springer, 1998.

[44] Geoffrey Leech, Paul Rayson, et al. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge, 2014.

[45] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[46] Jane Lin. Automatic Author Profiling of Online Chat Logs. Master's thesis, Naval Postgraduate School, Monterey, Califoria, 2007.

[47] Xiaofei Lu. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208, 2012.

[48] Ion Madrazo. Towards multipurpose readability assessment. Master's thesis, Boise State University, Boise, Idaho, 2016.

[49] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, page 92–101, New York, NY, USA, 2018. ACM.

[50] James Marquardt, Sergio Davalos, Tacoma Washington, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, Martine De Cock, and Tacoma Washington. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, pages 1129–1136, 2014.

[51] Ali Nemati. Gender and age prediction multilingual author profiles based on comments. In *FIRE (Working Notes)*, pages 232–239, 2018.

[52] Felicity O'Dell, John Read, Michael McCarthy, et al. *Assessing vocabulary*. Cambridge university press, 2000.

[53] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, pages 1–es, 2006.

[54] Nick Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *ICSC 2007 International Conference on Semantic Computing*, pages 235–241, 2007.

[55] Avar Pentel. Predicting age and gender by keystroke dynamics and mouse patterns. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 381–385. ACM, 2017.

[56] J. Piaget. *The Origins of Intelligence in Children*. The Norton library, N202. Norton, 1963.

[57] Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: an empirical study of common baselines. *Data mining and knowledge discovery*, 31(2):287–313, 2017.

[58] Pablo E. Román, Robert F. Dell, Juan D. Velásquez, and Pablo S. Loyola. Identifying user sessions from web server logs with integer programming. *Intelligent Data Analysis*, 18(1):43–61, 2014.

[59] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski, and Carol Tenopir. The google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, pages 290–310. Emerald Group Publishing Limited, 2008.

[60] Sophie Rutter, Nigel Ford, and Paul Clough. How do children reformulate their search queries?. *Information Research: An International Electronic Journal*, 20(1):n1, 2015.

[61] Sara Salehi, Jia Tina Du, and Helen Ashman. Examining personalization in academic web search. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, pages 103–111. ACM, 2015.

[62] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author profiling: Predicting age and gender from blogs. *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, 2013.

[63] Katherine Schaeffer. Discover research tools for studying texts. https://www.pewresearch.org/fact-tank/2019/08/23/most-u-s-teens-who-use-cellphones-do-it-to-pass-time-connect-with-others-learn-new-things/, 2019. (accessed 2019-09-09).

[64] Scholastic. Welcoming the Internet Into Your Classroom. available at: `https://www.scholastic.com/teachers/articles/teaching-content/welcoming-internet-your-classroom/`, 2018. (accessed March 11, 2018).

[65] Martin E. P. Seligman, David Stillwell, Lukasz Dziurzynski, H. Andrew Schwartz, Megha Agrawal, Achal Shah, Lyle H. Ungar, Stephanie M. Ramones, Margaret L. Kern, Michal Kosinski, and Johannes C. Eichstaedt. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *Public Library of Science ONE*, 8(9):e73791, 2013.

[66] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *Public Library of Science ONE*, 10(3):1–20, 2015.

[67] Jenny Tam and Craig H. Martell. Age detection in chat. *ICSC 2009 - 2009 IEEE International Conference on Semantic Computing*, pages 33–39, 2009.

[68] Edson R. D. Weren, Viviane Pereira Moreira, and José Palazzo Moreira de Oliveira. Using simple content features for the author profiling task notebook for pan at clef 2013. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, 2013.

[69] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 132–141, New York, NY, USA, 2009. ACM.

[70] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting user knowledge gain in informational search sessions. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pages 75–84. ACM, 6 2018.

[71] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1225–1226, 2011.

[72] Xiangmin Zhang, Jingjing Liu, Michael Cole, and Nicholas Belkin. Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology*, 66(5):980–1000, 2015.