UNIVERSITY OF CALIFORNIA
RIVERSIDE

Analysis and Querying of Health-Related Social Media

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Shouq Ahmed Sadah

June 2017

Dissertation Committee:

Dr. Vagelis Hristidis, Chairperson
Dr. Michalis Faloutsos
Dr. Tamar Shinar
Dr. Vassilis Tsotras

The Dissertation of Shouq Ahmed Sadah is approved:

_____

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgments

First, I would like to express my special appreciation and thanks to my advisor, Professor Vagelis Hristidis, for his support, motivation, patience, enthusiasm, and knowledge during the last five years. I appreciate all his ideas, advice, contributions, and willingness that allowed me to pursue research on topics for which I am truly passionate. Without his guidance and constant feedback, this Ph.D. would not be achievable.

My sincere thanks must also go to the members of my dissertation committee: Professor Michalis Faloutsos, Professor Tamar Shinar, and Professor Vassilis Tsotras for their time to offer me valuable comments toward improving my research.

I would also thank my lab-mates in the Database lab for all the support. To Moloud Shahbazi, who went with me through hard times, cheered me on, and celebrated our accomplishments together. To Matthew Wiley, Shewin Cheng, Nhat Le, and Mohiuddin Qader, for guiding and encouraging me during my research, and for the great moments we had together.

Finally, I would like to acknowledge with deep and sincere gratitude, the continuous support and unconditional love of my family. I am forever indebted to my parents, Ahmed and Khadijah, for giving me the opportunity and trust that have made who I am. I deeply thank my brother Mohammed, for his presence and being by my side throughout my Ph.D. I would also thank my siblings: Abdullah, Haifa, Alya, Rami, Raed, and Saeed, for their unfailing emotional support.

To my parents and family for all the love and support.

ABSTRACT OF THE DISSERTATION

Analysis and Querying of Health-Related Social Media

by

Shouq Ahmed Sadah

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, June 2017
Dr. Vagelis Hristidis, Chairperson

The increased popularity of social media and the copious amount of user-generated data in the last few years have impacted various aspects of individuals' lives. The use of social media for health care related purposes, which is the focus of this thesis, has increased exponentially.

This provides the researchers with a massive volume of data that can augment traditional health-related data sources (like electronic medical records) if properly mined and analyzed. Despite the advances in text analytics, it is challenging to analyze this data, due to its specialized vocabulary, the data collection, and the missing values.

In this thesis, we focus on two research directions: (a) Analyzing the demographics of users who participate in health-related social media, along with their posted content across a wide range of sources, and highlight specific health issues reported by users. (b) Effectively querying health-related social media or other health-related documents (can be generalized to the problem of querying annotated document).

Specifically, in our first contribution, we study the demographics of users who participate in health-related social media, to identify possible links to health care disparities. Using these demographics, our second contribution analyzes the content of posts grouped by demographic segments by implementing information extraction methods to extract medical concepts, identify top distinctive terms, and measure sentiment and emotion. We also extend our content analysis in the third contribution by studying the intent of posts generated by users for different data sources. Lastly, we focus on a specific domain, electronic cigarettes, and analyze the health-related effects reported by online users.

In the second direction of this thesis, we developed a query framework to help users efficiently explore health-related data, present in either online social media or other medical documents, by exploiting the relationships between the network users or the concepts inside the documents. Our solution is generalized to other domains with similar properties, such as general purpose social networks. We refer to this problem as keyword querying on graph-annotated documents, where we query documents annotated by interconnected entities, which are related to each other through association graphs. Our novel framework balances the importance of text relevance and semantic relevance through the graph.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The rising volume of internet activity over the past few years, with half of the world's population are online today [1], has affected nearly every aspect of our lives. Social media platforms with their nature of allowing people from various backgrounds to connect, collaborate, and engage in different ways, has annual growth of 21% since 2016, with billions of users who form 37% of the world's population [1]. Social media brings a new dimension to health care as it allows patients, health care professionals, and public to communicate and share health information with the opportunity of improving health outcomes. A national survey shows that 72% of adults have searched online for a range of health issues, while 26% of adults state they have watched or read other people's health experience, and 16% of adults have looked online to find others who share the same health concerns [2]. The massive user-generated data from health-related social media, if properly mined and analyzed, can be used efficiently by researchers, public and private heath care sectors to augment traditional health-related data sources (like electronic medical records), and improve quality of services and products. However, it is challenging to search and analyze this data, due to the informal writing style and specialized vocabulary used amongst health social media members, data collection and

the missing values, extracting health-related terms from social posts, and building domain-specific classifiers.

1.2 Research Problems

This thesis implemented two key approaches for advancing research in health care informatics, and analyzing the content of health-related social media to help users efficiently explore health-related data. Firstly, I analyzed three different types of health-related social media: general Web-based social networks, drug review websites, and health Web forums. I presented different aspects of the data analysis, including demographics, sentiment and emotions, top distinctive terms, top medical concepts, and users' intent. Secondly, I proposed an original query framework to answer keyword queries on graph-annotated documents. I next summarize the research of each chapter of this thesis.

A Study of the Demographics of Web-Based Health-Related Social Media Users

With regard to the first approach, I analyzed two different dimension of the data discovered on health-related social media, namely users' demographics and posts' content. I analyzed the users' demographics of health-related social media to identify possible links to health care disparities.

The challenge of health care disparities, where two population groups receive unequal services [3], has been monitored and analyzed across various dimensions of social determinants in health, including education and income, environmental hazards,

and health outcomes such as mortality, morbidity, and behavioral risk factors [4]. Nevertheless, despite the increasing volume of social media use during the contemporary period, the possible health care disparity has not been investigated in relation to social media engagement.

For this study, three different types of Web-based social media were utilized to obtain data: (1) general Web-based social networks, namely Google+ and Twitter; (2) drug review websites, and (3) health Web forums. We examined the following demographics attributes: gender, age, ethnicity, geographical location, and writing level. Because a number of the sources did not report all demographics attributes, we built and evaluated domain-specific classifiers to estimate the missing data when possible. Our findings revealed significant and unanticipated disparities of the various demographic groups' participation.

A complete analysis of the demographics data is presented in Chapter 2.

Demographic-Based Content Analysis of Web-Based Health-Related Social Media

Additionally, I analyzed Web-based health-related social media content in relation to the demographic data that had previously been collected and assessed, in order to identify popular topics discussed by certain demographic groups through different social media, which will assist with guiding research and educational activities.

Previous works have analyzed health-related social media and their content to evaluate their effectiveness for enhancing communication between patients and health care providers [5]–[8]. However, no previous research has investigated how various

demographic populations engage with health-related social media, thus we decided to assess this issue.

For this problem, we collected data from three types of health-related social media: (1) general Web-based social networks, namely Google+ and Twitter; (2) drug review websites, and (3) health Web forums, covering a total of 6 million users and 20 million posts. Our analysis considered five demographic attributes: gender, age, ethnicity, location, and writing level. For each demographic attribute, we analyzed the posts' contents across different dimensions: sentiment and emotion; top distinctive terms, and top medical concepts, including disorders and drugs. Our results can contribute to knowledge through various means, including guidance of educational initiatives, advertisement of associated products, assistance to funding agencies to better allocate resources, alongside an effective understanding of health disparities in health-related social media.

A complete analysis has been presented in Chapter 3.

Intent Classification of Health-Related Social Media

In this section, I analyzed the content of health-related social media in-depth, by classifying the intent of users in order to determine how they engage and share information across the different social media applications.

The use of health-related social media has increased in the last few year, with 72% of adults stating that they have searched for health condition information online [2]. Using health-related social media, patients seek online communities to interact with other

patients, as a means of sharing information and requesting or providing mutual support [9]–[11]. We analyzed two types of health-related social media: (1) general Web-based social networks, namely Google+ and Twitter, and (2) health Web forums. We randomly selected posts from each source, then manually identified and determined the intents based on the post content. For health Web forums, we identified four intents as follows: share experience, ask for advice, request/give support, and talking about family. For general Web-based social networks, we identified five additional intents: share news, jokes, ads, personal opinion, and educational materials. We labeled the posts according to each intent, and use supervised learning classifier to train the data if there were sufficient posts. The classifiers with greater accuracy were utilized to label the rest of our posts. We analyzed and categorized the content based on the associated demographic data when possible.

A complete explanation and analysis of this section is provided in Chapter 4.

Mining for Online Health-Related Effects Associated with Electronic Cigarettes

For this chapter, I analyzed the content associated with electronic cigarette- also known as e-cigarette- users, to better comprehend the symptoms and disorders associated with smoking e-cigarettes.

E-cigarettes have been designed to mitigate the health problems resulting from cigarette smoking. The ‘harm reduction’, as it is commonly advertised deliver the nicotine using battery-powered vaporization of a mixture of nicotine and propylene-

glycol solution [12]. Regardless, the health consequences of using e-cigarette products has not been studied in-depth.

For this research, we analyzed the data we collected from a reputable e-cigarette forum, and identified any stated health-related affects associated with smoking e-cigarettes. We analyzed the collected data further, by using a modified version of MetaMap tool[13], to extract references to medical concepts, alongside a measurement of the sentiments of all posts using a supervised learning classifier.

A comprehensive analysis of health-associated consequences of using e-cigarettes is presented in Chapter 5.

Querying Documents Annotated by Interconnected Entities

In this problem, I help users to retrieve the most relevant documents when they query a collection of documents annotated by interconnected entities.

Large number of applications have a collection of text documents that are annotated by entities, which are related to each other through association graphs. An example of such applications is PubMed documents (or Electronic Medical Records), where documents are annotated with a set of MeSH (Medical Subject Heading) concepts, and the associations between these concepts being defined by the MeSH ontology. A further example is social networks, where every post is annotated with the author's ID, with all the users connected through a friendship graph. For this research, we investigated the problem where a query specifies one or more graph entities, in addition to the keywords. Existing research has incorporated semi-structured data, such as controlled

vocabularies and knowledge bases, as a means of improving the quality of ranking by expanding the queries' entities [14]–[16]. However, in our problem the entities are provided by the users as a query component. Consequently, we proposed an original query framework, 'keyword queries on graph-annotated documents', which balances the importance of text relevance and semantic relevance.

A full presentation of the study of this problem is presented in Chapter 6.

The thesis is concluded in Chapter 7.

Chapter 2

A Study of the Demographics of Web-Based Health-Related Social Media Users

Background: The rapid spread of Web-based social media in recent years has impacted how patients share health-related information. However, little work has studied the demographics of these users.

Objective: Our aim was to study the demographics of users who participate in health-related Web-based social outlets to identify possible links to health care disparities.

Methods: We analyze and compare three different types of health-related social outlets: (1) general Web-based social networks, Twitter and Google+, (2) drug review websites, and (3) health Web forums. We focus on the following demographic attributes: age, gender, ethnicity, location, and writing level. We build and evaluate domain-specific classifiers to infer missing data where possible. The estimated demographic statistics are compared against various baselines, such as Internet and social networks usage of the population.

Results: We found that (1) drug review websites and health Web forums are dominated by female users, (2) the participants of health-related social outlets are generally older with the exception of the 65+ years bracket, (3) blacks are underrepresented in health-related social networks, (4) users in areas with better access to health care participate

more in Web-based health-related social outlets, and (5) the writing level of users in health-related social outlets is significantly lower than the reading level of the population. Conclusions: We identified interesting and actionable disparities in the participation of various demographic groups to various types of health-related social outlets. These disparities are significantly distinct from the disparities in Internet usage or general social outlets participation.

2.1 Introduction

Social media have been employed in many industries to engage consumers. The healthcare industry has moved at a slower pace in incorporating social media because of inherent risks such as patient privacy, but recently this rate has increased to fulfill the consumers' needs [17]. Moreover, some companies use social media to provide their employees with wellness videos to cut their health care costs [18].

At the same time, healthcare disparity is a well-studied problem in which two population groups receive unequal services [3]. This problem has been analyzed across various dimensions relating to social determinants in health, including: education and income, environmental hazards, and health outcomes including mortality, morbidity, and behavioral risk factors [19]. However, healthcare disparity has not been studied in terms of social media participation. This is important as Internet access and participation in health communities has the potential to improve health outcomes [4]. Hence, understanding the demographics of social outlets, which is the focus of this paper, may shed light to another facet of healthcare disparity.

To cover different types of online social outlets, we collected data from three types of sources: (i) general Online Social Networks, namely Google+ and Twitter (ii) drug review websites, and (iii) health web forums. We measure the following demographic attributes: age, gender, ethnicity, location, and writing level. Unfortunately, much of this information is unavailable for some, or all, of the sources. For that, we built and evaluated three classifiers for gender, ethnicity and writing level. User names were used for the gender and ethnicity classifiers. Writing level for users was calculated using modified reading level formula to ignore very long incomprehensible sentences. To extract the location of a post, we use a geocoding API.

2.2 Related work

2.2.1 Health-related social outlets analysis

Many researchers have explored the effectiveness of online social media in changing and improving the communication between providers and patients. According to Kane et al. [20], 60 million Americans are using Health 2.0 applications – social networks focused specifically on healthcare; further, approximately 40% of Americans find an opinion in social media is more trustworthy if it conflicts with a professional's opinion or diagnosis. Hackworth and Kunz [21], found that 80% of American adults have looked online for health-related topics. Recently, there is increased interest in analyzing the health-related content of social media [17]. Denecke and Nejdl [6], analyzed medical concepts mentioned in medical social media posts from different sources to differentiate between informative and affective posts. They found that patients and nurses tend to

share personal experiences, while physicians share health-related information. Lu et al. [7], studied the content of three disease-specific health communities and their relationship to five informative topics: symptoms, complications, examination, drugs and procedures. For example, users with breast cancer are more likely to discuss about examination, while users with lung cancer are more likely to discuss about symptoms. Wiley et al. [8], analyzed the content of online social media related to pharmaceutical drugs across several dimensions, including frequently mentioned diseases, keywords and sentiment. While the aforementioned work examined health-related social media and their content, none of them studied the demographics of the participating users, which is studied in this work.

2.2.2 Measure and estimate demographics of social outlets users

Survey- and classifier-based methods have been proposed: (a) Survey-based: In 2012, a Pew Research study showed that women, age 30 to 49, are more likely to participate in social media websites, where 75% of users are white [22]. eMarkter found that 68.9% of Hispanics use social media compared to 66.2% of the total population; further, they showed that Hispanics are more likely to compare products online while shopping and write reviews on products [23]. However, no research has focused on health-related social media. (b) Classifier-based: Mislove et al. [24], built methods to estimate both gender and ethnicity for Twitter users using the 1000 most popular first names reported by the U.S. Social Security Administration and frequently occurring surnames reported by 2000 U.S. Census. Gender and ethnicity methods used the reported first name and last name respectively. Mandel et al. [25], analyzed the tweets related to Hurricane Irene using Mislove's gender classifier. We build on Mislove's work when

creating our classifiers. While we also classify gender using first names, we extended these methods to screen names when first name is not present. A related work for estimating reading levels of the U.S. population [26] was presented to discuss limitations of low literacy patients. We measured the writing level based on this work since we didn't encounter any similar work.

2.3 Methods

2.3.1 Datasets

Our analysis used data collected from three different types of health social outlets: general social networks, drug review websites, and health web forums (Table 2.1). Google+ and Twitter were chosen as general social networks based on their popularity and number of users (we do not study Facebook, because it offers no public interfaces to access its data). For drug review websites and health web forums, three websites were selected for each, where we considered their breadth of topics and popularity. Figure 2.1 shows the overall process of our analysis, and table 2.1 shows key statistics of each source including number of users, number of posts and average sentence length. More information about the sources including start and end date is available in Appendix A.

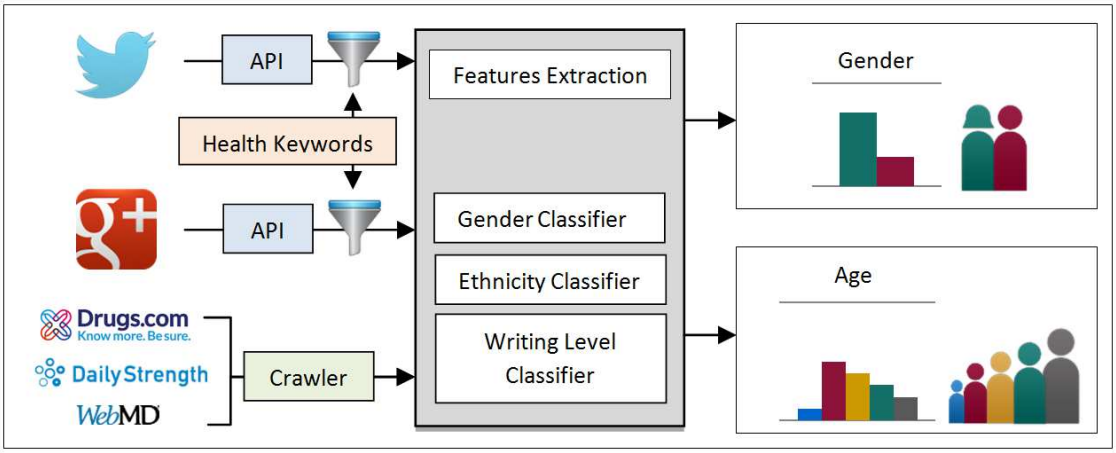

Figure 2.1 Overview of the data collection and analysis process.

Table 2.1 The total number of users, posts and average sentences length for each source.

Dataset	No. of users	No. of posts	Average sentence length (in words)
TwitterHealth [27]	5,095,849	11,637,888	10.82
Google+Health [28]	86,749	186,666	9.03
Drugs.com [29]	74,461	74,461	13.85
DailyStrength / Treatments [30]	213,524	1,055,603	11.92
WebMD / Drugs [31]	122,040	122,040	13.53
Drugs.com / Answers [32]	201,126	5,948,877	6.59
DailyStrength / Forums [33]	165,045	1,128,629	13.2
WebMD [34]	155,912	320,118	15.37

Table 2.2 shows which of the sources provide data for each of the five demographic attributes. Two demographics attributes are not present in any source: ethnicity and writing level; therefore, we created methods to automatically classify these attributes, along with gender for the sources where it is unavailable.

Table 2.2 List of all used sources with the available attributes. No indicates that the demographic attribute is not provided by the source. Yes indicates that the demographic attribute is provided by the source. Each classifier uses a distinct part of the user profile as denoted by the superscripts, where f stands for first name, l for last name, and s for screen name. The writing level classifier uses Flesch Kincaid measure based on all user's posts [35].

Dataset	**Age**	**Gender**	**Ethnicity**	**Location**	**Writing level**
TwitterHealth	NO	Gender classifier^{f}	Ethnicity classifier^{l}	YES	Writing level classifier
Google+Health	YES	YES		YES	
Drugs.com	NO	Gender classifier^{s}	NO	NO	Writing level classifier
DailyStrength / Treatments	YES	YES	NO	YES	
WebMD / Drugs	YES	YES	NO	NO	
Drugs.com / Answers	NO	Gender classifier^{s}	NO	NO	Writing level classifier
DailyStrength / Forums	YES	YES	NO	YES	
WebMD	NO	Gender classifier^{s}	NO	NO	

To filter health-related posts from Twitter and Google+, we built a list of 276 representative health-related keywords based on five categories. (i) Drugs: first we obtained a list of the 200 most popular drugs by prescriptions dispensed from RxList.com [36]. We then removed variants of the same drug (e.g., different milligram dosages) resulting in 125 unique drug names. (ii) Hashtags: we selected 11 popular health-related

Twitter hashtags such as #HCSM (Healthcare Communications & Social Media). (iii) Disorders: we selected 81 popular disorders such as cancer and Alzheimer. (iv) Pharmaceuticals: we selected the 12 largest pharmaceutical companies such as Pfizer. (v) Insurance: We selected 44 of the biggest insurances such as Medicare and Humana. A complete list of used keywords can be found in table B.1 of Appendix B.

We used the Twitter streaming API [37], with these keywords as filters, to obtain the relevant tweets for our TwitterHealth dataset. Our Google+Health dataset was collected via the Google+ API [38], where each health-related keyword was used as a query to find relevant posts. For the drug review websites and health web forums, we built custom crawlers in Java using the jsoup [39] library for crawling and parsing the HTML content. For each source, we collected the available data, including user information, posts, disorder or condition under which a discussion appears, keywords, tags, etc. We emphasize that we only collected publicly available data in accordance with each site's terms of use; no private data was collected.

2.3.2 User demographics estimation methods

We chose five demographic attributes as shown in Table 2.2: gender, age, ethnicity, location and writing level. Since these attributes are not available in every source, we created several classifiers to derive missing attributes as specified in Table 2.2. Note that we do not fill missing values of users for sources that provide this information for at least some of their users, e.g., if a user does not provide her age in Google+, we just ignore this user from the age-related analysis. Table D.1 of Appendix D shows the percentages of users who report each attribute in each source.

Gender

Four out of eight sources (Google+Health, DailyDtrength/Treatments, WebMD/Drugs, and DailyStrength/Forums) allow users to report their gender (as shown in Table 2.2). Approximately 80% of the users of these sources chose to report it; thus, the reported gender was used for these sources.

For the other sources where gender is not available, we extended the methods of Mislove et al. [24] to classify gender using the reported first name of users, if available; otherwise we extracted first names from user screen names. Note that screen names have not been used before, to the best of our knowledge, for gender estimation. In particular, we first collected the 1000 most popular male and female birth names reported by the U.S. Social Security Administration [40] for each year from 1935 to 1995. Thus, we collected the names of people currently (2014) having age from 19 – 79 years old, which constitute about 73.9% of the population [41]. There are 55,973 unique names in total. We further filtered this list to remove names with an aggregated frequency less than 10,000 or a discriminative gender probability less than 95%. The resultant list contained 1328 names. For TwitterHealth and google+, we checked if one of these 1328 first names is contained in the user-specified name to classify the user's gender. We first cleaned the first name by removing non alphabetical characters and then performed case-insensitive string matching. Gender classifier evaluation is reported in section C.1 of Appendix C; the accuracy ranges from 76% to 99%.

Age

Similarly, age was also reported in four sources (Google+Health, DailyStrength/Treatments, WebMD/Drugs, and DailyStrength/Forums); three sources display the age as a single number, whereas one source displays age as a range (e.g., 35-45). Approximately 61% of the users of these sources reported their age. When users provide an age range, the total number of users for each range is distributed uniformly to each year in the range. Ages are then grouped into five age groups: 0-17, 18-34, 35-44, 45-64, and 65 years and older. These age ranges are also used by the U.S. Census [42].

Ethnicity

The ethnicity of the users is not reported in any of the sources that we study; therefore, we created an ethnicity classifier similar to Mislove et al, [24]. The 2000 U.S. Census, which is the most recent available, reports the distribution of ethnicities for each last name (last names with less than 100 individuals were omitted) [43]. For example, the distribution for Hernadez is reported as 4.55% White, 0.38% Black, 0.27% Asian, and 93.81% Hispanic. We filtered this list to remove the last names with a frequency less than 1000, or where the discriminative probability of the majority ethnicity is less than 80%. We then use the ethnicity with the majority probability to classify ethnicity based on last name for sources that include the last name of users (Google+Health and TwitterHealth). We understand that race and ethnicity are not the same especially when referring to Hispanics, but in this paper, we try to simplify the presentation by only reporting ethnicity, that is, we do not distinguish groups like White Hispanic vs Black Hispanic, but only Hispanic. For the other sources (health web forums and drug review websites),

which do not have user names, we found that using the screen name for ethnicity estimation is inaccurate, and hence we do not report on the ethnicity of these sources. Ethnicity labeling and classifier evaluation is reported in section C.2 of Appendix C.

Location

Location is reported in four sources: the two general social networks (TwitterHealth, Google+Health), one drug review website (DailyStrength/Treatments), and one health web forum (DailyStrength/Forums); approximately 62% of users reported their locations. For TwitterHealth and Google+Health, users report their location using a single string (e.g., "NY, NY"). Thus, these strings are further processed to obtain structured locations (e.g., state: New York, city: New York). In particular, non-alphanumeric characters and extra spaces were removed, and location strings with a frequency less than 14 were removed. This left us with about 60% of TwitterHealth and Google+Health users with location strings. Each location string was mapped to a location (city, state, country) using the Google Geocoding API [38]. We focus on U.S. users and hence we remove users from other countries. DailyStrength/Treatments and DailyStrength/Forums list the user's city and state separately; thus, we use the reported state for these sources.

Writing Level

Different methods and formulas for measuring readability are available using different factors such as average number of syllables per words, average number of words per sentences, or average number of letters per words. In our work, we used the Flesch-

Kincaid Grade Level [35] formula to estimate the writing level (values generally correspond to school grades 1-12) of the users:

$$FKRA = (0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

where ASL is the Average Sentence Length, and ASW is the Average number of Syllable per Word.

Note that since we can only observe the text authored by users, we measure the writing level and not the reading level; however, we use the reading level formula since no alternative formula for the writing level exists. The writing level of a user is computed using the above equation by concatenating all of the user's posts and personal description; links and hashtags from tweets are removed, and users with less than 100 words in total are ignored. We found that very high reading level was being assigned to users that write very long incomprehensible sentences. This is a case that was not considered by the original FKRA formula which assumed that the text is grammatically and syntactically correct (e.g., the text of a novel). For that, we omit sentences with more than 30 words.

2.4 Results

To put our results in perspective, we compare them with other general demographics statistics. The population and Internet usage for each demographic group was obtained from the U.S. Census [41], [42], while other statistics for Twitter, and Google+ came from other sources [22], [43]–[45]. Further, we compare the demographics of the users participating in health-related discussions on Twitter and Google+ to the overall

demographics of the users in these sites. All our results are statistically significant, except the comparison between health web forums and drug review websites with respect to gender and age group (0-17). Also, there is no significant difference between Google+Health and Drug Review Websites for age group (35-44).

2.4.1 Gender

As shown in Table 2.3, the gender distribution in the population and Internet usage is almost the same, and there is a slight difference for general social networks. Our first key finding is that drug review websites and health web forums are dominated by female users; the number of female users is almost four times larger than that of male users. TwitterHealth and Google+Health have similar gender ratios when compared to the overall user base of Twitter and Google+.

Table 2.3 Gender distribution for TwitterHealth, Google+Health, drug reviews, health Web forums, compared to other relevant populations. The results in italics indicate results from this work. Non-italicized results are reported in the respective citations.

Source	F	M
Population [41]	51.05%	48.95%
Internet Use [42]	51.63%	48.37%
General social networks [22]	54.68%	45.32%
Twitter [44]	57.00%	43.00%
Google+ [44]	37.00%	63.00%
TwitterHealth	*51.81%*	*48.19%*
Google+Health	*35.36%*	*64.64%*
Drug Review Websites	*78.48%*	*21.52%*
Health Web Forums	*78.41%*	*21.59%*

The finding that women use health forums much more than men is partially supported by previous research, which shows that women report ill health more frequently than men [46]. In contrast, this is not true for Twitter and Google+, which are dominated by news exchanges [47].

2.4.2 Age

Table 2.4 reports the age distribution of users in the studied online social outlets and in other relevant sources, to put the results in perspective. Age groups were chosen based on Census; hence, we understand that the age ranges are not equal, but since our main goal is comparing the demographics of online health social outlets to other statistics such as internet usage, we chose to follow the Census age ranges in computer and internet access. Further, we provide population distribution in the same table to compare each group size with others. One-fifth of Internet users are in the group 0-17; this percentage drops to approximately 1% for drug review websites and health web forums. The majority of users on drug review websites are between 45 and 64 years old, and drug reviews have more users over 65 years than any other source; this is expected as older patients use more medications [48]. However, the percentage of drug review users above 65 is slightly lower than the percentage of internet users over 65, which means that older people still have low participation in Health 2.0 sites. Also, the 18-34 age group dominates health web forums, which is congruent with general social networks usage [45]. To summarize, our second key finding is that the participants of health-related social outlets are generally older than those of general-purpose social forums, but still

relatively low in the 65+ bracket; this is expected to change in the near future based on the participation statistics in the 45-64 bracket.

Table 2.4 Age distribution for Google+Health, drug reviews, health forums, and other relevant populations. The results in italics indicate results from this work. Non-italicized results are reported in the respective citations.

Source	0-17 years	18-34 years	35-44 years	45-64 years	65+ years
Population [41]	24.00%	23.11%	12.93%	26.53%	13.44%
Internet Use [42]	19.30%	27.55%	14.99%	28.36%	9.80%
General social networks [22]	14.58%	27.43%	20.68%	30.98%	6.32%
Google+ [49]	8.08%	71.61%	11.08%	7.82%	1.42%
Google+Health	*3.42%*	*53.21%*	*21.89%*	*19.02%*	*2.46%*
Drug Review Websites	*1.05%*	*31.13%*	*22.36%*	*36.84%*	*8.62%*
Health Web Forums	*1.03%*	*39.80%*	*25.81%*	*28.95%*	*4.41%*

2.4.3 Ethnicity

For the ethnicity and location analyses we focus on the US population, in order to compare to available U.S. census statistics. Table 2.5 shows the results of our ethnicity analysis. Recall that users' ethnicity in Google+Health and TwitterHealth is classified using our last name-based classifier. Our third key observation is Blacks are underrepresented in health-related social network discussions (Google+Health, TwitterHealth).

Table 2.5 Ethnicity distribution for TwitterHealth, Google+Health, and other relevant populations.

Source	Asian	Black	Hispanic	White
Population [50]	4.5%	12.2%	15.8%	65.1%
Internet Use [42]	5.46%	11.67%	13.94%	67.21%
General social networks [22]	5.26%	12.10%	14.53%	66.46%
Twitter [51]	N/A	9%	12%	71%
TwitterHealth	*3.24%*	*0.33%*	*23.46%*	*72.98%*
Google+Health	*5.6%*	*0.28%*	*17.4%*	*76.6%*

2.4.4 Location

In Figure 2.2, we show the distribution of users for each online health social outlets type, normalized by state population. Figure 2.2(A) shows the distribution of users in health web forums, figure 2.2(B) shows the distribution of users in drug reviews websites, figure 2.2(C) shows the distribution of users in TwitterHealth, and figure 2.2(D) shows the distribution of users in Google+Health combined.

To better understand these results, we created Table 2.6, which shows the correlation across all states between the normalized (by population) number of users in various health social outlets, and other societal measures. (More details are available in section D.2 of Appendix D). Our fourth key finding is that users in areas with higher income and more access to healthcare are more likely to participate in online health outlets, and particularly in web forums and drug review sites, which are the primary social sites for health-related information sharing [8]. Further, we see that in Twitter and Google+ the correlation with the number of physicians and education is higher.

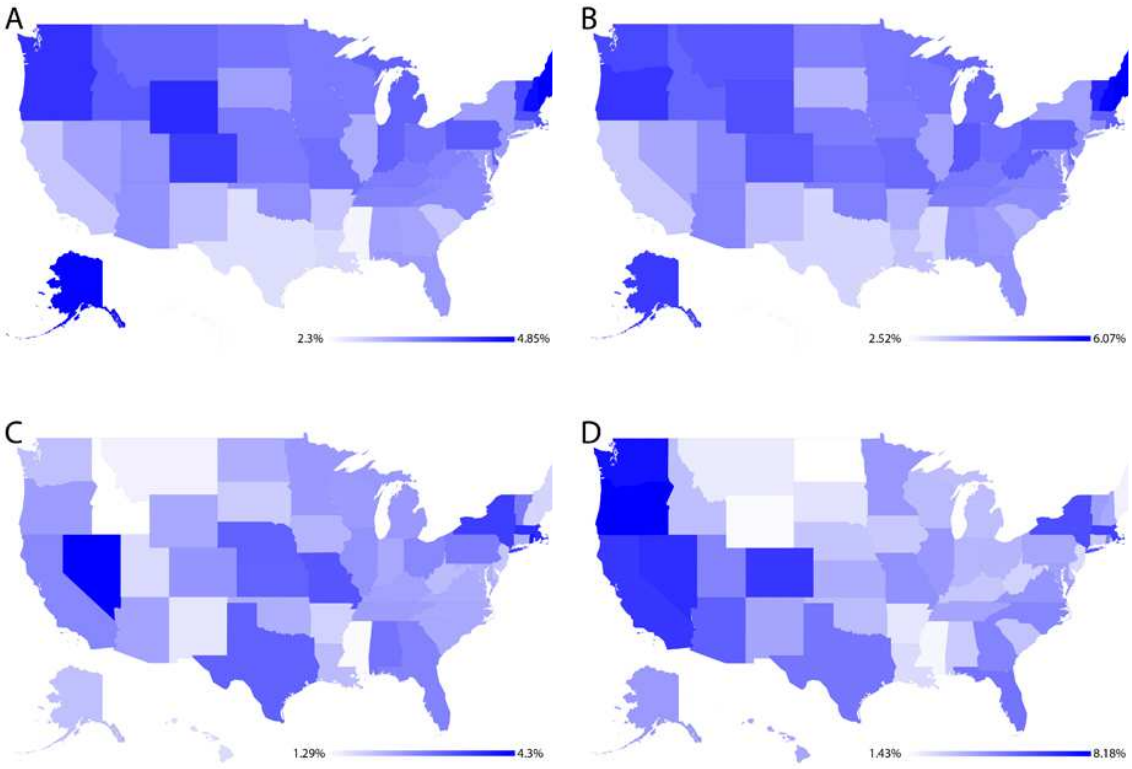

Figure 2.2 Per state capita number of users in (A) health web forums, (B) drug review websites, (C) TwitterHealth, and (D) Google+Health.

A reason could be that 59.1% of the about 878,194 US active physicians [52] participate in these networks [53], which is a significant number, as the geolocated subsets of the Google+Health and TwitterHealth datasets only contain 882,207 users in the U.S. The high correlation with education may be explained by the high percentage (91%) of Twitter users with college degree or higher [54].

Table 2.6 Correlation across all states between the normalized (per capita) number of users for each type of health social outlets, and each state's population, normalized number of Internet users, normalized number of physicians, normalized number of uninsured patients, average annual income and percentage of population with college degree or higher.

Correlation	Health Web Forums	Drug Review Websites	TwitterHealth	Google+Health	Google+
Internet usage [42]	0.19	0.28	0.01	-0.01	0.00
No. of physician [52]	0.37	0.19	0.88	0.80	0.44
Uninsured population [55]	-0.40	-0.40	-0.17	-0.11	-0.10
Annual Income [56]	0.38	0.27	0.17	0.25	0.26
Education (ratio of people with college degree) [57]	0.35	0.22	0.56	0.63	0.54

2.4.5 Writing Level

The writing level, as previously mentioned, is measured using a standard reading level formula that assigns a school grade to the given text. For example, when a person writes text at a 5th grade reading level, it implies that his or her writing should be understood by people that have passed the 5th grade. Table 2.7 reports our results for writing level of health social outlets users. We see that Google+Health users have generally higher writing level than the rest sources, which may mean that more of the Google+Health users are professional accounts.

Next, we try to put these findings in perspective. Unfortunately, related work only reports on reading levels (and not writing levels) of the U.S. population participating in social outlets. Thus, we compare our results in Table 2.7 to Figure 2.3, which reports the reading level of the general U.S. population [26].

Table 2.7 Writing Level distribution for TwitterHealth, Google+Health, drug reviews, and health forums.

Source	0 to 5	6 to 9	10 to 16
TwitterHealth	*37.77%*	*51.09%*	*11.13%*
Google+Health	*6.45%*	*55.63%*	*37.91%*
Drug Review Websites	*30.42%*	*66.17%*	*3.41%*
Health Web Forums	*28.79%*	*68.24%*	*2.98%*

Our fifth key finding is that the writing level in health social outlets (Table 2.7) is generally lower than the reading level of the population (Figure 2.3). Thus users/patients can easily comprehend the posts and hence benefit from the experiences of other users.

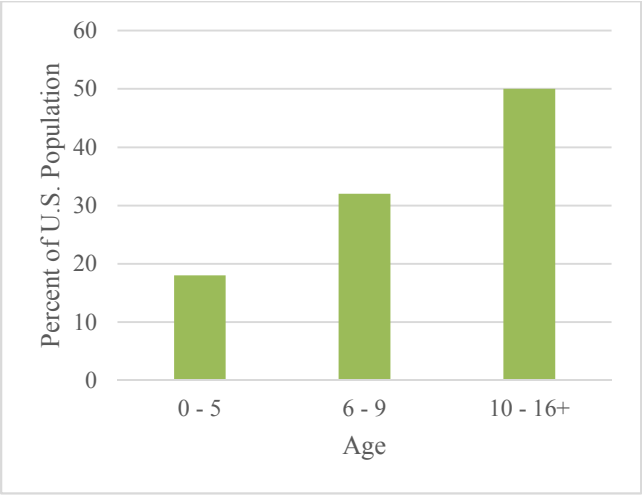

Figure 2.3 Reading level of U.S. Population [26].

The benefit of social interaction with respect to health empowerment has been demonstrated before [58]. In an online epilepsy community, 59% of patients found another patient who is experiencing the same symptoms, 58% had a better understanding of seizures, and 55% learned more about treatments and symptoms.

2.4.6 Statistical Significance Tests

Tables 2.8, 2.9, and 2.10 report the p-values for Pearson's Chi Squared test of independence and the Mann-Whitney U test. Note that we only compute significance values between sources that we have analyzed and not between our sources and sources analyzed by other works (such as Google+ [28]) since we don't have the raw data for those sources.

Table 2.8 p-values for Pearson's Chi Squared test of independence.

	Gender	**Age**	**Ethnicity**	**Writing Level**
TwitterHealth vs. Google+Health	< 0.001	N/A	< 0.001	< 0.001
TwitterHealth vs. Health Web Forums	< 0.001	N/A	< 0.001	< 0.001
TwitterHealth vs. Drug Review Websites	< 0.001	N/A	< 0.001	< 0.001
Google+Health vs. Health Web Forums	< 0.001	< 0.001	< 0.001	< 0.001
Google+Health vs. Drug Review Websites	< 0.001	< 0.001	< 0.001	< 0.001
Health Web Forums vs. Drug Review Websites	< 0.001	< 0.001	< 0.001	< 0.001

Table 2.9 p-values for Mann-Whitney U test

	TwitterHealth vs. Google+Health	**TwitterHealth vs. Health Web Forums**	**TwitterHealth vs. Drug Review Websites**
Gender - Male	< 0.00001	< 0.00001	< 0.00001
Gender - Female	< 0.00001	< 0.00001	< 0.00001
Age (0-17)	N/A	N/A	N/A
Age (18-34)	N/A	N/A	N/A
Age (35-44)	N/A	N/A	N/A
Age (45-64)	N/A	N/A	N/A
Age (>=65)	N/A	N/A	N/A
Ethnicity (White)	< 0.001	< 0.0001	< 0.00001
Ethnicity (Black)	0.6339	< 0.00001	< 0.00001
Ethnicity (Asian)	< 0.00001	< 0.001	< 0.01
Ethnicity (Hispanic)	< 0.00001	< 0.00001	< 0.00001
Writing Level (0-5)	< 0.00001	< 0.00001	< 0.00001
Writing Level (6-9)	< 0.00001	< 0.00001	< 0.00001
Writing Level (10-16)	< 0.00001	< 0.00001	< 0.00001

Table 2.10 p-values for Mann-Whitney U test

	Google+Health vs. Health Web Forums	**Google+Health vs. Drug Review Websites**	**Health Web Forums vs. Drug Review Websites**
Gender - Male	< 0.00001	< 0.00001	0.5797
Gender - Female	< 0.00001	< 0.00001	0.5797
Age (0-17)	< 0.001	< 0.001	0.5144
Age (18-34)	< 0.0001	< 0.00001	< 0.00001
Age (35-44)	0.01661	0.7747	< 0.00001
Age (45-64)	< 0.00001	< 0.00001	< 0.00001
Age (>=65)	0.01066	< 0.00001	< 0.00001
Ethnicity (White)	< 0.00001	< 0.00001	0.1316
Ethnicity (Black)	< 0.00001	< 0.00001	0.0944
Ethnicity (Asian)	< 0.00001	< 0.001	0.8054
Ethnicity (Hispanic)	< 0.001	< 0.00001	0.6503
Writing Level (0-5)	< 0.00001	< 0.00001	< 0.00001
Writing Level (6-9)	< 0.00001	< 0.00001	< 0.00001
Writing Level (10-16)	< 0.00001	< 0.00001	0.00516

2.5 Discussion

Our results can help healthcare providers customize educational campaigns for different groups. For example, white women should be informed to a larger extend on the possible misinformation spreading in health web forums, since they participate much more.

Regarding mitigating ethnicity-based healthcare disparities, we found that Twitter and Google+ are more effective in reaching out to Hispanics about healthcare offerings. However, this is not true for black ethnicity, who are not overrepresented in any health social outlet. This means that there is no single outlet to reach black population, which has been shown to receive worse health care by about 40% comparing to white population [59].

Advertisers may use our results to decide on the best sites to advertise their products; for instance, drug review websites are more appropriate than Google+ to advertise drugs for the 45-64 age bracket, but the opposite is true for the 18-34 age bracket. Further, drug reviews websites and health web forums are better to target female when advertising for their products than other health social outlets.

In the age results section, we found that younger groups (18-34 years old) participate in large numbers to health forums, which may sound counterintuitive. By analyzing posts for this age bracket, we found the most popular keywords are related to pregnancy such as birth control, ovulation, and miscarriage. On the other hand, their participation is lower for drug review websites. A possible explanation may be that often

patients who talk about pregnancy are not taking any drugs, compared to other conditions like diabetes, where drugs are more common.

We also attempt to explain the disparities in the participation in health social outlets based on socioeconomic factors through the state-level participation distributions. Our results in Table 2.6 show that less access to physicians does not lead to higher participation in health social outlets as one would expect. In contrast, it seems that the participation to such outlets is correlated with the access to healthcare and the average income.

The weak but positive correlation between income and participation to health web forums and drug review sites may be partially attributed to the higher internet usage of the more affluent groups, as shown in Table 2.6. Another possible explanation is that lower income or uninsured persons are more likely to be part of a community with healthcare disparities [50]. The positive correlation between education and participation in health social outlets, especially Google+Health and TwitterHealth, may be partially explained by the fact that people with college degree are less likely to be uninsured, since 10% only of college graduates are uninsured, compared to 40% of adults who have not graduated from high school [60]. In addition, 60% of uninsured people are from families with low income [61], and the group of people with income lower than 30K is the lowest group in terms of accessing health information [62], Hence, our results show that people with low income have less access to health information.

On the other hand, we found that the content in health social outlets is easy to understand for almost all users, given the low writing level. That is, the well-known

health literacy issue, which is more severe in low-income and lower education populations [4], does not seem to apply to online health social outlets. Of course, the low writing level does not address the issue of language, as many low income and education users in the U.S. do not speak English at home [62].

2.5.1 Limitations

Our ethnicity and gender classifiers are not perfect, as shown in Appendix C, and thus introduce an error into our analyses. This issue is less significant for gender, since out of all users included in our gender analysis for health web forums and drug review websites, a majority of the users (over 94%) report their gender, and hence the classifier was only used for 6% of users. Further, a majority of users in drug review websites and health web forums are female, and our gender classifier obtained an accuracy greater than 99% for females when using screen name.

Another limitation is the informal writing style of social media posts, as our writing level method uses the average sentence length, which expects that posts are properly punctuated. We addressed this limitation to some degree by only considering sentences of a reasonable length (less than 30 words). Estimating writing level could have been improved by considering other features like typos or spelling mistakes. Further, it would be useful to measure the quality of the posted information, in addition to just the writing level. This is a very hard problem, which we leave as future work.

Since all the attributes are reported by users, there is inevitably self-selection bias. In particular, gender, age, and location are not mandatory in any site. For instance, older people may choose not to report their age. Moreover, choosing to report the real names,

or posting profile pictures could also create self-selection bias in our gender and ethnicity classifiers. There may also be various types or degrees of bias across different outlets. For instance, WebMD users may use their real name less frequently than Twitter users. This in turn may bias the study results, especially for ethnicity where we depend completely on the classifier results.

2.6 Conclusion

We studied user demographics in online health social outlets, which we split into three different types: social networks, drug review websites, and health web forums. The distributions of the demographic attributes – gender, age, ethnicity, location and writing level – have been analyzed for each source type and compared with relevant baseline user distributions like Internet and general social outlets participation. The results reveal interesting and often unexpected disparities with respect to all demographic attributes.

Chapter 3

Demographic-Based Content Analysis of Web-Based Health-Related Social Media

Background: An increasing number of patients from diverse demographic groups share and search for health-related information on Web-based social media. However, little is known about the content of the posted information with respect to the users' demographics.

Objective: The aims of this study were to analyze the content of Web-based health-related social media based on users' demographics to identify which health topics are discussed in which social media by which demographic groups and to help guide educational and research activities.

Methods: We analyze 3 different types of health-related social media: (1) general Web-based social networks Twitter and Google+; (2) drug review websites; and (3) health Web forums, with a total of about 6 million users and 20 million posts. We analyzed the content of these posts based on the demographic group of their authors, in terms of sentiment and emotion, top distinctive terms, and top medical concepts.

Results: The results of this study are: (1) Pregnancy is the dominant topic for female users in drug review websites and health Web forums, whereas for male users, it is cardiac problems, HIV, and back pain, but this is not the case for Twitter; (2) younger users (0-17 years) mainly talk about attention-deficit hyperactivity disorder (ADHD) and

depression-related drugs, users aged 35-44 years discuss about multiple sclerosis (MS) drugs, and middle-aged users (45-64 years) talk about alcohol and smoking; (3) users from the Northeast United States talk about physical disorders, whereas users from the West United States talk about mental disorders and addictive behaviors; (4) Users with higher writing level express less anger in their posts.

Conclusion: We studied the popular topics and the sentiment based on users' demographics in Web-based health-related social media. Our results provide valuable information, which can help create targeted and effective educational campaigns and guide experts to reach the right users on Web-based social chatter.

3.1 Introduction

As Web-based social media are growing in popularity, the number of people who share their experiences or ask for support in health-related social media has also increased [63]. Fox and Jones have found that 41% of e-patients have read someone else's commentary or experience about health on a Web-based news group, website, or blog [64]. Kane et al [20] reported that more than 60 million Americans read or contribute to Health 2.0 apps, in which they consider these apps as their first source when gathering data and opinions. About 40% of Americans doubt a professional opinion when it conflicted with what they form from Web-based health social media [20].

One of the key benefits of health-related Web-based social media reported by researchers is the increased access to information to various demographic groups, regardless of age, education, income, or location [65]. However, previous work has

mainly relied on user surveys to study the effect of the use of social media to health-related factors such as psychological distress [66]. In addition, previous work does not reveal granular information on what disorders or other health topics are mostly discussed in the Internet by each demographic group, which would allow health care providers to create targeted and effective educational campaigns.

In this work, we conducted the first, to our best knowledge, large-scale data-driven comparative analysis of the content of health-related social media across various demographic dimensions—gender, age, ethnicity, location, and writing level. For each demographic group, we study the content of the posts across the following dimensions: sentiment, popular terms (keywords), and medical concepts (particularly disorders and drugs). Concepts refer to entries in the Unified Medical Language System (UMLS) vocabulary [67], whereas terms are just words from the posts' text that may or may not belong to any UMLS concept. We report results for 3 types of social media: (1) general Web-Based Social Networks, namely Google+ and Twitter, (2) drug review websites, and (3) health Web forums. The selection of social media types was based on their popularity and on our study of the literature on health-related social content [68]. The objective of this study was to identify which health topics are discussed in which social media by which demographic groups, to better guide educational outreach and research activities.

3.2 Related Work

3.2.1 Analysis of Health-Related Social Outlets

Different studies were established and conducted by researchers to study the effectiveness of Web-based social media in changing and improving the communication between providers and patients. Hackworth and Kunz [21] reported that 80% of Americans have searched the Internet for health-related information. Grajales et al [5] illustrated how, when, and why social media are used by health care sectors by conducting a narrative review of case studies, and they provided 4 recommendations that stakeholders may consider to engage with social media. Because analyzing the health-related content of social media is increased recently [17], Denecke and Nejdl [6] performed content analysis of medical concepts in different health-related social media sources. They presented a method to classify posts as informative or affective, and they found that doctors share health-related information, whereas patient and nurses are more likely to share personal experiences. Lu et al [7] analyzed the content of 3 disease-specific health communities including lung cancer, breast cancer, and diabetes and defined their relationship to 5 main informative topics: symptoms, complications, examination, drugs, and procedures. This study shows that examination is a hot topic for users with breast cancer, whereas symptoms are more likely to be discussed by users with lung cancer. Wiley et al [8] analyzed the content of drug-related chatter on various social media forums. The study demonstrates that Web-based social media's characteristics such as moderation affect the discussions in different ways including subjectivity and type of drugs discussed.

3.2.2 Measuring and Estimating Demographics of Users of Social Media

Krueger et al [69] studied the mortality attributable to low education level in the United States. They found people with less than high school degree have more mortality rate; thus, improving the US educational attainment could increase the survival in US population. A Pew research conducted in 2012 showed that white ethnicity represents 75% of social media websites users, where women in age group of 30-49 years participate more in these websites [24]. Another study by eMarkter found that Hispanic are more active in social media with 68.9% of them using social networks compared with 66.2% of total US population [25]. Mislove et al [70] estimated gender and ethnicity for Twitter users. The gender is estimated by using the reported first name and comparing it to the 1000 most popular first names reported by the US Social Security Administration, whereas ethnicity is estimated by using the reported last name and comparing it to the frequently occurring surnames reported by 2000 US Census. Using Mislove's gender classifier, Mandel et al [25] analyzed the tweets related to Hurricane Irene. Liu et al [71] proposed Natural Language Processing (NLP) methods to extract the demographics (gender, age, ethnicity) of users of social posts. Anderson-Bill et al [72] recruited Web-health users to examine their demographics, behavioral, and psychosocial characteristics, and they found that Web-health users are more likely middle-aged, upper class, and well-educated women. Although the aforementioned work examined health-related social media and their content, none of them studied how different demographics use Web-based social media, which is studied in this work.

In Chapter 2, we studied how many users from each demographic group (by gender, age, ethnicity, location, and writing level) participate in various social media, but it did not study the content of the posts, which is the focus of this paper. Some of the key findings of that work are: (1) drug review websites and health Web forums are dominated by female users; (2) the participants of health-related social media are generally older with the exception of the 65+ years bracket; (3) Asian and black ethnic groups are underrepresented in drug review websites and health Web forums, and blacks are also underrepresented in health-related Web-based social networks; (4) users in areas with better access to health care participate more in Web-based health social media; and (5) the writing level of users in health social media is significantly lower than the reading level of the population.

3.3 Methods

3.3.1 Key Challenges

A key challenge is to estimate the demographic group, for example, gender, of a Web-based user when this information is not explicitly stated. Another challenge in this work is the extraction of medical concepts from social posts, given that existing tools such as MetaMap focus on biomedical text, which is generally generated by researchers or practitioners; therefore, we filtered out some misclassified concepts generated by this tool to work on health social media posts. Another challenge has been the time to extract the medical concepts from the social posts. In this paper, we process more than 20 million posts, which would take several months to parse on a single machine. For that, we

have parallelized this into 10 machines that extracted all concepts in about 1 month. To extract popular terms for each demographic group, we use stemming to merge together terms with the similar root.

3.3.2 Datasets

As summarized in Table 3.1, for general social networks, we chose Twitter and Google+ for their popularity and number of users (we did not include Facebook as it does not provide public data). For the other 2 types, we selected 3 different websites for each one to ensure diversity. More information about the sources including start and end date is available in Table A.1 and A.2 of Appendix A. Because Twitter and Google+ are general social networks, we filtered the posts using 276 representative health-related keywords as follows: (1) Drugs: from the most prescriptions dispensed from RxList.com, we selected the 200 most popular drugs [36]. By removing the variants of the same drug (eg, different milligram dosages), the final list of drugs contained 125 unique drug names. (2) Hashtags: from Twitter Hashtags, we selected 11 popular health-related Twitter hashtags such as #BCSM (Breast Cancer and Social Media). (3) Disorders: 81 popular disorders were selected such as AIDS and asthma. (4) Pharmaceuticals: the 12 largest pharmaceutical companies were selected such as Novartis. (5) Insurance: 44 of the biggest insurances were selected such as Aetna and Shield. The rationale of selecting the keywords was to cover as much as we can by including popular drugs and disorders, popular health-related hashtags in Twitter, and other related health keywords that can help increase the number of the posts related to health, similar to previous work on Twitter filtering [8]. A complete list of used keywords can be found in Table B.1 of

Appendix B, and all terms' frequencies for both sources can be found in Table B.2 of Appendix B.

Table 3.1. List of all used sources with their number of posts and with the available demographic attributes.

Dataset	No. of posts	Gender[a]	Age[a]	Ethnicity[a]	Location[a]	Writing level
TwitterHealth [27]	11,637,888	Gender classifier	NO	Ethnicity classifier	YES	Writing level classifier
Google+Health [28]	186,666	YES	YES		YES	
Drugs.com [29]	74,461	Gender classifier	NO	NO	NO	Writing level classifier
DailyStrength/ Treatments [30]	1,055,603	YES	YES	NO	YES	
WebMD/Drugs [31]	122,040	YES	YES	NO	NO	
Drugs.com/ Answers [32]	320,118	Gender classifier	NO	NO	NO	Writing level classifier
DailyStrength/ Forums [33]	5,948,877	YES	YES	NO	YES	
WebMD [30]	1,128,629	Gender classifier	NO	NO	NO	

[a]*NO* indicates that the demographic attribute is not provided by the source and no classifier is used due to low accuracy. *YES* indicates that the demographic attribute is provided by the source. More details on the demographic classifiers are available in the Chapter 2.

Then, to filter out Twitter using the health-related keyword list, we used the Twitter streaming Application Program Interface (API) [37] to extract the relevant tweets for TwitterHealth. Google+Health posts were collected via the Google+ API [38], in which the health-related keyword list was used in the queries to obtain relevant posts for

Google+. For the other drug review websites and health Web forums, we built a crawler for each website in Java using the Java library jsoup [39] for extracting and parsing hypertext markup language content. For each website, we crawled and collected the available data, including public user information, posts, disorders, conditions, keywords, tags, rating, and so forth. We emphasize that we do not collect or use any private data, and we only collected publicly available data in accordance with each site's terms of use. Figure 3.1 shows the overall process of our analysis.

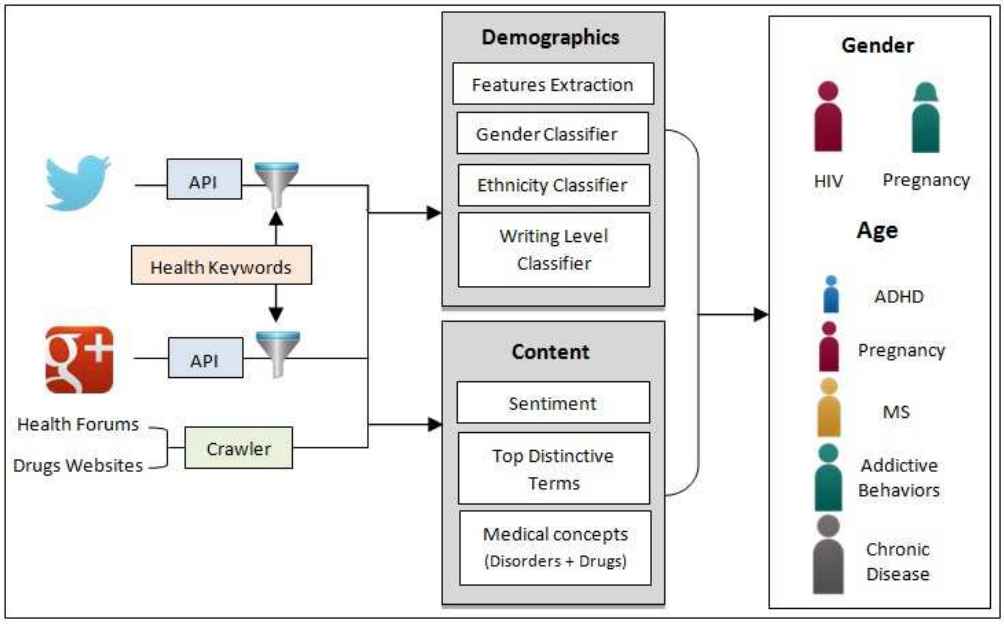

Figure 3.1 Overview of the data collection and analysis process.

3.3.3 Demographic Data Computation

The demographic data (age, gender, ethnicity, location, writing level) of users are extracted from the data or estimated through classifiers as discussed in Chapter 2 where more statistics of the collected posts such as dates and number of users are also reported.

As summarized in Table 3.1, gender attribute is either reported by the source or generated by a classifier that uses the first name to distinguish between male and female. Age and location, on the other hand, are used as they reported by the source; however, location was further processed to map user's input into geolocations using Google API [38]. Because ethnicity is not reported in any source, we used a classifier that uses the last name to predict ethnicity for users in sources that provide the last name. For writing level, we measured each user writing level using a modified version of Flesch–Kincaid Grade Level [35].

3.3.4 Sentiment and Emotion

To compute sentiment and emotion, we map each phrase in the post to a phrase from a sentiment lexicon. We use a sentiment lexicon, SentiWordNet [73], and an emotion lexicon, NRC word-emotion lexicon [74]. These 2 lexicons were selected owing to their effectiveness and popularity in previous studies [6], [75], [76] and because they cover complementary aspects. We use the SentiWordNet dictionary for sentiment, which assigns positive, negative, and objective score to each term where the sum of all 3 scores equals 1. Because SentiWordNet uses senses and part of speech, the Stanford CoreNLP Trigger [77] was used to tag each word with its part of speech tag. All words in posts and SentiWordNet were then stemmed to remove words variation. The longest possible match is then used to map each phrase in posts to a phrase from SentiWordNet, and after that, each post's sentiment is calculated by averaging scores of all phrases. For each source, the total sentiment score for each demographic attribute is measured by averaging all posts' scores associated with that attribute and normalized by the number of posts of the

attribute. For emotion, we use the NRC word-emotion lexicon, which measures anger–fear, trust–disgust, and anticipation–surprise.

3.3.5 Top Distinctive Terms

The content of all sources was analyzed to get the top distinctive terms for each source. All posts are first filtered to remove stop words and then stemmed using Porter stemmer [78]. From these words, we considered only the ones that occur in at least 0.01% of the total number of posts that are annotated for a given demographic attribute value (or 30 if 0.01% is less than 30). That is, if less than 0.01% of posts from users who reported their gender contain a term, this term is not reported in either male or female group analysis. Then, for each demographic attribute value, that is, male, we normalized the number of occurrences for each term in that attribute value by the number of users posts in the same attribute value to get the frequency, for example:

$$\text{Freq}_{male}(\text{headache}) = \text{No. of occurrences (headache) in male} / \text{No. of male posts} \quad (3.1)$$

To get the top 10 distinctive terms for each demographic attribute, we then calculated the relative difference as follows:

$$RelDif_{male}(headache) = [Freq_{male}(headache) - AvgFreq_{gender}(headache)] / AvgFreq_{gender}(headache) \quad (3.2)$$

Where $AvgFreq_{gender}(headache)$ is the average frequency of the word headache in all posts by male or female users. For example, $AvgFreq_{location}(headache) = [Freq_{Northeast}(headache) + Freq_{Midwest}(headache) + Freq_{South}(headache) + Freq_{West}(headache)]/4$. Finally, we only display health-related terms in each demographic group that have a

relative difference greater than 0.1; that is, we decided to hide results with a difference of less than 10% from the average score, which we believe is intuitive.

3.3.6 Medical Concepts

To annotate posts with corresponding medical concepts from the UMLS [79], the MetaMap tool [13] was used to represent each post as a set of medical concepts. Because MetaMap was originally built to extract concepts from biomedical text generated by researchers or practitioners, it is not perfect to annotate social media posts [80]. Therefore, we manually removed some annotations that were misclassified by MetaMap as following: (1) we order generated concepts by their frequencies for each source systematically, (2) we analyze each phrase that was mapped for each concept, and (3) we delete the misclassified UMLS concepts from the results. For example, the letter "i" mapped to (immunologic factor) and word bad mapped to (organic chemical). Such mistakes were deleted from MetaMap annotations to improve accuracy. In UMLS, we have 15 semantic groups (eg, Disease or Anatomy), and each concept in UMLS is associated with one or more semantic types, where each semantic type belongs to 1 semantic group. In this part, we analyzed only 2 semantic groups including drugs and disorders, and we reported the top distinctive drugs and disorders for each demographics using the same threshold and method used in finding top distinctive terms (Equation 3.2).

3.4 Results

In this section, we present our results for sentiment and emotion, top distinctive terms, and medical concepts by each demographic group. Two medical concept types

were considered and reported to avoid less interesting results: disorders and drugs. For each demographic group, we show the top distinctive disorders and drugs using Equation 3.2 that have a relative difference more than 0.1. Some demographic attribute values are not reported owing to small number of users (age group (0-17) and (65+) in Google+Health), or demographic attribute is not reported by the source (all age groups in TwitterHealth), or because users talk about unrelated health topics (writing level (0-5) in TwitterHealth talk about astrology), or the relative difference (Equation 3.2) for the top findings is less than 0.1.

3.4.1 Gender

In Table 3.2, we summarize the top distinctive (highest relative difference according to Equation 3.2) terms by gender; note that some demographic attributes such as female in Google+Health do not have distinctive terms. Because Twitter and Google+ are more news-based social media, many health posts share news in different areas including politics and sports—we excluded them to include health-related keywords only. Our first key finding is that male users in TwitterHealth tend to talk more about the reproductive system, tumor and AIDS, and health insurance, whereas female users talk about headache and emotion. In drug review websites and health Web forums, female users tend to talk more about pregnancy-related topics, whereas male users discuss pain drugs, cholesterol, and heart problems.

Table 3.2. Top 10 distinctive terms by gender.

Gender	TwitterHealth	Google+Health	Drugs	Forums
Male	Prostate, Gay, Testicular, Viagra, Tumor, AIDS, Obamacare, Marijuana, Medicare, Insurance	Pharmacology, Encephalomyelitis, Amphetamine, Pertussis, Fukushima, Pfizer, Novartis, Neutrophil, Biomed, Viagra	Wife, Oxycontin, Urine, Lisinopril, Cholesterol, Hydrocodone, Disc, Spinal, Libido, Diovan	HIV, Wife, Tinnitus, Gay, Cholesterol, Artery, AA (alcoholic anonymous), Valium, Cardiologist, Alcohol
Female	Cry, Migraine, Moody, Frown, Pound, Laugh, Nap, Eczema, Headache, Tension	N/A	Ovulation, IUI (intrauterine insemination), Pregnancy, Clomid (used to cause ovulation in women), IVF (in vitro fertilization), Pregnant, Birth, Boyfriend, BC (birth control), Fibromyalgia	Miscarry, PCOS (polycystic ovary syndrome), Endometriosis, Lupron, Uterus, Hysterectomy, Infertility, Ovarian, Rheumatologist, Progesterone

In Table 3.3, we summarize top distinctive disorders by gender. Male users in drug review websites mainly talk about back pain and blood pressure, whereas female users talk about pregnancy. In health Web forums and websites, male users discuss heart

problems and panic topics, and female users talk more about skin disorders, headache, and chronic fatigue disorders. In TwitterHealth and Google+Health, top disorders discussed by male users can be classified as sexually transmitted diseases, including AIDS and herpes, whereas female in TwitterHealth as seen in the top distinctive terms discuss topics related to headache and feelings.

Table 3.3. Top 5 distinctive disorders by gender.

Gender	TwitterHealth	Google+Health	Drugs	Forums
Male	Acquired Immunodeficiency Syndrome (AIDS), HIV seropositivity, Cerebrovascular accident (stroke), Incised wound, Herpes NOS	Gonorrhea, Marijuana abuse, Sexually transmitted diseases, Malignant neoplasm of lung, Infantile neuroaxonal dystrophy	Low back pain, Dry cough, Blood pressure finding, Back pain, Diabetic	Atrial fibrillation, Codependency, Panic attacks, Diabetes, Marijuana abuse
Female	Migraine disorders, Emotional, Headache, Pain NOS adverse event, Asleep	Chronic fatigue syndrome	Gravidity; Endometriosis, site unspecified; Yeast infection; Fibromyalgia; Hot flushes	Dermatitis herpetiformis, familial; Lupus vulgaris; Lupus erythematosus, systemic; Fibromyalgia; Migraine disorders

Table 3.4 summarizes top distinctive drugs by gender. In drug review websites, the top drugs discussed by female users are related to pregnancy including birth control and ovulation stimulation, whereas male users talk mainly about drugs related to blood pressure. In health Web forums ,male users discuss depression-related drugs and alcohol topics. In TwitterHealth, not many distinctive drugs were found for female and male users, whereas in Google+Health, different drugs and chemicals were reported. Sentiment and emotion were evaluated for all sources. Because the results look similar between gender groups, we summarize the results in Tables E.1 and E.2 of Appendix E.

Table 3.4. Top 5 distinctive drugs by gender.

Gender	TwitterHealth	Google+Health[a]	Drugs[a]	Forums[a]
Male	Viagra	Aldosterone, DC101 monoclonal antibody, Bicarbonates, Aspartame, Methamphetamine[1]	Low-density lipoproteins, Plavix, Bystolic[6], Oxycodone, Opiates	Alcohols, Xanax[4], Detox adjuvant[2], Prozac[4], Dietary lead
Female	Trivalent influenza vaccine	Thioctic acid, Detoxadjuvant[2], Seroquel	Yaz[3], Implanon[3], Tamoxifen[2], Estrogens, Clomid[3]	Plaquenil, Diamox, Topamax, Concerta, Synthroid

[a]Some of the drugs are coded to match the corresponding disorders they treat: [1]ADHD, [2]Cancer, [3]pregnancy, [4]depression, [5]MS, [6]BP, heart problem and cholesterol, [7]Diabetes.

3.4.2 Age

Table 3.5 summarizes the top 10 distinctive terms for each age group. Generally, for younger groups (0-17 years), ADHD and skin problems are popular topics in drug review websites, whereas in health Web forums, they talk more about parents and homosexuality. For age groups of 18-34 years in drug review websites and health Web forums, the main topics discussed are related to relationships, pregnancy, or getting pregnant using simple intervention methods, or family members; whereas the same groups in Google+Health talk about different aspects including vitamins and sleep disorders. Age group of 35-45 years also discusses pregnancy topics but using sophisticated intervention methods including in vitro fertilization. Age group of 45-64 years, as in Table 3.4, discusses topics related to chronic diseases including fibromyalgia, disc, and cholesterol, and it also discusses other topics including addiction to smoking, alcohol, and menopause. HIV also appears to be a popular topic for that group in Google+Health and health Web forums. Finally, people aged older than 65 years also talk more about chronic diseases and heart-related problems including drugs that can help mitigate the pain. We see that most topics are more likely discussed by women because drug review websites and health Web forums are dominated by female users [21].

Table 3.6 summarizes top distinctive disorders by age. In drug review websites, the young age group of 0-17 years talks more about skin disorders and mental disorders, whereas the same age group in health forums websites discusses mainly mental disorders. Age groups of 18-34 years and 35-44 years in drug review websites talk about pregnancy

and mental disorder topics. Older age groups in both sources tend to talk about diabetes, heart diseases, and muscles pain.

Table 3.5. Top 10 distinctive terms by age.

Age, years	Google+Health	Drugs	Forums
0-17	N/A	Concerta, Acne, ADHD, Birth, Wash, Lip, Prescribed, Boyfriend, Skin, Scar	Lesbian, Bullying, Buddy, Gay, Mum, Crush, Suicide, Rape, Teen, Dad
18-34	Supplement, Arthritis, Weight, Vitamin, Headache, Hospital, Friend, Food, Love, Skin	Clomid (used to cause ovulation in women), Ovulation, Phentermine, Calorie, Pregnancy, Gym, Pregnant, Baby, BC (birth control), Workout	BC (birth control), Clomid (used to cause ovulation in women), Ovulation, PCOS (polycystic ovary syndrome), TTC (trying to conceive), Miscarried, Fiance, Baby, Pap, Conceive
35-44	Vitamin, Sleep, Food, Parkinson, Friend, Healthcare, Community, Vaccine, Pain, Insomnia	IVF (in vitro fertilization), IUI (intrauterine insemination), Clomid, Ovulation, Marriage, Divorce, Mania, Narcotic, Lithium, Kid	IVF (in vitro fertilization), IUI (intrauterine insemination), BFP (big fat positive), BFN (big fat negative), PG, Stbx (Soon-to-be-ex), Lupron, HCG, Infertility, Fertile
45-64	Syndrome, Death, Chronic, Diet, Anxiety, Hospital, HIV, Infect, Treatment, Flu	Menopause, Fibromyalgia, Oxycontin, Chantix, AA (alcoholic anonymous), RA (rheumatoid arthritis), Disc, Narcotic, Heat, Chronic	Menopause, Grandson, HIV, Disc, Tinnitus, Lesion, Liver, Cholesterol, Enzyme, Colon
65+	N/A	Diovan, Lisinopril, Neuropathic, Urine, Ankle, Cholesterol, Stroke, Arthritis, BP (blood pressure), Cancer	COPD (chronic obstructive pulmonary disease), Valium, PD (panic disorder), Caregiver, Retire, Oxygen, Transplant, Chemo, Cardiologist, Grandchildren

Table 3.6. Top 5 distinctive disorders by age.

Age, years	Drugs	Forums
0-17	Acne vulgaris, Acne, Attention deficit hyperactivity disorder, Mood swings, Feeling suicidal (finding)	Depressed mood, Incised wound, Mental depression, Fear (finding), Emotional distress
18-34	Endometriosis, site unspecified, Gravidity, Panic attacks, Anxiety attack, Manic	Gastritis, Asthma, Panic, Anxiety disorders, Observation of attack
35-44	Endometriosis, site unspecified, Manic, Manic mood, Addictive behavior, Chronic pain	Autistic disorder; Disability; Lupus erythematosus, systemic; Attention deficit hyperactivity disorder; Pressure (finding)
45-64	Hot flushes, Chronic pain, Fibromyalgia, Night sweats, Nerve pain	Codependency; Gastritis; Fibromyalgia; Lupus vulgaris; Lupus erythematosus, systemic
65+	Muscle cramps in leg, Dry cough, Lassitude, Diabetic, Blood pressure finding	Atrial fibrillation, Diabetic, Panic attacks, Cerebrovascular accident, Gastroesophageal reflux disease

In Table 3.7, we summarize all age groups' top drugs. For the younger group of 0-17 years in drug review websites, top drugs discussed are the ones related to ADHD. Age group of 18-34 years in drug review websites discusses pregnancy-related drugs, whereas for age group of 35-44 years, the top drugs are related to MS disorder. This group of 35-44 years in health Web forums tends to share information about ADHD drugs. Older age users (65+ years) discuss drugs related to heart problems, blood pressure, diabetes, and cholesterol.

Table 3.7. Top 5 distinctive drugs by age.

Age, years	Drugs[a]	Forums[a]
0-17	Accutane, Concerta[1], Vyvanse[1], Strattera[1], Implanon[3]	Commit Lozenge, Relate—vinyl resin, Vent, Zoloft[4], Topamax
18-34	Clomid[3], Phentermine, Seasonique[3], Lupron[2], Yaz[3]	Human papilloma virus vaccine, Topamax, Diamox, Adderall[1], Antibiotics
35-44	Clomid[3], Rebif[5], Avonex[5], Tysabri[5], Lortab	Concerta[1], Melatonin, Diamox, Plaquenil, Adderall[1]
45-64	Tamoxifen[2], Avonex[5], Oxycontin, Savella, Soma	Smoke, Hydrocortisone, Cymbalta[4], Lyrica[7], Alcohols
65+	Plavix[6], Diovan[6], Actos[7], Hydroxymethylglutaryl-CoA reductase inhibitors, Lipitor[6]	Metformin[7], Carbohydrates, Oxygen, Sugars, Xanax[4]

[a]Some of the drugs are coded to match the corresponding disorders they treat: [1]ADHD, [2]Cancer, [3]pregnancy, [4]depression, [5]MS, [6]BP, heart problem and cholesterol, [7]Diabetes.

Sentiment and emotion were evaluated for all sources. Because the results look similar among age groups, we summarize the results in Tables E.3 and E.4 of Appendix E. One key finding from the emotion results is that older people in Google+Health and drug review websites express less anger, whereas younger people in drug review websites express more anger.

3.4.3 Ethnicity

Only TwitterHealth and Google+Health have a large enough number of users whose ethnicity we can estimate (see Table A.2 in Appendix A), and hence, we only report finding for these outlets. In Table 3.8, we summarize top disorders for each ethnicity except black owing to the small number of users. As a key finding of top disorders, fibromyalgia is one of the top disorders that white and Hispanic users discuss in TwitterHealth, heart and kidney diseases are discussed more by Asian users, and headache and sleeplessness are 2 of the top disorders discussed by Hispanic users. The other ethnicity-based results exhibit less variance among the ethnicity groups, and hence, we report in Tables E.5, E.6, E.7, and E.8 of Appendix E.

Table 3.8. Top 5 distinctive disorders by ethnicity.

Ethnicity	TwitterHealth	Google+Health
White	Fibromyalgia, Presenile dementia, Leukemia, Migraine disorders, Mental disorders	Binge eating disorder, Diabetic neuropathies, Marijuana abuse, Neuropathy, Crohn disease
Asian	Heart diseases, Food poisoning, Obesity, Herpes NOS, Stress	Kidney diseases, Myopia, Fatigue, Hemorrhage, Hypersensitivity
Hispanic	Headache, Fibromyalgia, Sleeplessness, Mental depression, Insomnia adverse event	Herpes zoster disease, Diarrhea, Suicide, Lupus vulgaris, Osteoporosis

3.4.4 Location

Table 3.9 summarizes the top disorder results for all sources. Focusing on drugs and forums, which have been shown to have more useful information regarding one's health [8], our key finding is that users in the Northeast talk more about traditional physical disorders including diabetes and heart conditions, users in the Midwest discuss about weight loss, users in the South about fibromyalgia, and users in the West discuss mental disorders and addictive behaviors.

Table 3.9. Top 5 distinctive disorders by location.

Location	TwitterHealth	Google+Health	Drugs	Forums
Northeast	N/A	Inflammatory bowel diseases, Crohn disease, Occupant of van injured in transport accident, Kidney diseases, Prostate carcinoma	Asleep, Seizures, Patient outcome—died, Memory observations, Fatigue	Diabetes, Atrial fibrillation, Gastroesophageal reflux disease, ACHE, Lupus vulgaris
Midwest	Migraine disorders, Primary malignant neoplasm	Confusion, Marijuana abuse, Van der Woude syndrome, Injury wounds, Cataract	Hemorrhage, Body weight decreased, Hot flushes, Weight loss adverse event, Xerostomia	Asthma, Migraine disorders, Pressure (finding), Autistic disorder, Cerebrovascular accident

South	N/A	Diabetic neuropathies, Binge eating disorder, Neuropathy, Alzheimer's disease pathway KEGG, Diarrhea	Fibromyalgia, Drowsiness, Edema, Pruritus, Manic	Codependency, Shot (injury)
West	Presenile dementia, Heart diseases, Mental suffering, Herpes NOS, Obesity	Sexually transmitted diseases, Bipolar disorder, Myocardial infarction, Vitality, Acquired immunodeficiency syndrome	Post-traumatic stress disorder, Sleeplessness, Anxiety attack, Addictive behavior, Suicidal	Marijuana abuse, Addictive behavior, codependency, Autistic disorder, Lupus erythematosus, systemic

The other location-based results including sentiment, emotions, top distinctive terms, and top distinctive drugs exhibit less variance among the location groups, and hence, we report them in Tables E.9, E.10, E.11, and E.12 of Appendix E, as the variations across locations are not significant.

3.4.5 Writing Level

Table 3.10 and 3.11 summarizes the emotion results for all sources. For shortness, only 3 emotions are listed here: anger, trust, and anticipation, as the other 3 (fear, disgust, and surprise), are complementary to these, respectively. We see that users with lower writing level express more anger, with the exception of drug review websites, whereas

people with higher writing level express less anger. Due to low variance among writing levels, the other results for writing level including sentiment, top distinctive terms, top distinctive disorders, and top distinctive drugs can be found in Tables E.13, E.14, E.15, and E.16 of Appendix E, respectively.

Table 3.10. Emotion for each demographic grouped by source for TwitterHealth and Google+Health.

Writing level	**TwitterHealth**			**Google+Health**		
	Anger (%)	Trust (%)	Anticipation (%)	Anger (%)	Trust (%)	Anticipation (%)
0-5	N/A	N/A	N/A	**38.2**[a]	68.5	71.2[a]
6-9	41.0[a]	**44.3**[a]	75.3[a]	34.6[a]	67.9	75.8[a]
10-16	**34.1**[a]	**55.2**[a]	81.9[a]	**31.0**[a]	66.6	79.1[a]

[a]Represents the values with high significance ($P \leq .05$) compared with the union of the other age groups.

Table 3.11. Emotion for each demographic grouped by source for Drugs and Forums

Writing level	**Drug**			**Forums**		
	Anger (%)	Trust (%)	Anticipation (%)	Anger (%)	Trust (%)	Anticipation (%)
0-5	31.6[a]	66.8[a]	72.9[a]	**34.3**[a]	78.1[a]	71.5[a]
6-9	31.4[a]	67.7[a]	73.2[a]	31.7[a]	77.8[a]	72.6[a]
10-16	29.9[a]	73.1[a]	72.9[a]	**27.4**[a]	77.2[a]	75.4[a]

[a]Represents the values with high significance ($P \leq .05$) compared with the union of the other age groups.

3.5 Discussion

3.5.1 Notable Results

Our results provide valuable information that can help reach the right demographic group for each health condition. For example, to reach young users (aged 0-17 years) with ADHD, one should go to drug review websites. This finding can be a result of the increased percentage of children with ADHD recently (9%), compared with 2000 when it was 7% [81]. Similarly, to reach users of age group 18-34 with sleep disorder, one should go to Google+. We also found that the age group of 35-44 years discusses drugs associated with MS disorder, which agrees with the average age of clinical onset of MS, which is 30-33 years, and the average age of diagnosis, which is 37 years [82]. Because older age groups as our results show tend to discuss chronic diseases such as diabetes, heart problems, and cholesterol, health professionals and educators can target these groups in drug review websites and health Web forums to increase national awareness and decrease disease-related deaths.

Furthermore, a surprising finding is that, despite the fact that women suffer from back pain more than men [83], our results show that men discuss back pain more than women. Because 76% of all adults who have HIV are men [84], our results support this fact where HIV is one of the top discussed topics in TwitterHealth.

Our results also found that users in Western states discuss mental disorders and addictive behaviors including alcohol and marijuana as Table E.11 of Appendix E shows. This finding is associated with the fact that 5 of top 10 states with high marijuana use are in the West area [85]. Midwest users discuss weight loss more than the other regions

according to our results, which can be related to the fact that the Midwest is the second (slightly trailing the South) highest region in terms of obesity, with more than 25% of the adults being obese (body max index of 30+) [86].

3.5.2 Applications

There are several ways to leverage our results. Our findings can help health care providers and public health officials create targeted and effective educational campaigns, guide advertisers for different topics discussed by different demographic groups, help funding agencies allocate their research funds to have a larger impact on the society's top health issues, and help understand health disparities in Web-based health social media.

For instance, to reach pregnant or trying-to-get-pregnant women, advertisers should go to health Web forums and drug review websites instead of Google+ for advertising related products. This finding is supported by the fact that drug review websites and health Web forums are dominated by female users. Also, this finding may indicate that there is a need for more definitive and authoritative sources of such information.

Our results can also help understand health disparities in Web-based health social media. Users with higher writing level are less angry when discussing health-related issues, which may be linked to the fact that people with lower level of education receive lower quality of health care [87] and have higher mortality rate [69].

These demographics-specific findings can be used in targeted educational campaigns, which are recently becoming the focus of several research efforts. As an example, Whittaker [88] shows how a smoking cessation intervention using mobile

phones for young adults can be effective by sending general health videos messages and setting a quit date. Furthermore, Opel et al [89] show how social marketing can be used to increase immunization rates, where they explained how social marketing techniques can capture attention and motivate the targeted population to change. Patel et al [90] performed a systemic review to evaluate the effect of applications of contemporary social media on clinical outcomes in chronic disease. The study shows that providing social, emotional, and experiential support in current social media can help improve the patient care. Valle et al [91] evaluated a Facebook-based intervention that aims to increase the physical activity of young adult cancer survivors, which shows a potential for increasing the physical activity compared with Facebook-based self-help. A review of health interventions in Web-based social networks is presented in the study by Maher et al [92] where it is shown that several studies included in the systematic review reported significant improvement in health behavior or outcomes.

3.5.3 Limitations

For the general social networks, Google+ and Twitter, we used 276 health keywords and phrases as we described in the Methods section to filter the posts. These keywords and phrases miss some consumer phrases or abbreviations, such as ivf (in vitro fertilization) and iui (Intrauterine insemination). Unfortunately, we must select a relatively small set of keywords, given the rate constraints of the APIs of the social media.

Owing to the fact that ethnicity was estimated using a classifier (Chapter 2), we were not able to confidently compute the ethnicity of enough users to have reliable results

for several cases. For that, we omit results for black users. Furthermore, we do not report ethnicity results for drug review websites and health Web forums because these sources do not provide users' last names. Another limitation is self-selection bias because all demographic attributes (explicitly reported or classified) are reported by users. For instance, a user may choose to report age or last name (which is used to classify ethnicity). For example, people who trust the opinion of other users or experts participate more in social networks, whereas people who have less trust might not share their private experiences.

For extracting medical concepts, we do not handle all abbreviations. We handle some cases through manual rules, for example, Metamap would map "I" to iodine. Also, MetaMap is not perfect for annotating social media posts; thus, we removed annotations that look incorrect as the previous example. Moreover, when computing the top distinctive terms, we do not handle variations of terms, that is, "iui" and "Intrauterine insemination" are considered different terms. We do a manual postprocessing to address this issue for the top results. In measuring the sentiment of posts, the sentiment lexicon "SentiWordNet" was not built specifically for social or medical text. For example, some words such as "omg" or "lol" are not mapped to any word in the lexicon; thus, not all terms in the posts are assigned a sentiment.

3.6 Conclusion

We analyzed the content of Web-based health social media based on users' demographics. Three different types of Web-based health social media were considered: social networks, drug review websites, and health Web forums. For each demographic

attribute—gender, age, ethnicity, location, and writing level—we evaluated sentiment and emotion, and we extracted top distinctive terms and medical concepts, specifically disorders and drugs. Our results are both expected and surprising and show several key findings for each demographic attribute. For example, the dominant topic for female users in drug review websites and health Web forums is pregnancy, whereas for male users, it is cardiac problems, HIV, and back pain. Attention-deficit hyperactivity disorder and depression-related drugs are the main topics discussed by younger users (0-17 years), MS drugs are discussed more by users of age 35-44 years, and alcohol and smoking are mainly discussed by middle-aged users (45-64 years). Users from the Northeast United States talk about physical disorders, whereas users from the West United States talk about mental disorders and addictive behaviors. Finally, users with higher writing level express less anger in their posts. These key findings can help experts reach the right users in many ways, including creating targeted and effective educational campaigns by health care providers, advertising related products, allocating funds for the right research by funding agencies, and understanding health disparities in Web-based health social media.

Chapter 4

Intent Classification of Health-Related Social Media

Background: The rising volume of web-based health social media activity, where users connect, collaborate, and engage, has increased the significance of analyzing how people use them.

Objective: The aim of this study is to classify the intent – e.g., share experiences, seek support – of users who participate in web-based health social media, and study the effect of the user demographics to the posting intent.

Methods: We analyzed two different types of health-related social media: (1) health Web forums including WebMD and DailyStrength; and (2) general Web-based social networks Twitter and Google+. We identified several post intents, and built classifiers to automatically detect the intent of posts. These classifiers where used to study the distribution of intents for various demographic groups.

Results: The results of this study are: (1) General social networks Twitter and Google+ are mostly used to share health-related news and educational material; (2) Half of the posts in WebMD and Dailystrength are sharing experiences, for both male and female users; (3) Male users ask for medical advice more often than female users in WebMD; and (4) Half of the posts in DailyStrength are about sharing experience, regardless of the age group or location.

Conclusion: We studied and analyzed the intent of users participating in health-related social media. Our results can guide health care providers and practitioners to create effective and targeted health care campaigns.

4.1 Introduction

The ongoing increase in the use of health-related social media, especially in health care contexts, has increased the importance of harvesting and analyzing its content. Health-related social media is mainly used to increase interactions with other, and sharing and retrieving health messages [63]. Users in different health-related social outlets share their information, family members, or friends to seek help for a wide range of health issues [63]. In the United States, more than 60 million Americans have read or collaborate in Health 2.0 applications [20]. In addition, 40% of Americans doubt a professional opinion when it conflicted with what they form from Web-based health social media [20]. Comparing to traditional communication methods, health-related social media widening access for health information for public regardless of their race, age, locality, and education [63].

In this work, we study the intent of posts in different health-related social media. We analyzed two types of health-related social media: (1) health Web forums, including WebMD and DailyStrength; and (2) general Web-based social networks, namely Google+ and Twitter. We randomly selected posts from each source, then manually identified and determined the intents based on the post content. For health Web forums, we identified four intents as follows: share experience, ask for advice, request/give

support, and talking about family. For general Web-based social networks, we identified five additional intents: share news, jokes, ads, personal opinion, and educational materials. We labeled the posts for each intent, and use supervised learning classifier to train the data if there were sufficient posts. The classifiers with greater accuracy were utilized to label the rest of our posts. We finally analyzed the demographic-based content when possible.

4.2 Related Work

Many conducted studies and research have been established to extract meaningful information from health-related social media, including demographics, diseases, drugs, and so forth. Hackworth and Kunz [21] reported that 80% of Americans have searched the Internet for health-related information. Sadilek et al. [93] studied the spread infectious diseases by analyzing Twitter data using SVM model. Wiley et al [13] studied the impact of different characteristics of various social media forums on content. Nikfarjam et al. [94] proposed a machine learning-based tagger to extract adverse drug reaction (ADR) from health-related social media. Eichstaedt et al. [95] predicted the count-level heart disease mortality by capturing the psychological characteristics of community through expressed text in Twitter. Chapter 2 and 3 analyzed the demographic of health-related social media, and also performed a demographic-based content analysis to extract top distinctive terms, top drugs and disorders, and sentiment and emotion. Krueger et al [69] studied the mortality attributable to low education level in the United States, where they found people with less than high school degree have more mortality rate. Mislove et al.

[24] estimated the gender and ethnicity of Twitter users using reported first name and last name. Anderson-Bill et al [72] examined demographics, behavioral, and psychosocial characteristics of recruited Web-health users.

4.3 Methods

4.3.1 Datasets

As shown in Table 4.1, for health Web forums, we selected 2 different websites, WebMD and DailyStrength. The reason for selecting two health Web forums is to cover different types of health Web forums, where WebMD is used to ask specific and medical related posts, while DailyStrength has broad topics including medical conditions and life challenges. For general social networks, we chose Twitter and Google+ for their popularity and number of users. More information about the sources including start and end date is available in Table A.1 and A.2 of Appendix A. We used 267 representative health-related keywords to filter Twitter and Google+ because they are general social networks as follows: (1) Drugs: from the most prescriptions dispensed from RxList.com, we selected the 200 most popular drugs [36]. By removing the variants of the same drug (eg, different milligram dosages), the final list of drugs contained 125 unique drug names. (2) Hashtags: from Twitter Hashtags, we selected 11 popular health-related Twitter hashtags such as #BCSM (Breast Cancer and Social Media). (3) Disorders: 81 popular disorders were selected such as AIDS and asthma. (4) Pharmaceuticals: the 12 largest pharmaceutical companies were selected such as Novartis. (5) Insurance: 44 of the biggest insurances were selected such as Aetna and Shield.

Table 4.1. List of all used sources with their number of posts and with the available demographic attributes.

Dataset	No. of posts	Gender[a]	Age[a]	Ethnicity[a]	Location[a]
TwitterHealth [27]	11,637,888	Gender classifier	NO	Ethnicity classifier	YES
Google+Health [28]	186,666	YES	YES		YES
DailyStrength [33]	5,948,877	YES	YES	NO	YES
WebMD [34]	1,128,629	Gender classifier	NO	NO	NO

[a]NO indicates that the demographic attribute is not provided by the source and no classifier is used due to low accuracy. YES indicates that the demographic attribute is provided by the source. More details on the demographic classifiers are available in the paper by Sadah et al [96].

To filter out Twitter using the health-related keyword list to retrieve the relevant tweets for TwitterHealth, we used Twitter streaming Application Program Interface (API) [37]. For Google+, we used Google+ API [38] to extract the relevant posts for Google+Health, by using health-related keyword list in the queries. For health Web forums, WebMD and DailyStrength, we built a crawler for each website in Java using jsoup [39], a library to extract and parse hypertext markup language content. For each website, we collected available data, including posts, user information, keywords, tags, and so forth.

4.3.2 Identifying Intents

From each source, we randomly selected 1000 posts, and we then manually identify the different intents of shared content for each type of health-related social media. As

shown in Table 4.2, we identified 9 different intents. The first 4 intents are identified for the two types of health-related social media, health Web forums, and general social networks. Because Twitter and Google+ are more news-based social media, we identified 5 more categories to cover the different categories from these sources.

Table 4.2. List of all identified intents.

	Intent	**Example**
Health Web Forums – General Social Networks	Share Experiences	"I could not work after Tylenol." Or "I have taking Lipitor every day."
	Ask for Specific Medical Advice or Information	"Is honey allowed for diabetics?"
	Request or Give Psychological Support	"I hope your diabetes is under control." or "We're thinking of you."
	About Family (Not About Self)	"My son is now nine months old and teething like crazy."
General Social Networks	Share News	"Kaiser Permanente Invites Software Developers To Build Apps - Forbes. http://feedly.com/k/Zojwq"
	Jokes	"Got any jokes about Sodium Hypobromite? NaBro."
	Advertisements	"Check out these two vitamins for one recipe! http://bit.ly/1471dbn"
	Personal Opinion	"Main frustration of lupus is losing the ability to do things that used to be normal"
	Educational Material	"Side Effects of Alzheimer's and Dementia Drugs http://bit.ly/cK7L1f"

4.3.3 Identifying Intents

We asked three graduate students to label the selected data, and we used the majority vote as the final results for each source. As shown in Table 4.3, the distribution of intents in each source is different, where share experiences category is more in health Web forums.

Table 4.3. Percentages of intents in each source from the labeled data.

Intent	WebMD	DailyStrength	TwitterHealth	Google+Health
Share Experiences	55%	59.4%	14.8%	13.5%
Ask for Specific Medical Advice or Information	48%	25.1%	0.6%	2.1%
Request or Give Psychological Support	15.8%	13.1%	1.8%	1%
About Family (Not About Self)	17.3%	10.9%	1%	4.4%
Share News	N/A	N/A	11.2%	32.5%
Jokes			7.06%	5%
Advertisement			5.2%	14.9%
Personal Opinion			7%	15.8%
Educational Material			7.2%	28.8%

To train each category classifier, we extracted 9 different features as shown in Table 4.4, where 4 of them are word vector, and the rest are numeric. The feature board name is extracted from the post URL, and for the positive and negative emoticons, we created a dictionary of positive and negative emotion icons and applied it on a post's title and body text to count the numbers. In order to get the word vector features, we used

StringToWordVector class filter from Weka machine-learning toolkit v. 3.8.1 [97], which filters strings into N-grams using WordTokenizer class, with the following settings: (1) convert all words to lower case, and (2) perform TF/IDF transformation. After getting all features, we used Weka's AttibuteSelection filter with evaluation method of Info Gain and search method of Ranker with threshold 0, to get the most important features (features with positive InfoGain) to train each classifier.

Table 4.4. All classifiers training features.

<table>
<tr><th></th><th>Word Vector</th><th>Numeric</th></tr>
<tr><td>WebMD</td><td rowspan="2">Title + body + board name</td><td rowspan="4">Number of question marks
Number of exclamation marks
Number of positive emoticons
Number of eegative emoticons
Number of URLs
Post Length</td></tr>
<tr><td>Dailystrength</td></tr>
<tr><td>Google+</td><td>Title + body</td></tr>
<tr><td>Twitter</td><td>body</td></tr>
</table>

To build the classifiers, we excluded the intents where the percentage is less than 10%, and for the rest, we split the labeled data to three datasets as follows: (1) training dataset (800 posts), (2) validation dataset (100 posts), and (3) test dataset (100 posts). To build the classifiers, we train our data using Random Forest classifier by varying three different parameters: maximum depth, number of trees, and number of features. For each parameter, we consider the range from 1 to 35, that is, there are 42,875 (=353) combinations. For all combinations, we train the model using the training dataset, and evaluate the accuracy on the validation dataset to select the combination with the highest accuracy. Then, we test the model on the test dataset using same parameters. Table 4.5 and 4.6 show the classifiers' accuracy for each source. In table 6, we show only the classifiers for categories that have more than 10% of labeled data.

Table 4.5. Classifiers accuracy for health Web forums.

Intent	WebMD		DailyStrength	
	Accuracy	**Weighted Accuracy**	**Accuracy**	**Weighted Accuracy**
Share Experiences	83%	82%	78%	76.7%
Ask for Specific Medical Advice or Information	89%	88.5%	84%	68%
Request or Give Psychological Support	86%	56.25%	79%	48.73%
About Family (Not About Self)	94%	87.24%	91%	63.13%

We only consider the classifiers that have weighted accuracy higher that 75%; therefore, we don't use any classifier from TwitterHealth and Google+Health to further analyze the demographics. For WebMD and DailyStrength, we used the classifiers with higher accuracy to label the rest of data in both sources, and we report the results in the Results section.

Table 4.6. Classifiers accuracy for general social networks.

Intent	TwitterHealth		Google+Health	
	Accuracy	**Weighted Accuracy**	**Accuracy**	**Weighted Accuracy**
Share Experiences	93%	68%	89%	72.67%
Share News	N/A		73%	64.28%
Advertisement	N/A		76%	55.65%
Personal Opinion	N/A		82%	51.13%
Educational Material	N/A		77%	65.48%

4.4 Results

In this section, we present the intents results by each demographic when possible. As shown in Table 4.5 and Table 4.6, we only use select 4 intent classifiers: (1) for WebMD we have the following intent classifiers: share experience, ask for specific medical advice, and about family, and (2) for DailyStrength we have share experience intent classifier. The weighted accuracy for both TwitterHealth and Google+Health classifiers are not satisfying, however, it's worth noting from the labeled data that these two sources are popular for sharing news and educational material, which is different from health Web forums.

4.4.1 WebMD

As shown in Table 4.1, WebMD has a gender that was predicted by the gender classifier in from Chapter 2. Therefore, we report here the distribution of the three selected classifiers from Table 4.5: share experiences, ask for specific medical advice or information, and about family (not about self). Figure 4.1 shows the percentages of posts shared by male and female for each category, where almost half of the posts shared by male and female are about share experiences.

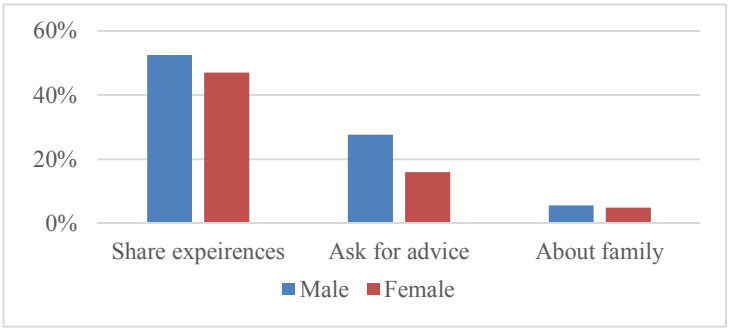

Figure 4.1. Share experiences, ask for specific medical advice or information, and about family categories distribution by gender

4.4.2 DailtStrength

There are three demographic attributes reported by DailyStrength as shown in Table 1: gender, age, and location. For these demographics attributes, we report the results for the category "share experience", as it's the only classifier with accuracy higher than 75%. Figure 2 shows the category distribution by the different demographic attributes, where mostly half of the posts shared by each group are classified as sharing experiences.

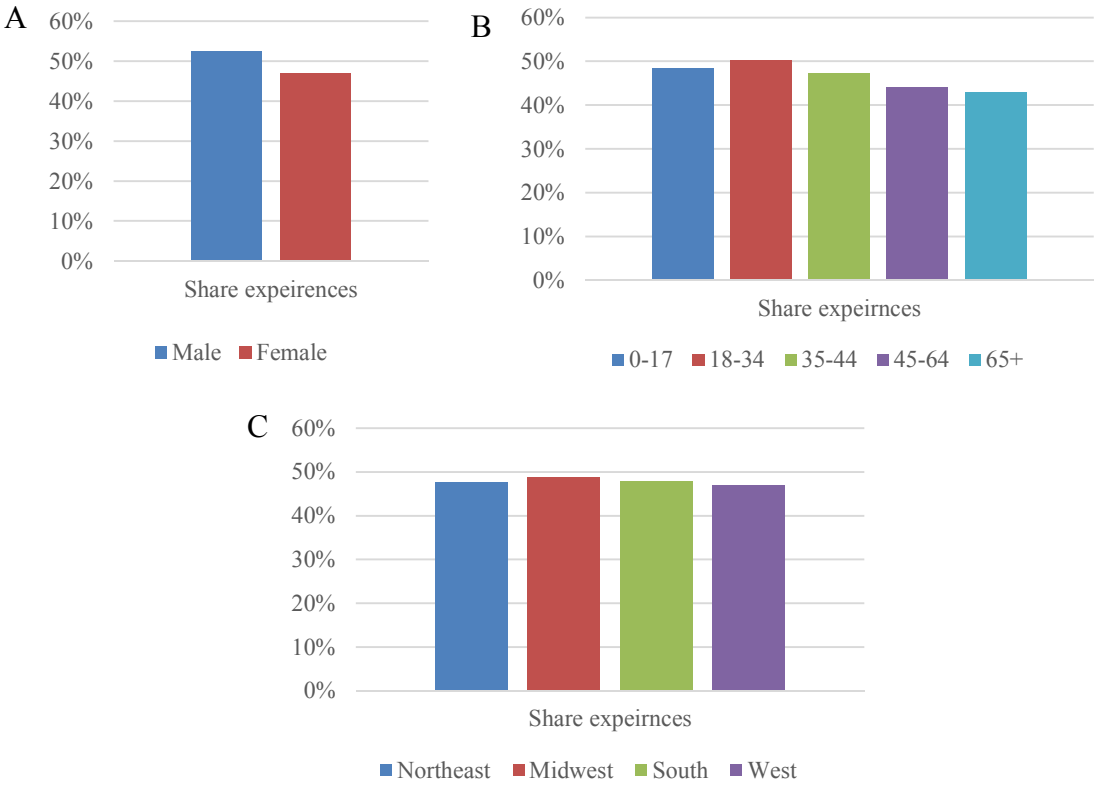

Figure 4.2. Share experiences category distribution by (A) gender, (B) Age, and (C) Location

4.5 Discussion

Our results provide useful information that can help health care providers to reach the right demographic group. For example, researchers can use health Web forums when looking for clinical trials, where nearly half of posts are sharing experiences. Moreover, demographic-specific results can help guide the targeted educational campaigns. As an example, male users in WebMD ask more specific medical advice questions than female. The two types of health-related social media are using differently, which reinforces the fact that Twitter and Google+ are more news based social media.

4.5.1 Limitations

As users of health-related social media use informal writing style, our selected 276 words to filter Twitter and Google+ as describe in the Methods section may not cover all health-related posts. For example, the abbreviation “iui” (Intrauterine insemination), which is widely used in health-related posts but not included in the health-related keyword list. Another limitation is the different uses of terms used to filter Twitter and Google+. For example, word “cancer” yield many tweets that talk about zodiac signs.

4.6 Conclusion

In this study, we analyzed the intent of content shared in two different types of health-related social media: health Web forums and general social networks. For the two type of health-related social media, we manually identified four posts categories including: share experiences, ask for specific medical advice, request or give psychological support, and about family, and we additionally identified five categories

for general social networks including: share news, jokes, advertisements, personal opinion, and educational material. After labeling a randomly selected data for each source, we built classifiers for each category. We finally made a demographic-based content analysis when possible.

Chapter 5

Mining for Online Health-Related Effects Associated with Electronic Cigarettes

Background: Previous online infodemiological studies reported negative health-effects associated with electronic cigarette (EC) use. Automated methods can be valuable to retrieve health data for further analysis.

Objective: Our purpose was to extract and collate a large collection of online forum posts related to health-effects associated with EC use between 2008-2015. We mined a major EC online forum that contained various sub-forums, and focused on the posts in a sub-forum containing health effect data.

Methods: Data were annotated with a set of medical concepts from the Unified Medical Language System (UMLS), using a modified MetaMap tool. Over 1 million posts were collected on potential symptoms and disorders associated and/or affected by electronic cigarette use.

Results: Health effects data were grouped into symptoms or disorder data and were categorized into 12 organ systems/anatomical regions. Overall, most posts for symptoms and disorder data contained negative sentiment (%) across 7 years. Effects were most often reported in the neurological, mouth and throat, and respiratory systems for both symptoms and disorders. Additionally, users often reported paired symptoms of health effects within these categories (i.e. coughing and headache).

Conclusion: This study provides additional data on the short-term health effects reported by EC users over a period of time. Online forums provide a unique repository of data that can be useful for tracking health sentiment, understanding adverse effects and identifying potential symptoms associated with EC use.

5.1 Introduction

Since their introduction nearly a decade ago, electronic cigarettes (EC) have gained worldwide popularly without prior knowledge of their effects on human health. EC production is not currently well-regulated, and quality control during manufacture has frequently been questioned [98]–[102]. Therefore, a wide range of research concerning the health effects associated with EC use has been conducted in the past several years. These studies include online informatics and survey studies [103], [104], short-term physiological assessments of EC use on human health [105], and in vitro and in vivo cytotoxicity studies [106]–[112].

The Internet has become a useful source of information that can be mined for data dealing with human health. Online health forums in particular are a useful repository for human health data that can be mined to understand health effects users may experience and effects that may be underreported in the peer reviewed literature. This infodemiological approach has yielded new information EC topography and the on the effects of EC use on health [113]–[115]. Previous studies that mined Internet data on EC puffing topography showed that puff duration is about twice as long in EC users as in conventional smokers [116], [117]. In addition, topography is highly variable among EC

users who generally intake much larger volumes of aerosol than smoker's intake cigarette smoke [117]. Our prior study on online heath forums mined information manually from major EC websites and identified a number of negative and some positive health effects that users attributed to EC use [118].

EC users can post data relevant to health symptoms they experience with EC usage on various online EC forums. Some of these forums have been in existence since the introduction of EC and therefore contain health related data spanning more than 7 years. In this study, we followed-up on a previous infodemiological study in which we acquired and analyzed self-reported health effects posted on the Internet forums by EC users. We used enhanced automated measures to collect a large dataset from which we mined and sorted the various symptoms and disorders that EC users associated with EC. The data collected by using informatics tools agree with our previous study that wide-ranging symptoms can accompany EC use, and enabled us to track sentiment (positive, negative, neutral) of EC use across multiple years. These data support the idea that EC use is not free of adverse health effects and that it is important to continue tracking a range of symptoms that are often reported by users in the forums.

5.2 Methods

In this section, we describe our collected data, and the methods we used to extract the medical concept and measure the sentiment.

5.2.1 Datasets

We collected our data from a large e-cigarette discussion Web forum between January 2008 and July 2015. We analyzed the layout of the website and built a crawler in Java using the Java library jsoup [39], which is designed to extract and parse information from HTML pages. The posts were collected from seven sub-forums. The total number of discussion threads is 44,222, and the total number of posts is 1,450,896. However, since the primary goal of this paper is to study the short-term health effects produced by e-cigarette, we only focus on posts that belong to the health sub-forum, which has 2,330 discussion threads and 41,216 posts. We emphasize that all collected data are publicly available, including discussion threads and users' information.

5.2.2 Medical Concepts

We used a modified version of the MetaMap tool [119] to annotate each post with a set of medical concepts from the Unified Medical Language System (UMLS). The UMLS [120] is a repository of a large number of biomedical controlled vocabularies. In UMLS, there are 15 high-level semantic groups, which were created to help reduce the complexity by grouping the semantic types [121]. In this work, we analyze two semantic types, "disorder or syndrome" and "sign or symptoms", which belong to the "Disorders" semantic group. Each concept in UMLS can be assigned to multiple semantic types, but only to one semantic group [121]. Since MetaMap was built to annotate the natural language text in biomedical academic publications, it is not very effective out-of-the-box on social media posts, as it successfully maps the medical terms most of the time, but not the descriptive or non-medical terms [80]. To improve the tool's mapping efficiency, we

manually examined and removed misclassified UMLS concepts generated by MetaMap by performing the following steps:

1. For the two semantic types we analyze, disorders and symptoms, we order the concepts by their frequencies.
2. We analyzed the different terms mapped to each concept.
3. We removed the misclassified concepts from our results. Examples of misclassified concepts include:
    a. “mod”, which refers to vape mods, was mapped to “Type 2 diabetes mellitus” (C0011860)
    b. “ect”, which is a type of vape mods, was mapped to “Benign Rolandic epilepsy” (C2363129)
    c. “pic” was mapped to “Punctate inner choroidopathy” (C0730321)

For each semantic type, we reported the most frequent disorders and symptoms by year and overall.

5.2.3 Sentiment

To measure the positive and negative health effects produced by e-cigarette use, we used a supervised learning classifier (Random Forest) on a set of manually labeled posts to predict the sentiment for unseen posts. We randomly selected 1080 posts, and asked three labelers to categorize them as following:

- Negative: if a post clearly contains a health effect or unpleasant experience or complaint that co-occurred with the use of e-cigarette.

- Positive: if a post clearly mentions a health improvement or a recovery from previous health effects when switching from smoking analogs to e-cigarettes, or a good experience with e-cigarettes including products.
- Neutral: if a post doesn't express any sentiment.

Our interpretation of positive and negative is different from typical sentiment classifications, and mainly focuses on health-related effects. We first asked the labelers to categorize 400 posts, and then we measured the intercoder reliability between the labelers. Using "ReCal" [122], an online tool to calculate the reliability for the masses, the agreement was 80.53% using the "Average Pairwise Percent Agreement" measure. Due to the high agreement, the rest of the posts were split evenly among the labelers to categorize. Table 5.1 shows the class distribution of our sample data with examples for each class; 44.81% of posts were labeled as negative, 38.51% as neutral, and 16.67% as positive.

Using Weka machine-learning toolkit v. 3.8.1 [97], we first filter our sample data after many experiments using StringToWordVector class filter, which filters strings into N-grams using WordTokenizer class, with the following settings: (1) convert all words to lower case, (2) remove stop words, (3) stem words using Weka built-in stemmer, (4) keep only terms that appear at least twice, and (5) retain unigram, bigram, and trigram. We then split the sample data as following: (1) 962 posts for the training test, and (2) 118 posts for the test set. After that, we train our data using Random Forest classifier; however, the classifier's initial accuracy was not satisfactory.

Table 5.1. Sample data summary.

Class	No. of Posts	Example
Positive	180	"Welcome to ECF! I've only been vaping for 2 1/2 weeks, but I've already noticed a big difference in my lungs (after 20+ years of smoking). For example, I had a chest cold when I started, and in the past, once a cold moved into my chest it took a couple of months to get rid of it. … E-cigs are pretty darn amazing, IMHO."
Neutral	416	"I dont think there are any tests since flavoring were not meant to be inhaled. I think we are taking our chances untill some evidence comes out… "
Negative	484	"Hi Everyone, I have been using e-cigarrette for the past 2 months and very dissappointed that I have to stop, reason being my teeth, gums are sensitive and my tooth cracked yesterday, I have to have a crown fitted.8-o. I think that the nicotine is seriously not good for the mouth. My husband and work collegue have also reported sore gums, little sores in the mouth. …"

To improve the classifier's accuracy, we need to address a well-known issue in our sample data, which is the imbalanced class distribution [123]. The Positive class as seen in Table 5.1 only covers 16.67% of the data, while the Neutral class covers 38.52%s and the Negative class covers 44.81%. Thus, we oversampled the Positive class by duplicating the posts which they were labeled Positive in the training set only. Table 5.2 shows the new class distribution for the training set, namely Training (extended). Another approach we used to improve the accuracy is annotating all the posts in the sample data

with the ancestors of the medical concepts mentioned in the posts. For example, if "pneumonia" is mentioned in a post, then the post will be appended with "Disorder of lung".

Table 5.2. Training data summary.

Class	Training	Training (extended)
Positive	16.53%	28.37%
Neutral	39.19%	33.63%
Negative	44.28%	38.00%

After using the new training set, the classifier's accuracy increased from 66.95% to 75.42%. Table 5.3 reports for each class three different measures, including precision, recall, and F-measure. As seen in the table, the classifier is most accurate on Negative class (F-measure=0.79), followed by Positive and Neutral classes.

Table 5.3. Test data classification accuracy.

Class	Precision	Recall	F-measure	N
Positive	0.73	0.72	0.74	21
Neutral	0.67	0.77	0.71	39
Negative	0.84	0.74	0.79	58
Average	0.76	0.75	0.76	118

5.2.4 Data Categorization and Analysis

After data were collected iteratively they were sorted on MS Excel spreadsheets. All health-related effects (symptoms and disorders) reported and/or associated by EC users in posts, were grouped according to the organ system/anatomical region, which we

defined as “systems”, as previously described by [103]. When a symptom could have been associated with more than one system, the effect was assigned to the system for which it had the strongest fit (eg, improved sense of taste was assigned to sensory but could have been mouth/throat). Frequency distributions for the overall grouped data in each system for symptoms and disorders were plotted using GraphPad Prism. Additionally, the sentiment for each post were grouped and sorted for each post across seven years (2008-2015). We sorted them according to their positive, neutral and negative sentiment as described.

5.3 Results

5.3.1 Overall Frequency of Reported Symptoms and Disorders

To analyze the frequency of reports for various symptoms and disorders, the data in Figures 5.1-5.4 were condensed by linking all health effects into structural or physiological systems (e.g., sore throat was classified into mouth and throat). The five systems that contained the most reports for symptoms were: neurological (N=3623), respiratory (N=1995), digestive (N=1637), mouth and throat (N=1390), and integumentary (N=853) (Figure 5.1). The top five systems for disorders were respiratory (N=2972), mouth and throat (N=1986), neurological (N=1143), integumentary (N=1123) and immune (N=739) (Figure 5.2). For both symptoms and disorders, a majority of the posts were associated with negative sentiment across all systems (Figure 5.1 & 5.2).

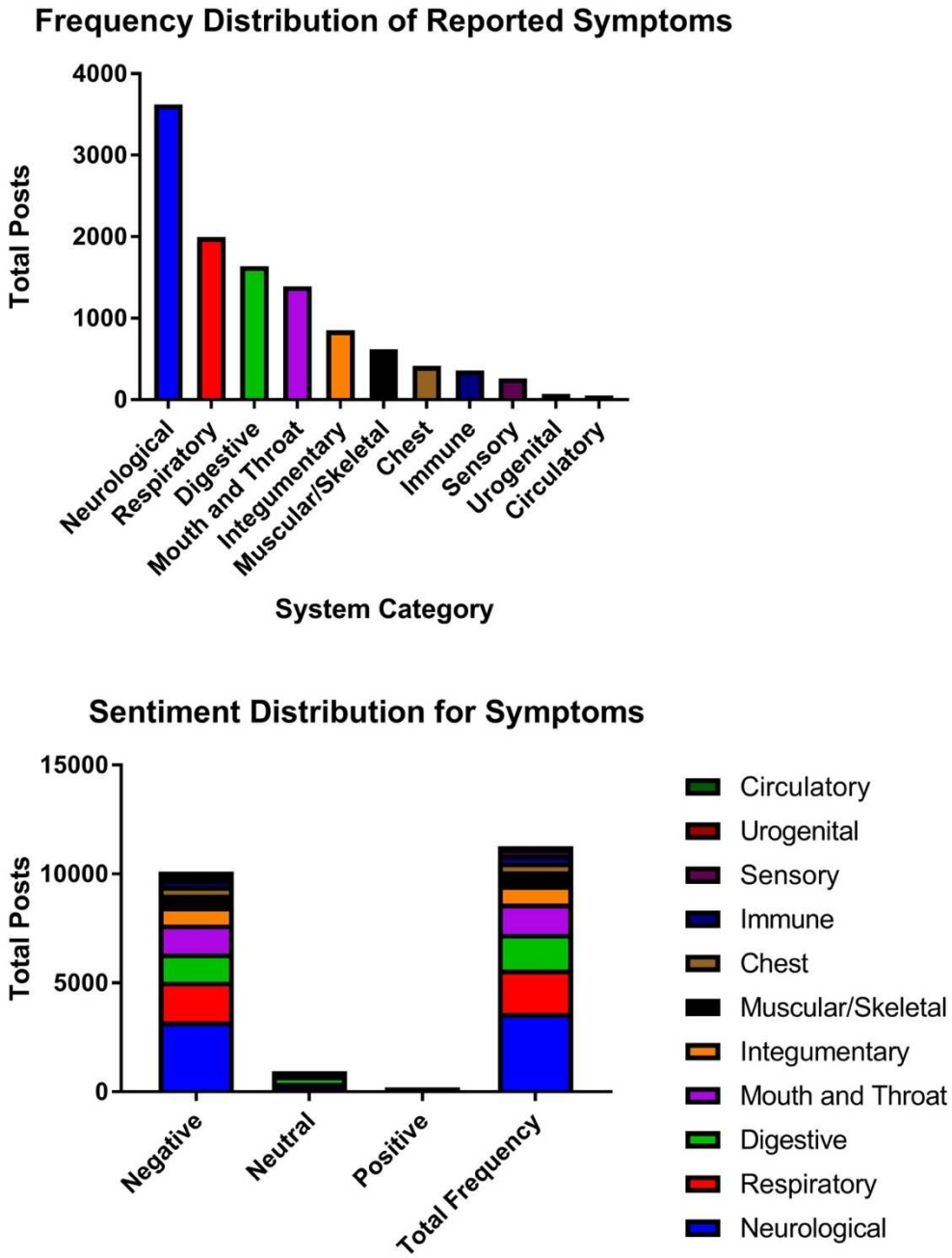

Figure 5.1: Frequency distribution of reported symptom posts grouped into their related systems or anatomical regions (above). The sentiment distribution (positive, neutral, and negative) for each category is shown, along with total frequency of posts (below).

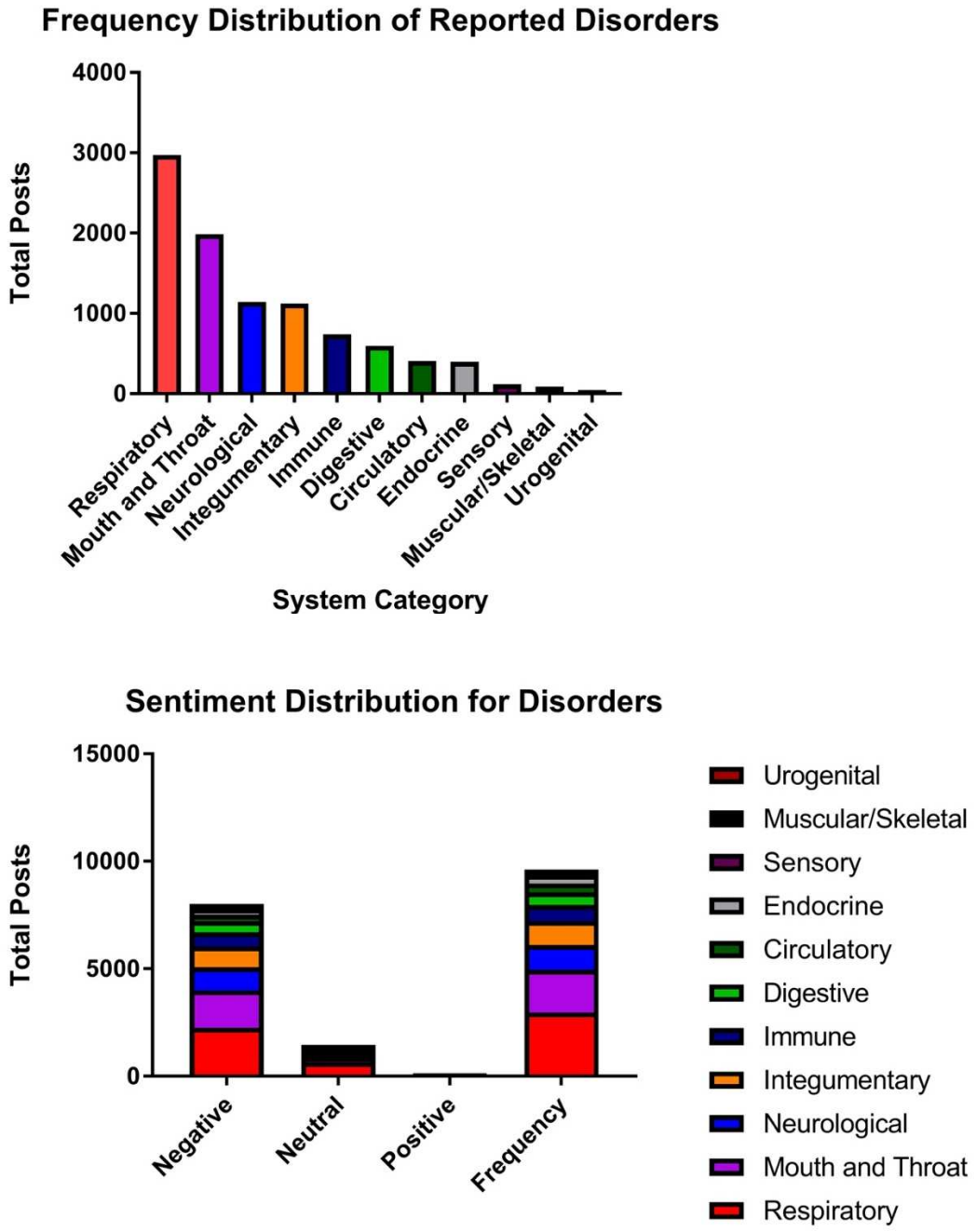

Figure 5.2: Frequency distribution associated with reported disorder posts are grouped into their related systems or anatomical regions (above). The sentiment distribution (positive, neutral, and negative) for each category is shown, along with total frequency of posts (below).

5.3.2 Symptom and Disorder Frequency and Sentiment Distribution Over Time

Data were collected for posts between 2008 to 2015. The posts for symptoms and disorders were grouped according to their years for analysis. Across all years for both symptoms and disorders, we found the frequency distribution of reports per year. Additionally, the posts for symptoms and disorders were categorized and their frequency per year were summarized in a stacked bar graphs for each year. Mostly negative sentiment was associated with the posts in each system or anatomical region. Typically for both the top five categories mentioned above, the results remained consistent over time (Figure 5.3 & 5.4).

For the symptoms (Figure 5.3), the posts with the most reports were consistently neurological, respiratory, digestive, integumentary, and mouth and throat. For all years except 2008, the neurological and respiratory systems were the top two systems. The other three systems (digestive, integumentary, and mouth and throat) alternated in some years but were generally in the top five systems with the most posts.

Similarly, the posts containing disorders associated with EC use (Figure 5.4), had consistent results for their top five system categories with the most posts. For the respiratory and mouth/throat were the top two systems reported in between 2008 through 2012. Alternating in the top five disorders were the integumentary, neurological, and muscular/skeletal systems.

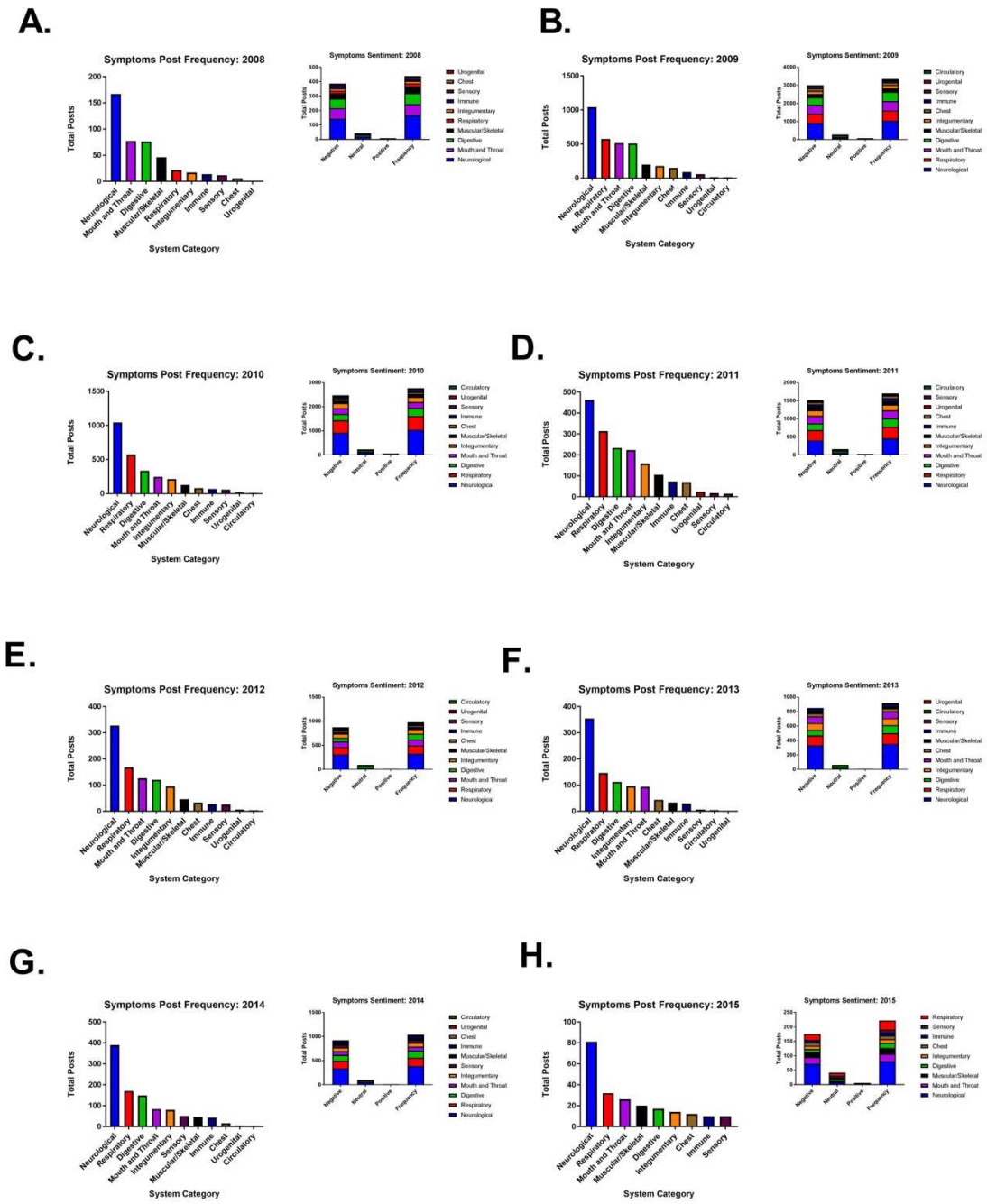

Figure 5.3: (A-H) Breakdown of frequency distribution of reported symptom posts from 2008 to 2015 grouped into their related systems or anatomical regions, along with sentiment distribution for each category (positive, neutral, and negative).

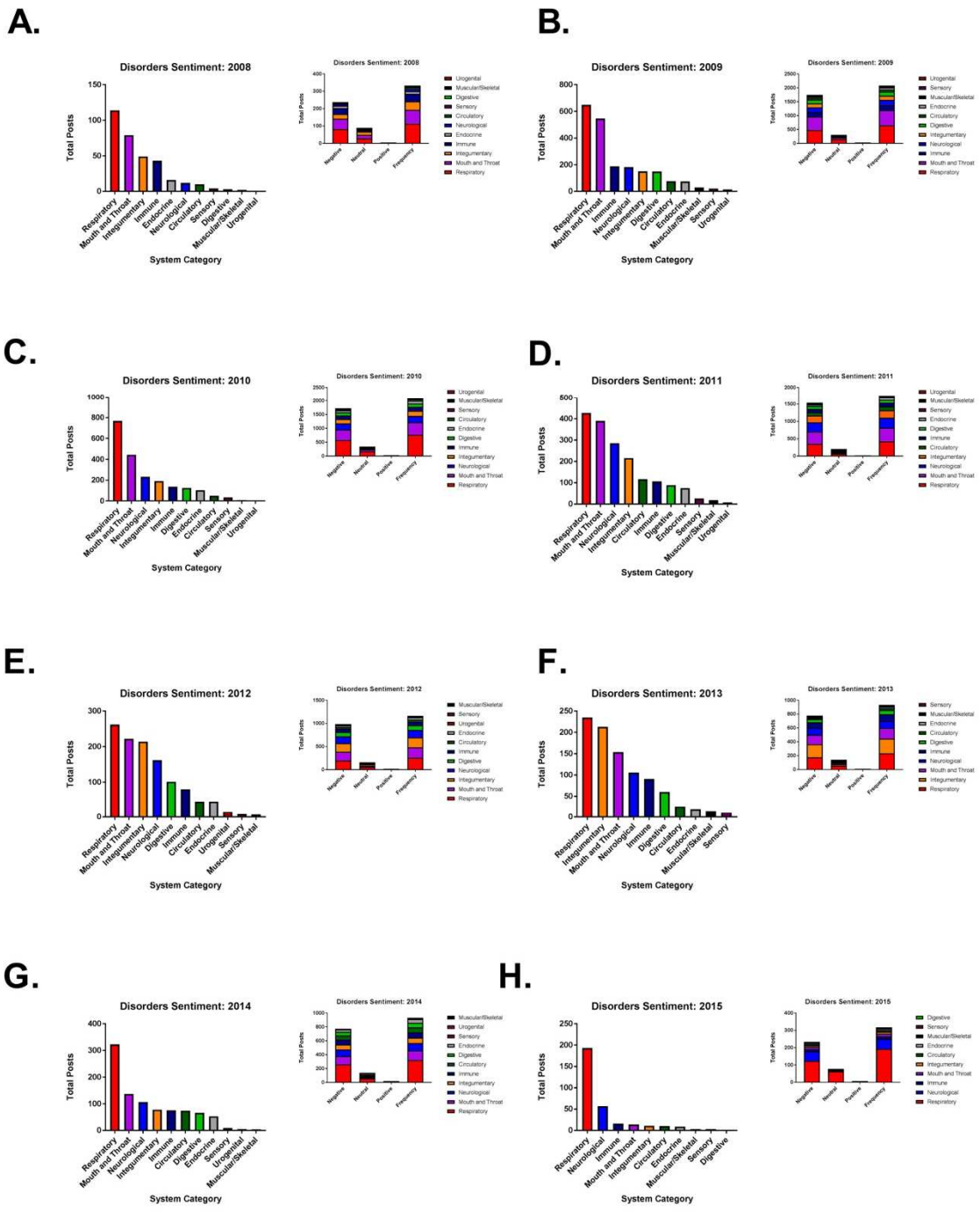

Figure 5.4: (A-H) Breakdown of frequency distribution associated with reported disorder posts from 2008 to 2015 grouped into their related systems or anatomical regions, along with sentiment distribution for each category (positive, neutral, and negative).

There was some increased reported of positive effects that can be noted in 2015 for the disorders (Figure 5.4). It should also be noted however, that we only have partial reporting for the year of 2015, so the results for both symptoms and disorders may not have captured all posts for these respective categories.

5.3.3 Identification of Top Reported Symptoms in Systems with Most Reports

Heatmaps were generated to view the specific symptoms for the top five systems that had the most reported health effects (Figure 5.5 & 5.6). Because most symptoms were associated with negative sentiment, we chose to visualize the symptoms associated with negative reporting.

For analysis, the total post number was converted to a log scale value to better visualize the frequency of posts. A red-gray-blue color scheme was used to show infrequent post counts (blue) to those with the highest frequency (red). Those symptoms that are white in the heat maps were associated with neutral or positive posts.

Numerous symptoms were reported for each system, and the frequency of reports varied. Typically, more than half of symptoms had few/trace reporting. Approximately a quarter of symptoms for the top five categories had mid-range post frequency between 10-100 posts (gray color scale). In the neurological system the most common symptoms reported were: headaches (N=939), fatigue/tired/malaise (N=468), nausea (N=290), dizziness (N=183) and lightheadedness (N=113) (Figure 5.5). In the respiratory system the most negative effects were reported for coughing (N=852), wheezing (N=298), dyspnea (N=235), and excessively deep breathing (N=112).

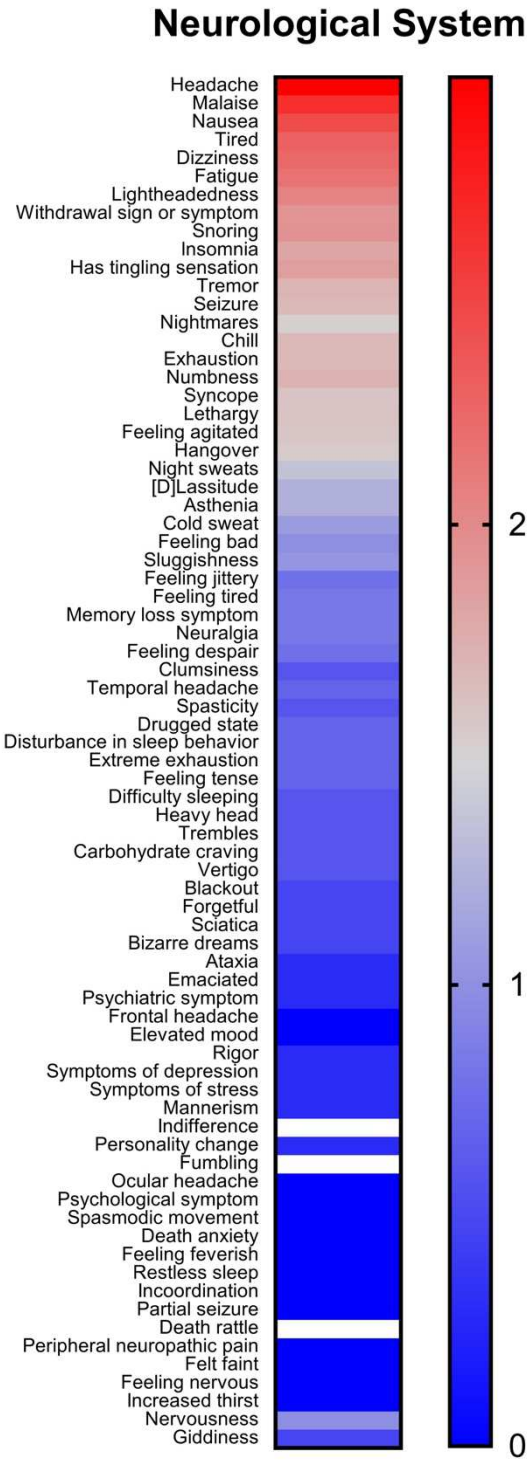

Figure 5.5: Heatmap of all symptoms reported in the neurological system. Post count were converted to log scale with from greatest (red) to least (blue).

The most reported symptoms in the digestive system were: heartburn (N=327), flatus (N=303), cramping (N=176) and constipation (N=113). In the mouth/throat and integumentary systems pain in throat (N=643), harsh voice quality (N=175), pharyngeal dryness (N=147), itching skin (N=565), and dry skin (N=121) were commonly reported (Figure 5.6).

5.4 Discussion

The short and long term health effects associated with EC is an important public health issue for product users, health professionals, and regulatory agencies. To better understand health effects associated with EC use, we have taken a combined infodemiological and informatics approach to automatically mine information from a large pool of data collected from an online health forum. Our data, which are based on over 1 million online forum posts, showed that a variety of symptoms and disorders were found in posts linked to EC. Additionally, the sentiment data for these posts revealed that most posts were linked to negative sentiments associated with EC.

Our data are in general agreement with a previous study that manually examined health posts from online EC forums [118]. Using an automated data mining method, we were able to look at many more posts, and our results in this larger study are similar to those reported in manually mined data. Generally, the systems most effected for symptoms and disorders were neurological, respiratory and mouth and throat. For both symptoms and disorders, more posts were negative (> 89%%) than positive (<1%), probably because individuals tend to post their negative health problems on websites.

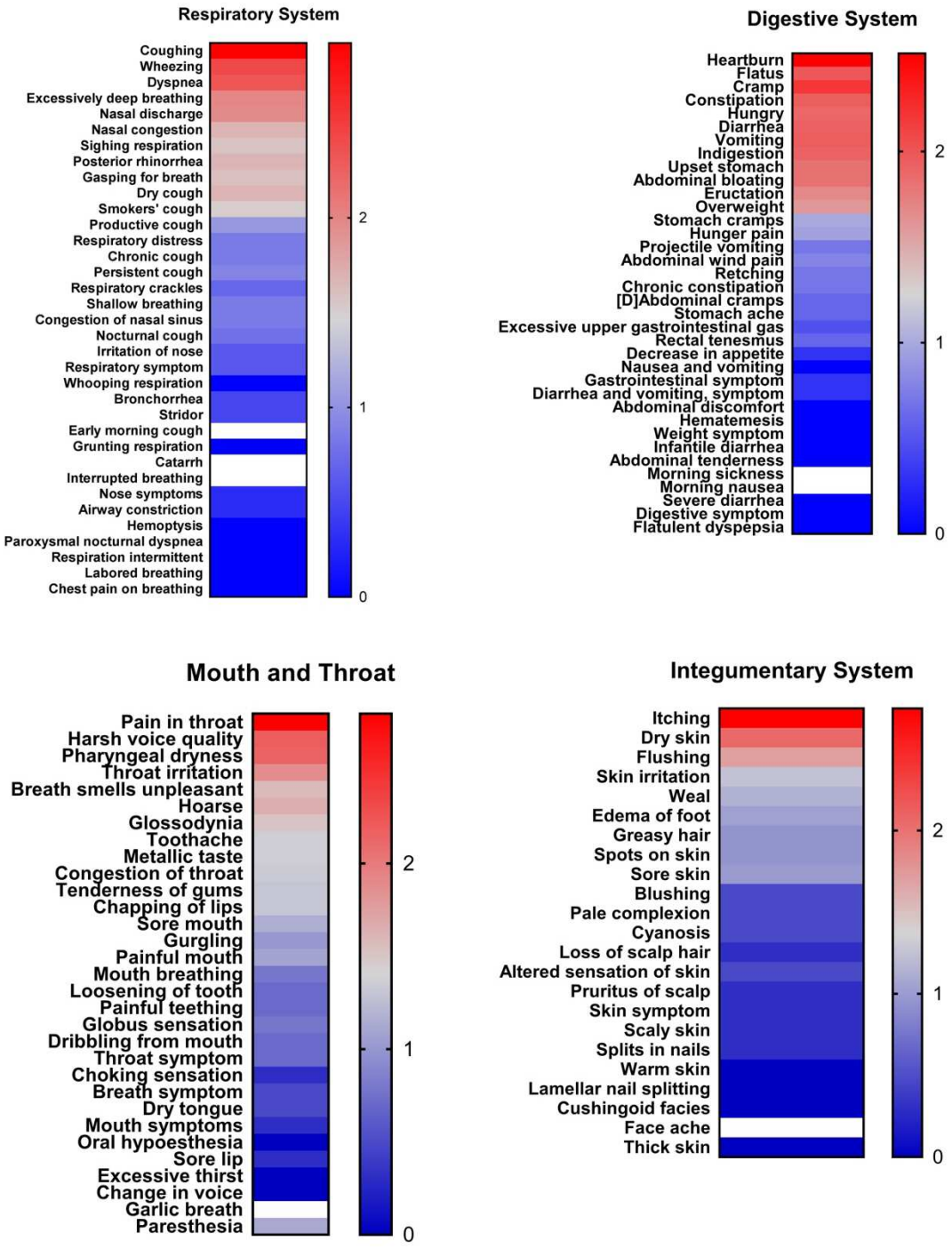

Figure 5.6: Heatmap of all symptoms reported in remaining top systems (respiratory, digestive, mouth and throat, and integumentary). Post count were converted to log scale with from greatest (red) to least (blue).

Also, the posts were sorted according to key terms linked to their descriptions were determined to have a negative connotation.

In our previous study [118], most posts were associated with negative responses to EC use in the respiratory and mouth and throat systems, which was similar to the reporting in the top posts of the disorders of this larger study. In this study, neurological symptoms were associated with EC use and were similar to those we previously reported such as headache and dizziness. The difference in system reporting's could be associated with how the terms were classified/annotated to fit their categories. In our previous study, we found many of the symptoms similarly categorized, however we only examined one sub forum and looked at primary posts (excluding linked posts to primary user posts). If keywords were matched to subsequent posts, they may have been extracted and placed into their appropriate bins.

Although there are relatively few case report studies dealing with EC, those that do exist are consistent with our study in that the systems most often effected by EC use in the existing case reports were the respiratory, circulatory, mouth and throat, digestive, and neurological systems [124], [125].

Several prior studies have been reported on the effects EC have on blood pressure and abnormal heart rate of young adults [126], [127]. For increased heart rate, one study observed increased levels of plasma nicotine were during e-cigarette use. In addition, as in our previous infodemiological study, we found that EC users from online forums reported relatively few different symptoms for the circulatory system such as pounding heart, abnormal heart beat, fluttering heart, and widened pulse pressure. In a previous

published case report, one young male adult suffered from myocardial infarction after use of EC with nicotine in the refill fluid [128]. These online reported symptoms could have significantly long-term health impacts if left untreated.

Nicotine can potentially be an abusive substance and can cause neurological damage to youth and adolescents [129]. Most users use EC with nicotine and artificial flavorings contained in the refill fluids. Previous studied of poison control centers have reported and exponential increase in nicotine poisonings [130] which are often underreported in the peer reviewed literature.

Several prior health-related studies and case reports related to EC have dealt with findings in the respiratory, mouth/throat and digestives system. These were three of five systems that had the most reported symptoms collated data. Typically, users reported coughing/phlegm symptoms. For mouth/throat different pain in throat/mouth and other irritation were often reported, and for the digestive system most users reported heartburn, flatus, and cramping. In the disorders, some posts revealed pneumonia, gastrointestinal and other related mouth ailments, which have been reported in the EC literature case reports [125]. Several human patients have experienced lipoid pneumonia, pleurisy or other lung defects. Two cases of ulcerative colitis have been found in literature [131], [132] which we were able to find hundreds of posts in disorders related to gastrointestinal illness. While for mouth and throat, there have only been mechanical injuries reported (i.e. explosion injuries to the mouth due to EC), many dental professionals are concerned about the adverse effects EC may have on the tooth and gum structure [133].

A recent report on EC usage in the adolescent and young adult populations also warn against the dangers of using EC for these populations, due to potential adverse health effects to the neurological system. One case report from the literature reported a combined neurological/circulatory disorder associated with short term EC use (reversible cerebrovascular syndrome).

There are several components of EC that could contribute to negative health events. Previous studies have reported cytotoxicity related to particular EC flavorings, potentially harmful substances in the aerosol and refill fluids [99], [101], [134]–[137], nicotine overdose and poisoning from EC refill fluids [112], [138]–[142], and general harm from mechanical injury [133], [143]. Also, individual user puffing topography and style of use, choice of EC device and EC refill fluids can all have different contributing factors to various health effects reported [117].

In our paired symptoms listing (not shown here), we found that EC symptoms are not always reported discretely. Users can report multiple symptoms affected by EC use, as we had documented in our previous study as well [118]. In our total collection of posts, we saw the dramatic number of negative sentiment posts compared to positive sentiment, suggesting that EC users have experienced or write more about negative/neutral experiences associated with EC use compared to positive events.

This study was able to track data for the past 7 years by using high-throughput processing to extract and sort amassed data. Using the definitions for user reported symptoms and disorders, we systematically sorted posts to each category and further grouped them into their respective years which they were associated. Future studies

employing similar methods can also take into account potential linked-health histories for certain users who have posted on multiple occasions across different years. We treated these posts discretely, but future histories linking users to multiple posts could potentially create online medical histories for individuals. This can be useful to view progression of conditions or benefits from EC use. Also, this can be used to better monitor user health associated with EC use, or even general health monitoring of health from online databases. As more measures become available to track and evaluate health data online, the Internet becomes increasingly valuable repository where users have identified and self-reported their health effects.

5.5 Conclusion

This study contributes to the growing body of knowledge linking use of EC to adverse effects and negative symptoms. It brings further awareness to the different systemic effects EC can have and also reiterates the importance of using online social media and forums to mine data for health effects attributed to EC use. When we previously reported, the respiratory, mouth and throat and neurological systems had the most posts for adverse effects linked to E-cigarette uses. Now, in literature EC can trigger severe conditions such as lipoid pneumonia and other respiratory distress [144]–[149]. Also, there is growing concern for the effect of EC on the mouth, teeth and gums which are continually being investigated [133]. We have also seen that irreversible damage caused by nicotine to youths is an important public health awareness concern that should continually be addressed [129]. More symptoms and disorders reported will need to be

confirmed with further laboratory assessments and case reports in the literature, however these data can still be useful in monitoring the effects users post or mention in social media outlets and forums.

Chapter 6

Querying Documents Annotated by Interconnected Entities

In a large number of applications, from biomedical literature to social networks, there are collections of text documents that are annotated by interconnected entities, which are related to each other through association graphs. For example, PubMed articles are annotated by Mesh terms, which are related through ontological relationships, and social posts are related through the friendship graph of their authors. To effectively query such collections, in addition to the text content relevance of a document, the semantic distance between the entities of a document and the query must be taken into account.

In this paper, we propose a novel query framework, which we refer as keyword querying on graph-annotated documents, and query techniques to answer such queries. Our methods automatically balance the impact of the graph entities and the text content in the ranking. Further, we propose several indexing schemes and early termination algorithms to generate the top-k results.

Our thorough qualitative and quantitative evaluation on real datasets shows that our methods improve the ranking quality and the execution time compared to baseline ranking systems.

6.1 Introduction

Much research has studied how to query interconnected documents, such as Web pages [150], relational databases (tuples are the documents) [151]–[153], or XML data (XML elements are the documents) [154]. In these settings, the assumption is generally that the user submits a keyword query and the system combines the text similarity with the graph structure to rank documents or collections of documents.

However, this paradigm misses the quite common scenario where the relationships do not exist directly between the documents, but between graph entities contained in the documents. As a first example, consider PubMed documents (or Electronic Medical Records), which are annotated manually or automatically by a set of MeSH (Medical Subject Headings) or SNOMED-CT [155] concepts, where the associations between the concepts are defined by the MeSH and SNOMED-CT ontologies. As another example, consider the posts of a social network, which contain the id of their author (and possibly of the recipients too), and the users are connected through a friendship graph. In the former case, the graph entities are the concepts, and in the latter case the graph entity is the user who submits the query.

We specifically study the problem where the query, in addition to keywords, specifies one or more graph entities (or simply "entities") of interest. For example, as shown in Figure 6.1, in the case of PubMed, the query may specify the concept (entity) "Heart Valve Finding (C5)" and the keywords "dyspnea" and "fever". This means that the user is interested in documents that have been annotated by the concept "Heart Valve Finding (C5)" or a similar one, and are relevant to the keywords "dyspnea" and "fever".

In the social networks domain, the user who submits the query becomes the entity, meaning that he is more interested in documents posted by closer friends. For example, if user John in Figure 6.2 submits a query with keyword "birthday", he likely prefers a post related to "birthday" by his direct friend Mike than by his farther friend Bob (assuming other factors such as freshness, user influence and text similarity are the same).

We refer to the above type of queries as keyword queries on graph annotated documents, to differentiate them from the more traditional keyword queries on interconnected data mentioned above. In addition to the biomedical data and social networks, the keyword queries on graph-annotated documents can be found in other domains. In e-commerce, products have a description and are annotated by a product category (e.g., "SLR Camera"); the category becomes the entity. In spatial databases, each document may have a location and a graph of (e.g., roads) that defines the distances between them; the user's location becomes the entity.

A query may have several entities from the same or different association graphs. For example, a user may be searching in a health forum for information related to concept "Heart Valve Finding (C5)" (first entity), and have a preference towards posts from closer friends (second entity), and also specify a set of keywords.

A key challenge in effectively answering keyword queries on graph-annotated documents is to balance the importance of the various graph entities and the keyword terms. For example, in the above social network query by John, if there is a post by Bob that is very relevant to keyword "Obama", it is likely a better result than a post of Mike that is less relevant to Obama, if Obama is a global (not local) topic in the social network.

In contrast, if John specifies keyword “birthday”, he is likely interested in birthday posts of his friends only.

A second key challenge is how to execute keyword queries efficiently with graph entities, given that we don’t know how far we have to go on the association graphs (in the social network application the friendship graph) to find the top results. This is different from the traditional query expansion [156] problem, where we typically select a few terms to add to the query.

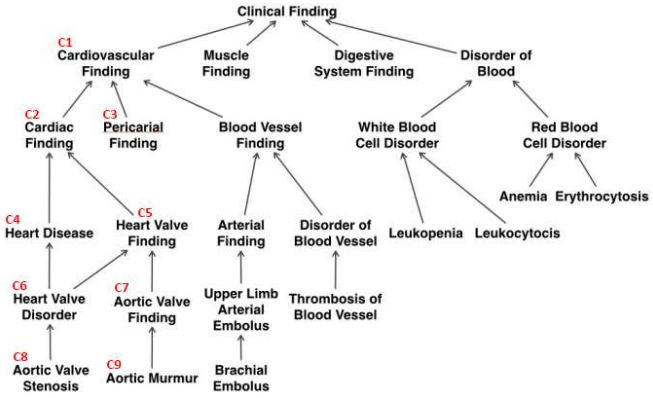

Figure 6.1: A subgraph of the SNOMED-CT ontology

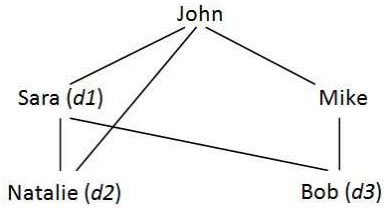

Figure 6.2: Example of social network graph showing the post IDs for users

In this work, we make the following contributions:

• We define the keyword queries on graph-annotated documents problem, and propose an effective framework that intelligently balances graph entities relevance with key- word relevance (Sections 6.3 and 6.4, respectively).

• We propose several indexing schemes and efficient algorithms to answer keyword queries on graph-annotated documents (Section 6.5).

• We evaluate the results quality of our proposed frame- work, comparing to several baseline methods through a comprehensive user study (Section 6.6.1).

• We evaluate the time performance of our proposed frame- work on two real datasets (Section 6.6.2).

Section 6.2 presents the related work and we conclude in Section 6.7.

6.2 Related Work

6.2.1 Keyword search in databases

There is much work on key-word search on interconnected entities. The key goal is to find subtrees of connected entities that collectively contain all the keywords [151], [153], [154], or to leverage the links to rank the importance of each entity [152]. However, they do not consider contained graph entities or even ontological relationships, as they focus on exact matches.

6.2.2 Ontology-based query expansion

XontoRank [157] exploits ontological relationships to answer keyword queries on XML documents. For that, it precomputes the semantic distance for each pair of keyword

and concept up to a specified semantic distance threshold, and uses this information to expand the query keywords with relevant concepts. Precomputing all pairs of distances is infeasible in our setting as we do not have a distance threshold. Further, their query only contains keywords and not graph entities, which are necessary in our setting (e.g. to specify the user who submits a query in a social network). The same limitations also apply to Arvanitis et al. [158], which proposes methods to find electronic medical records (EMRs) similar to a query EMR. Another difference is that XOntoRank uses tree traversal algorithms to find subtrees that contain all keywords, whereas we are searching collections of documents.

Xiong and Callan [15] use semi-structured data sources such as controlled vocabularies and knowledge bases to improve the quality of ranking – entities from external sources are used as objects connecting query and documents. Their proposed technique "EsdRank" annotates the query using related objects from external data to improve retrieved documents. Tonon et al. [16] answer keyword queries using an inverted index and semi-structured data by expanding queries entities to improve search effectiveness. Dalton et al. [14] proposed entity query feature expansion (EQFE), which uses semi-structured data (Google Freebase) to annotate query and documents with features from entities, including structured attributes and text in order to maximize the retrieval effectiveness. Considering query term dependencies, Nikolaev et al. [159] account for full query term dependencies and sequential query term dependencies when expanding the query using semi-structured data, and use number of statistical and linguistic features to estimate their probabilistic mapping onto the fields of semi-

structured data. The aim of these works is to effectively expand the query using the semi-structured data entities, whereas in our setting the entities are provided by users as part of the query, and our focus is on (a) balancing the impact of entities and keywords and (b) studying the time performance.

6.2.3 Top-K algorithms

Several top-k algorithms have been proposed to combine ranked lists [160]. However, these works generally assume that we know the set of ranked lists, whereas in our case we have to consider jumping from the list of an entity to the lists of its neighbors on the association graph. Nevertheless, our top-k algorithm build upon the early termination and threshold computation ideas of previous work [161].

6.2.4 Search in social networks

Bonaque et al. [162] presented a new data model called S3, which captures the properties of rich sources like social networks, including social, structured, and semantic dimensions. They also proposed a top-k algorithm called S3K that retrieves the most relevant documents or document fragments by considering the three dimensions captured by S3. However, the structure captures only one entity relation between the document and the user since it's mainly for social networks, while in our work a document may be annotated by more than entity. Moreover, their queries consist only one entity called "seeker", while in our work we can have one or more query entities from one or multiple association graphs. Maniu and Cautis [163] proposed a top-k algorithm that relies exclusively on weight of tagging systems. Given a query consisting of a user and keywords (tags in this case), the algorithm shall return top-k relevant documents having

the highest score with respect to the keywords and the proximity (or social) scores between the users. Hence, items tagged by closer users are given more weights, while in our work we balance the importance of annotation entities with query keywords.

Table 6.1: Association graphs

Association Graph	**Entity**	**Association Edge**
ontology	concept	semantic relationship
social network	user	friendship
spatial objects	object location	distance

6.3 Problem Definition and Semantic

Let D be a collection of documents. Each document $d \epsilon D$ is defined as a tuple $(d.w, d.u)$, where $d.w$ is its textual content and $d.u$ is a set of graph entities by which d is annotated.

The graph entities are related to each other through one or more association graphs. An association graph $G = (N, E)$ consists of a set of entities (nodes) *N*, and a set of association edges *E*. The nodes and edges in various association graphs are shown in Table 6.1. An ontology graph is often a Directed Acyclic Graph (DAG), where *N* is a set of concepts and *E* is a set of relations between the concepts. On the other hand, a social network graph could be an undirected (e.g. friendship connections in Facebook) or directed graph (e.g. follower/followee connections in Twitter), where *N* is a set of users and *E* is a set of relationships between users. In the rest of the paper, for simplicity of presentation, we assume that there is only one association graph; extending the algorithms for multiple association graphs is straightforward.

Figure 6.1 shows a subset of an association graph based on the SNOMED-CT ontology, where concepts are related through "is-a" relationships. The SNOMED-CT ontology [155] contains more than 300K medical concepts that are connected through various relation- ship types including the "is-a" relationship (e.g. head "is-a" body-part). In our experiments, we only consider the directed "is-a"' relationships for SNOMED-CT. For each document $d \epsilon D$, the textual data is analyzed to extract the UMLS concepts using the MetaMap tool [13]. Note that we only indexed concepts that correspond to SNOMED-CT concepts.

Example 6.1: Consider the following documents that correspond to substrings of PubMed abstracts:

d1: "A 59-year-old male had a latent epicardial mass discovered at cardiovascular imaging during the assessment of an aortic murmur (C9)"

d2: "Closure of an atrial septal defect with a one-way flap (C6) patch in a patient with severe pulmonary hypertension (C1)"

d3: "A 63-year-old man who was admitted to the emergency department with new and spontaneous onset of fatigue, dyspnea, and palpitations (C1)"

The concepts detected in each document are underlined, and the concept ids (e.g., C9) correspond to the ones in Figure 6.1.

Example 6.2: Consider social posts, which are annotated by the id of their author. In contrast to ontology graphs, each document here can be annotated with only one entity. Sup- pose we have the following posts and the social network graph shown in Figure 6.2.

d1: Sara: “Obama to announce $600 million in grant programs to prepare workforce for jobs”

d2: Natalie: “Michael Bloomberg Pledges Million to Push Gun Control”

d3: Bob: “Obama Supporters Don’t Know Obama”

A keyword query on graph-annotated documents $Q = (Q.w, Q.u)$ consists of $Q.w$, a set of query keywords, and $Q.u = (Q.u_1, \ldots, Q.u_m)$, a set of graph entities, from one or more association graphs (we focus on one association graph as explained above).

A Top-k keyword query on graph-annotated documents returns a ranked list of the k most relevant documents from D based on a similarity function that combines both the graph entities and the textual similarity.

Example 6.1 (cont’d): A query is Q = ({“cardiovascular”}, {Heart Valve Finding (C5)}), where “cardiovascular” here is the keyword, and Heart Valve Finding is the concept (graph entity).

Example 6.2 (cont’d): if user “John” submits keyword query “Obama policies”, the corresponding keyword query with graph entity is Q = ({“Obama”, “policies”}, {John}).

Our contribution in this paper is to balance the impact of the query keywords and the query entities in the ranking (Section 6.4) and to compute the top-k documents efficiently (Section 6.5).

6.4 Ranking Semantics

To keep the ranking model generic in terms of combining functions, we define separately the impact of the query keywords $IRScore(d.w, Q.w)$ and graph entities $Dist(d.u, Q.u)$ and combine them by a monotone aggregate function. We emphasize that our focus is on (a) the effective balancing of the keywords and entities impacts and (b) the efficient query execution, and not on the best way of defining the semantic distance between two graph nodes, which has been studied extensively as discussed in Section 6.2. The proposed algorithms can be adapted for a wide range of monotone impact and combining functions. The monotonicity for the graph entities is defined on the path length between two entities, whereas for the keywords impact is defined on the term frequencies or other text features.

The score of the document d for query Q is:

$$score(d, Q) = f(Dist(d.u, Q.u), IRScore(d.w, Q.w)) \quad (6.1)$$

The combining function f may include other features such as document or user popularity. We adopt a previously proposed combining function that multiplies the impact of the two components and uses a decay factor for the entities distance [154] (originally used in the context of XML documents):

$$score(d, Q) = \alpha^{Dist(d.u, Q.u)} \times IRScore(d.w, Q.w) \quad (6.2)$$

where $\alpha < 1.0$ is the distance decay factor in the association graph G. A key challenge, which we tackle in Section 6.4.1, is the computation of α.

For the purpose of the experiments, to compute $Dist(d.u, Q.u)$, we build upon previous work [157], [164], and define it as the sum of the shortest path distances between each of the query's graph entities in $Q.u$ and their closest document entity in $d.u$. In other words, $Dist(d.u, Q.u)$ is the sum of the number of edges between every entity in $Q.u$ and its closest entity in $d.u$. Formally, $Dist(d.u, Q.u)$ is defined as follows:

$$Dist(d.u, Q.u) = \sum_{q \in Q.u} G.ShortestPath(q, d.u) \tag{6.3}$$

where *ShortestPath* computes the length of the shortest path in association graph between an entity q and its closest entity in set $d.u$. In the case of multiple association graphs, the score is defined as following:

$$score(d, Q) = \prod_{for\ each\ G} \alpha_G^{G.Dist(d.u, Q.u)} \times IRScore(d.w, Q.w) \tag{6.4}$$

where α_G is the decay factor for association graph G.

An example of a specific text ranking function $IRScore(d.w, Q.w)$ used in our experiments is as follows:

$$IRScore(d.w, Q.w) = \sum_{t \in Q.w} tf(d.w, t) \times idf(t) \tag{6.5}$$

where the normalized term frequency is defined as

$$tf(d.w, t) = \frac{f(t, d.w)}{Sum_{s \in d.w} f(s, d.w)} \tag{6.6}$$

and $f(t, d.w)$ is the frequency of term t in $d.w$, and

$$idf(t) = log\ \frac{|D|}{Count(d \in D : t \in D)} \tag{6.7}$$

where $|D|$ is the number of documents in collection D. Other Information Retrieval keyword similarity functions such as BM25 are possible [165].

6.4.1 Computation of α Parameter

In this subsection, we explain how α is computed to balance the relevance of the graph entity distance with the key- word similarity. We argue that the following intuition holds, which we also evaluate in Section 6.1: If the documents that match the query keywords have similar content regardless of their distance to the query's graph entities, then the distance should have a smaller importance.

Specifically, if the association graph is a social network, this means if user John specifies keyword "Obama" and his friends do not have any consistent political views (e.g., some are Republican, some Democrat, and some undecided), then John would likely be interested in posts about Obama coming from both close friends and the rest of the network. Thus, the importance of $Q.u$ is higher. In contrast, if John's friends discuss a topic about Obama (e.g., his immigration views), which is distinct from the general chatter about Obama on the whole network, then John would prefer posts from his friends

rather than from the rest of the network; thus, the importance of $Q.w$ is higher. As another example, for query "birthday party", if John's friend had a party, then John would be most interested in posts about that party and not about a random party on the network.

To achieve the above intuition in the social network application, we compute the content difference between the local community and the whole network for the set of documents that match the query keywords $Q.w$. A popular measure of the difference of two sets of documents is the Kullback Leibler (KL) divergence [166], which measures the difference between two probability distributions, specifically the distribution of terms in the posts relevant to $Q.w$ from the user's social neighborhood, and the distribution of terms in the posts relevant to $Q.w$ in the whole network:

$$KL(R_Q^u, R_Q) = \sum_{v \in vocab} R_Q^u(v) log \frac{R_Q^u(v)}{R_Q(v)} \tag{6.8}$$

where R_Q^u and R_Q are the probability distributions of the relevant posts in the neighborhood of the user *u* and in the whole social network (hence the latter may be precomputed as a set of term, probability, pairs), respectively. Let *D* be the set of all posts in the social network, and let D_Q be the subset of *D* that contains at least one of the keywords in $Q.w$, and D_Q^u be the subset of D_Q posted by users with distance up to *T* from the query user u. Suppose we have *n* query terms in $Q.w$, we compute the exact value of R_Q as:

$$R_Q(v) = \frac{\sum_{d \in D_Q} tf(d, v)}{\sum_{d \in D_Q, v' \in vocab} tf(d, v')} \tag{6.9}$$

To compute R_Q^u, we concatenate the text of all posts that are relevant to $Q.w$ (e.g., that contain all terms in $Q.w$) posted by users with distance up to T from the query user u. We set the threshold T = 1 to only consider direct friends.

$$R_Q^u(v) = \frac{\sum_{d \in D_Q^u} tf(d, v)}{\sum_{d \in D_Q^u, v' \in vocab} tf(d, v')} \tag{6.10}$$

As an example, the word "Peter" may appear with probability $R_Q^u(Peter) = 0.001$ in the user's neighborhood and with probability $R_Q(Peter) = 0.00005$ in the whole social network. To incorporate the *KL* measure in our scoring function (Equation 6.2), we need first to normalize it for each query *Q* between (0, 1) since *KL* is unbounded. Therefore, we define using *KL* as follow:

$$\alpha = e^{-KL(R_Q^u, R_Q)} \tag{6.11}$$

That is, the larger *KL* means the two sets of documents are different, and hence the posts from user's neighborhood are more preferable. The same rationale applies to several other types of association graphs, such as the ontology graph discussed above.

Table 6.2: Indexes and algorithms that use them (x denotes that an index is used by an algorithm)

Index	Description	Access	Entity-first	Term-first	Parallel
TF Lookup index	$(term, documentID \rightarrow tf$	random	x		
Term index	$(term) \rightarrow [documentID, tf]$	sequential		x	x
Entity index	$(entity) \rightarrow [documentID]$	sequential	x		x
Distance lookup index	$(entity, documentID \rightarrow dist$	random		x	x

6.5 Indexes and Algorithms

In this section, we first present necessary indexes to store the data for the proposed algorithms; then we describe our three proposed algorithms to generate top-k results efficiently for keyword queries on graph-annotated documents.

6.5.1 Indexes

As it is shown in Table 6.2, we created four different indexes: 1. The *TF lookup* index, which is used for random accesses to get term frequency given a term and a document ID. 2. The Term index, is an inverted index to map each term to a list of $(documentID, tf)$ pairs sorted by term frequency in descending order. 3. The *Entity* index, that maps each entity to a set of documents IDs. 4. The *Distance lookup* index is an adaptation of Akiba's exact shortest-path index, which allows computing pairwise shortest paths on large graphs [167].

Briefly, Akiba's shortest-path computation is based on distance labeling of vertices, in which the algorithm conducts a breadth-first search from all vertices with pruning to build an index. In the labeling method, they precomputed a label $L(v)$ for each vertex v in graph G that contains pairs of (u, δ_{uv}), where u is a vertex and δ_{uv} is the distance between v and u in graph G. The pruning reduces the number of labels, which yields fast preprocessing time, small index space, and fast query time. To answer a distance query between two vertices v and u, a merge-join algorithm is used to find the minimal distance between the labels of vertices.

To adapt this method to our datasets, we first built the index using the association graph, and then we added all the documents to the index as vertices, where each document's label is a union of the graph entities labels' in which the document is attached to.

If two pairs share the same vertex from different entities, we keep the pair with the minimum distance. Adding documents labels after building the index of the association graph is more useful when new documents are added to the dataset, where the labels can be added separately instead of building a whole new index.

Considering example 6.1, suppose we have the following two labels $L(C1), L(C6)$ for concepts $C1$ and $C6$ from Figure 6.1, respectively:

$$L(C1) \rightarrow [(C2, 1), (C3, 1), (C4, 2), (C5, 2)]$$

$$L(C6) \rightarrow [(C4, 1), (C5, 1), (C8, 1), (C7, 2)]$$

This means that the shortest path from $C1$ to $C2$ has length 1, from $C1$ to $C3$ length 1, from $C1$ to $C4$ length 2, and so on. Thus, the label set of document *d2* that is annotated with $C1$ and $C6$, is the union of $L(C1)$ and $L(C6)$ labels:

$$L(d2) \rightarrow [(C2,1),(C3,1),(C4,1),(C5,2),(C8,1),(C7,2)]$$

Which means the shortest distance from *d2* to $C2$ is 1, and so on. Note that, if a vertex is in both lists, the vertex with the minimum shortest-path stays in the union. In example above, $(C4,2)$ and $(C5,2)$ are removed from the union of the $L(C1)$ and $L(C6)$.

The following theorem shows that the above method of taking the union of the labels of the graph entities of a document allows computing optimal shortest paths between that document and any entity in the association graph.

Theorem 6.1 *To compute the shortest path of a document from a single entity, the document only requires to maintain the labels that are associated with its entities.*

Proof Sketch: For example, suppose a document *d* includes two entities $C1$ and $C2$, where each entity has the following list of labels:

$$C1 \rightarrow [(C0,2),(C3,4),(C5,3),(C10,1)]$$

$$C2 \rightarrow [(C0,1),(C4,3),(C5,6),(C8,4)]$$

Suppose $C9$ is in the list of $d \implies \exists$ node C_x, such that shortest path from $d \xrightarrow{C9} C_x$. Therefore, shortest path from $C1$ or $C2$ to C_x goes through $C9$, which is not possible since neither one has $C9$.

In the next section, we explain how each algorithm uses a subset of indexes above to compute *top-k* results.

6.5.2 Algorithms

We present three algorithms to compute exact top-k results: 1. *Entity-first*, 2. *Term-first*, and 3. *Parallel*. Briefly, the Entity-first algorithm uses Entity index to match the query's graph entities $Q.u$, and then performs lookup using *TF lookup* index to find the term frequency of each query keyword for the matched documents. The Term-first algorithm uses Term index to match query keywords $Q.w$, and then does lookup using *Distance lookup* index to match the query's graph entities $Q.u$. The *Parallel* algorithm uses both Entity and Term indexes to efficiently find the *top-k* relevant documents.

Table 6.3: Main variables used in our algorithms

candidates	Set of documents which are candidates of the query answers
top-k	Set of *top-k* documents
th	The maximum possible score of unseen documents
min-tk	The minimum score of current *top-k* documents (the min value is used if the exact document score not known yet)
max-m	The maximum possible score of the documents in candidates (i.e., max of documents' maximum values)
dist	Current sum of the distances (i.e. of last accessed documents) from all query's graph entities

Each algorithm exploits early termination through computing a threshold for the unseen documents, in a way in- spired by the threshold algorithm.

Table 6.3 lists the main variables we used in our algorithms. We maintain minimum and maximum scores for each document d in the candidate documents for *top-k* results, where minimum is computed by replacing unknown values with 0 in Equation 6.2, and maximum is computed by replacing the unknown values with current list values. It is worthy of mention that in this paper, we do not limit the distance computation by a threshold.

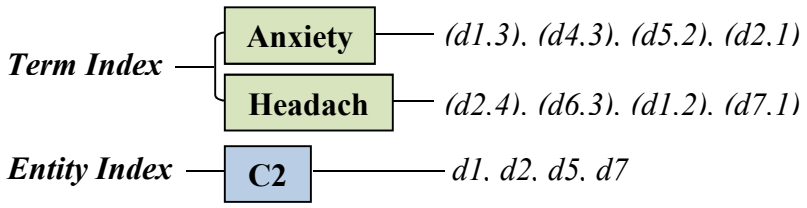

Figure 6.3: Example of Term and Entity indexes from Table 6.2

Entity-first Algorithm

In this method, the association graph is loaded into memory to access the neighbors of the query's graph entities whenever needed. As shown in Table 6.2, we use two indexes. We first access in parallel the lists of document IDs of the query's graph entities $Q.u$ using the Entity index. For each retrieved document, we perform a random access using *TF lookup* index to get the tf score and compute the document's score using Equation 6.2. Once the document list of an entity is exhausted, we move to the lists of its neighbors, which we access through the association graph. The algorithm terminates when the maximum possible score of seen documents max-m and the maximum possible score of unseen documents threshold are less than the minimum *top-k* value *min-tk*.

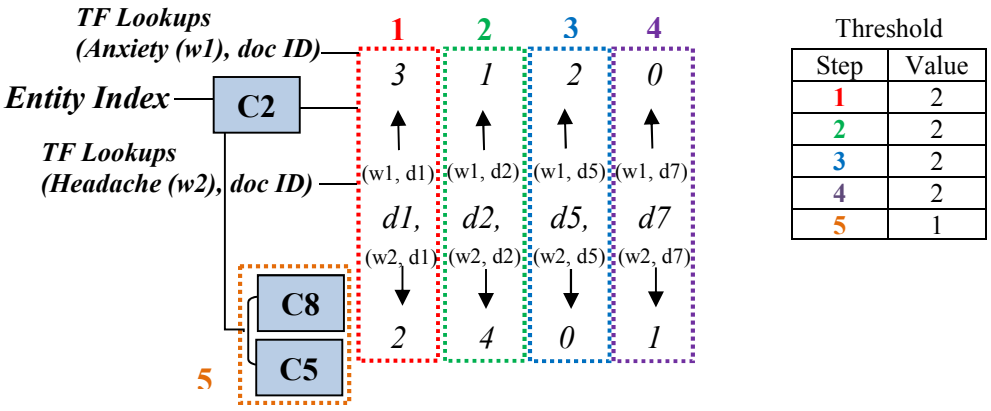

Figure 6.4: Entity-first algorithm

To better illustrate the algorithm, consider the following example:

Example 6.3: Suppose a user submits query *Q*=({"anxiety", "headache"},{cardiac finding (C2)}), where "anxiety" and "headache" are the query keywords, and "cardiac finding(C2)" is the entity shown in Figure 6.1. Figure 6.3 shows an example of retrieved document lists from Entity and Term indexes. The entity-first algorithm starts by retrieving one or more lists of document IDs annotated by the query's graph entity (*C2*) Figure 6.4. Next, the algorithm processes document *d1*, and then does random access using *TF lookup* index to get the tf scores of the document *d1* with both query keywords ("anxiety", "headache") - step 1 in Figure 6.4. After retrieving the tf scores of *d1*, the algorithm sets the threshold and computes the minimum and maximum scores for *d1* using Equation 6.2. If the minimum score of *d1* is higher than the threshold, then the algorithm adds it to top-k or keeps it in the candidates when its maximum score is higher than the minimum score of current *top-k* documents. The algorithm continues to process the rest of the documents *d2, d3, d4* in the same way. If all the documents that are

annotated by query's graph entity *C2* have been processed, and the termination condition has not satisfied yet, the algorithm will then access the neighbors of *C2* to retrieve their document lists, and updates the threshold accordingly.

Term-first Algorithm

In this algorithm, two indexes are used: *Term* and *Distance looku*p as mentioned in (Table 6.2). For each query keyword in $Q.w$, the algorithm accesses in parallel a list of document IDs and tf scores pairs using the *Term* index. For each seen document, we do a random access for each query's graph entity in $Q.u$, to find the distance between the query's graph entity and the document. Similar to the *Entity-fir*st algorithm, this algorithm terminates when the maximum possible score of seen documents max-m and the maximum possible score of unseen documents th are less than the minimum *top-k* value *min-tk*.

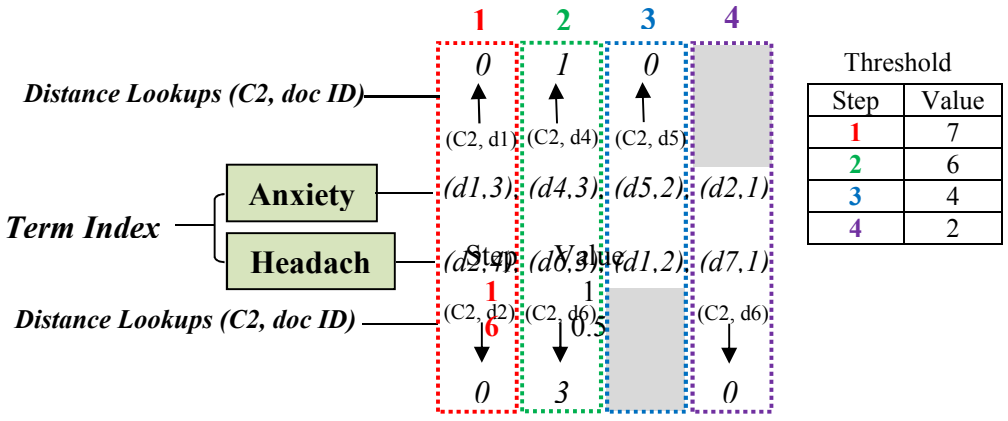

Figure 6.5: Term-first algorithm

Considering Example 6.3, the algorithm first retrieves the lists of the query's keywords ("anxiety", "headache") Figure 6.5. Then, the algorithm processes a document from each list at the same time *(d1, d2)*, and then accesses the *Distance lookup* index to get the distance between the documents and the query's graph entity C2. Next, the algorithm sets the threshold and computes the minimum and maximum scores for *(d1, d2)* using Equation 6.2. Similar to the previous algorithm, if the minimum score of either d1 or d2 is higher than the threshold then the algorithm adds it to *top-k* or keeps it in the candidates when its maximum score is higher than the minimum score of current *top-k* documents. The second step is similar to step 1, where the algorithm processes *(d3, d6)*. Steps 3 and 4 are also similar to the first two steps, except that the algorithm does not do Distance lookup for *(d1, d2)* since it has already done in the first two steps. The algorithm continues the same steps until the termination condition satisfied, or all the term lists are exhausted.

Parallel Algorithm

In this method, two indexes Entity and Term are used as pointed in Table 6.2. Simultaneously, the algorithm retrieves at least two lists at the same time for each query's graph entity and query's term from both indexes. Similar to the previous methods, this algorithm terminates when the maximum possible score of seen documents max-m and the maximum possible score of unseen documents *th* are less than the minimum *top-k* value min-tk. Parallel is a non-random access algorithm that reads both the *Term* and *Entity* indexes in parallel. As we discuss later, in the special case where the term lists are exhausted, we perform random access using the *Distance lookup* index.

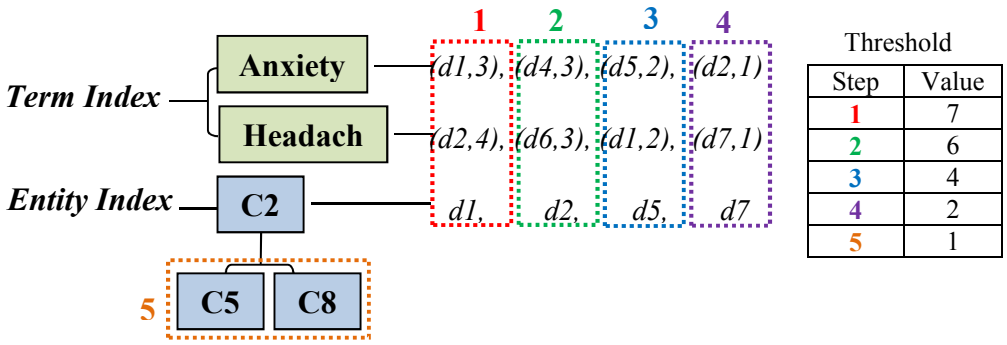

Figure 6.6: Parallel algorithm

Considering Example 6.3, the algorithm retrieves the lists of the query's keywords ("anxiety", "headache") and entity *C2* as shown in Figure (6.6). In the first step, the algorithm first processes a document from each set, *d1* and *d2* here, and computes their minimum and maximum scores along with setting the threshold value. If the minimum score of any document is higher than the threshold, the algorithm adds it to top-k or keeps it in the candidates when its maximum score is higher than the minimum score of current top-k documents. The algorithm continues to process the documents in the same way, and updates the scores of the documents based on their existence in other lists. If all the lists attached to *C2* have been retrieved and the termination condition has not been satisfied, the algorithm will then access the neighbors of *C2* (step 5 in Figure 6.6). Another possible case is when the term lists are exhausted, i.e. there are no more lists to retrieve. In this case, we use the *Distance lookup* index to find the distance between all the documents in candidates and the query's graph entities to compute the exact score and return *top-k* results.

In the next section, we present the qualitative experiments to measure the precision of *KL-based* method, and we also present the time performance of the three algorithms.

Table 6.4: Description of datasets

Property	Entities	Documents
Health Web Forums	296,433 concepts	2,961,526
Twitter	18,492 users	221,643

6.6 Experiments

In this section, we present the experimental evaluation of our proposed algorithms. First we describe our datasets, and then we present two types of experiments: qualitative and time performance.

Datasets: We conducted our experiments on two real datasets: (i) Health Web forums, and (ii) Twitter.

The health Web forums dataset was obtained from various websites including dailystrength.org support groups, Webmd.com, and Drugs.com. The collected posts from these sources were further parsed and annotated with medical concepts corresponding to SNOMED-CT ontology using MetaMap tool [13]. We don't evaluate the precision of the annotated data since it is out of our scope. Posts with less than 10 words were ignored as they are often spam or convey no useful information.

The Twitter dataset was obtained using the Twitter Streaming API [37]. Since the relations between users are bidirectional, i.e. each user has followers and followings, we discarded all unidirectional relations to convert the graph to undirected one, that is, we only keep an edge between two users if they follow each other. The goal of only

considering bidirectional relations is to define the local and global communities for each user. Since there is no large public Twitter dataset that has users' connections, and the rate limit to get user's connection from Twitter REST API is strict (one request/minute for each followers and followings), we only use a small dataset. Similar to health Web forums, we ignored tweets with less than 5 words.

Table 6.4 shows a description of our datasets, including the number of the association graph entities and documents.

Setup: All indexes and algorithms were implemented in Java. The experiments were conducted on a 12 core AMD Opteron(tm) Processor 6168 using 64 GB of RAM. All index structures are disk resident, and stored on a Cassandra database.

6.6.1 Qualitative

In this first set of experiments, we evaluate our proposed ranking method, which computes α parameter using the *KL* strategy based on user's community, as we discussed in Section 6.4.1. To achieve this, we used 20 queries, where each query consists of a user who submits the query, and a list of keywords. For that, we selected 15 different users from Twitter dataset, where each user has at least 20 friends and 10 tweets. For each query Q, we combined the query keyword $Q.w$ with the user id as the graph entity $Q.u$, and then we computed the top-3 results by using 6 different methods:

1. Two baselines:
    - $IRscore$ baseline: computes document scores using the text similarity only, and ignore the distance to the $Q.u$

- Distance baseline: orders the documents by their distance from the user; for ties orders by decreasing document freshness

2. Static α parameter, using Equation 6.2. We consider α =0.1, 0.5, and 0.9

3. Adaptive (query-specific) α using *KL* divergence, as shown in Equation 6.10

After finding the top-3 results for each of the methods, we took the union of the results and conducted a user study. We asked seven students to imagine that they are the selected Twitter users $Q.u$, and to mark the top-3 relevant tweets for each of the 20 queries (the distance of each tweet to the user is also displayed). To help them get an idea of what their friends and the rest of the network talk about, we provided them with the following information:

1. Top 20 tweets of the local community (immediate network) that contain the query keyword

2. Top 20 tweets of the global community (all users in the network) that contain the query keyword

After the students selected the top-3 results for each query, we took the majority voting to define the top-3 "ground truth" results for each query, and then we compared all the methods with the students' selections in terms of accuracy, that is, how many of their top-3 results are in the ground truth top-3.

Table 6.5 shows the accuracy of the 20 queries. We see that using KL-based method to compute α parameter achieved accuracy of 88.33% comparing to the students' selections, which is 51.66% improvement over $IRscore$ baseline and

66.66% improvement over Distance baseline. Our method using α = 0.5 also achieved accuracy of 76.67%, with improvement of 43% over $IRscore$ baseline and improvement of 55% over Distance baseline.

To intuitively explain the role of *KL* divergence in computing α parameter, consider the query keyword "Michael", where *KL* divergence is high between the local and the global communities, specifically *KL*=1.11. The reason for the high *KL* value is that the local community for the user who submitted the query keyword "Michael" talks more about Michael Bloomberg, while the global community talks more about Michael Tsarion and Michael Donnor. Since the two communities are different, tweets from local community are preferred, and hence α here equals 0.33 by using Equation 6.10. Another example is query keyword" Constitution", where *KL* divergence here equals 0.05. Thus, both local and global communities talk about the same constitution, which means the user is more likely interested in tweets from both local and global communities when selecting top relevant tweets. To avoid computing the exact *KL*, we only consider the top recent tweets in both communities (1000 tweets for local and 5000 tweets for global).

6.6.2 Time Performance

In this section, we evaluate time performance by varying several parameters. Table 6.6 shows a list of the parameters and their ranges, where default values are shown in bold. For all experiments, we vary one parameter while using the default of other parameters. Each experiment is the average running time of 100 queries. As what we mentioned before, we don't limit the distance computation by a threshold in the association graph since this is not applicable for Parallel algorithm. We compare our

algorithms to the lookup algorithm, which processes all the documents that are sorted by the documents ids and attached to the $Q.w$, and does lookup to find the distance between the documents and the query's graph entities. Specifically, it first reads the whole term lists and gets the exact term score of each document that contains at least one of the query terms, and then looks up the distance to the query's graph entities using the distance lookup index.

Table 6.5: Query keywords and number of matches per ranking method

	IRscore Baseline	**Distance Baseline**	**α=0.1**	**α=0.5**	**α=0.9**	**KL-based**
Obama	1	0	0	3	1	1
Forest	2	1	3	3	2	3
Tax	0	1	2	3	2	3
Wednesday	0	0	3	3	0	3
Madrid	1	0	1	3	2	3
Prince	0	2	3	3	1	3
Bundy	1	0	2	3	1	2
Michael	0	2	3	3	0	3
Sonia	3	0	0	0	3	3
Gun	3	0	2	2	3	3
Happy	1	0	3	3	3	3
Mystery	2	1	3	3	2	3
Food	0	3	3	1	0	3
Hope	0	0	3	3	1	1
Constitution	3	0	1	1	3	3
Beautiful	0	1	3	1	2	2
Photo	0	1	3	3	0	3
Law	2	0	1	2	2	2
Market	2	1	3	2	2	3
Dream	1	0	1	1	2	3
Total	22	13	43	46	32	53

Figure 6.7 and Figure 6.8 show the results for health Web forums and Twitter, respectively. In general, Parallel algorithm out- performed the other two algorithms in both datasets, with an exception where k=1.

Table 6.6: Values for parameters. The default value is in boldface.

Parameter	Range
k	1, 10, 20, 50, 100
α	0.1, 0.3, 0.5, 0.7, 0.9
No. of terms	1, 2, 3, 4, 5
No. of graph entities	1, 2, 3, 4, 5
Entity frequency	low, medium, high
Term frequency	low, medium, high

Varying k The experimental results when varying k are shown in Figures 6.7a and Figure 6.8a for health Web forums and Twitter, respectively. In both datasets, Parallel algorithm performed better comparing to the other two algorithms, as it minimizes the random accesses. We also evaluated the Lookup algorithm on health Web forums dataset, which is not TA-based algorithm.

Varying α Since our focus here is on time performance, we don't compute the α parameter for each query, but instead, we use fixed values to measure the runtime for each algorithm. Figures 6.7b and 8b show the effect on performance when varying the α parameter. Similar to varying k, Parallel algorithm performed better in both datasets. In health Web forum, Entity-first algorithm is slower with a high α value, where more documents become candidates to the query answer as what is shown in Figure 6.9.

Varying the number of terms Figures 6.7c and 6.8c show the impact of changing number of terms in both datasets. When there is one term, Term-first performed better in both datasets. A probable reason is that Term-first does less Distance lookup when there is only one query keyword, while all other experiments have two query keywords as default. Other than that, Parallel algorithm achieved better results in both datasets.

Varying the number of graph entities We only vary the number of graph entities in health Web forums, since Twitter queries have only one entity, which is the user who submits the query. Figure 6.7d shows the results of varying the number of entities. Parallel algorithm achieved better results than the other two methods.

Varying graph entity and term frequencies We vary the frequency of both term and graph entity by splitting the frequency to three types: low, medium, and high. In low frequency, the term or graph entity have between 6-15 sets of documents, where each set has 200 documents. Medium frequency means there are 60-200 sets of documents, while high frequency means there are 450-1020 sets of documents. In this part, we only vary the frequency in health Forum dataset, since Twitter dataset is small. Figure 6.7e shows the results for varying both concept (graph entity) and term in health Web forums. Entity-first algorithm total runtime is similar despite the frequency; however, it is faster when both concept and term frequencies are medium. Term-first and Parallel algorithms total runtime increase when both concept and term frequencies increase. Also, both algorithms are faster when the concept frequency is high and the term frequency is low, and the probable reason is that both algorithms terminate when all term lists exhausted.

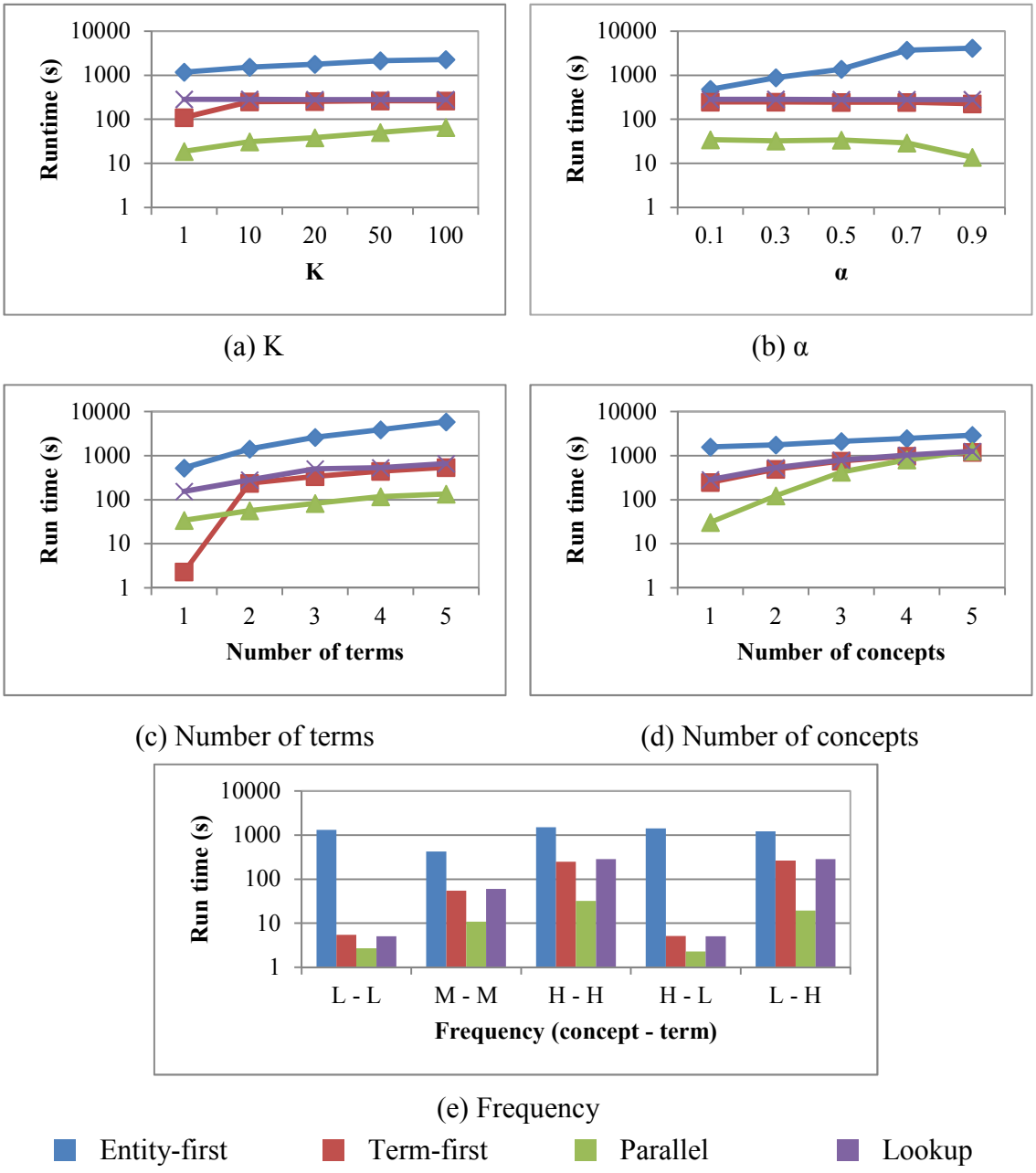

Figure 6.7: Health Web forums time performance

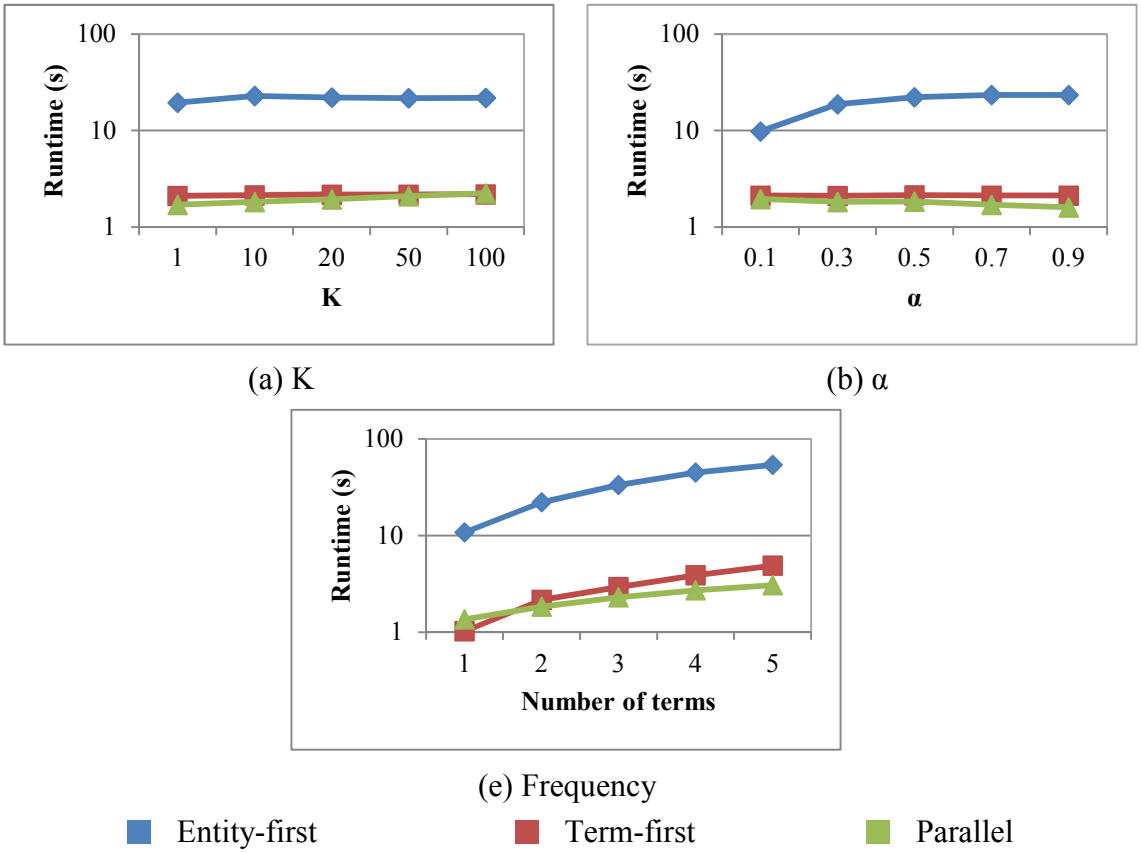

Figure 6.8: Twitter time performance

6.7 Discussion

In summary, our experiments show that:

- Our proposed *KL-based* method to compute α parameters, and hence balance the importance of the graph entities relevance and the text relevance, achieved high precision comparing to all other methods.
- In terms of time performance, Parallel algorithm performed better comparing to Term-first and Entity-first algorithms, except when k=1 where Term-first performed better.

The rationale behind the superior performance of Parallel is its non-random access nature, which yields less disk accesses and hence faster running time despite the number of documents it processes. On the other hand, Entity-first algorithm accesses disk n times for every new document added to the candidates, where n is the number of query keywords; thus, increasing the running time significantly

6.8 Conclusion

In this work, we proposed a novel query framework for querying collections of graph-annotated documents, which we refer as keyword querying on graph-annotated documents. Our method automatically balances the importance of the graph entities relevance and the text content relevance. We presented several indexing schemes and early termination algorithms to generate the *top-k* results. Our qualitative experiments show that the *KL-based* method achieved an average accuracy improvement of 60% over baselines. Moreover, our time performance experiments show that the *Parallel* algorithm is significantly faster than the *Entity-first* and *Term-first* algorithms in most settings.

Chapter 7

Conclusion

In this thesis, we presented two key approaches for advancing research in health care informatics, and analyzing the content of health-related social media to help users efficiently explore health-related data.

In Chapter 2, we analyzed the demographic of three different types of Web-based social media: (1) general Web-based social networks, namely Google+ and Twitter; (2) drug review websites, and (3) health Web forums. We examined the following demographics attributes: gender, age, ethnicity, geographical location, and writing level., and we built and evaluated domain-specific classifiers to estimate the missing data when possible. Our findings revealed significant and unanticipated disparities of the various demographic groups' participation.

In Chapter 3, we analyzed Web-based health-related social media content in relation to the demographic data, in order to identify popular topics discussed by certain demographic groups through different social media, which will assist with guiding research and educational activities. Similar to Chapter 2, we collected data from three types of health-related social media: (1) general Web-based social networks, namely Google+ and Twitter; (2) drug review websites, and (3) health Web forums. Our analysis considered five demographic attributes: gender, age, ethnicity, location, and writing level. For each demographic attribute, we analyzed the posts' contents across different

dimensions: sentiment and emotion; top distinctive terms, and top medical concepts, including disorders and drugs. Our results can contribute to knowledge through various means, including guidance of educational initiatives, advertisement of associated products, assistance to funding agencies to better allocate resources, alongside an effective understanding of health disparities in health-related social media.

In Chapter 4, we analyzed the content of health-related social media in-depth, by classifying the intent of users in order to determine how they engage and share information across the different social media applications. We analyzed two types of health-related social media: (1) general Web-based social networks, namely Google+ and Twitter, and (2) health Web forums. For health Web forums, we identified four intents as follows: share experience, ask for advice, request/give support, and talking about family. For general Web-based social networks, we identified five additional intents: share news, jokes, ads, personal opinion, and educational materials. We used supervised learning classifier to train a randomly selected and labeled data if there were sufficient posts. The classifiers with greater accuracy were utilized to label the rest of our posts. We finally analyzed and categorized the content based on the associated demographic data when possible.

In Chapter 5, we analyzed the content associated with electronic cigarette- also known as e-cigarette- users, to better comprehend the symptoms and disorders associated with smoking e-cigarettes. For this research, we analyzed the data we collected from a reputable e-cigarette forum, and identified any stated health-related affects associated with smoking e-cigarettes. We analyzed the collected data further, by using a modified

version of MetaMap tool[13], to extract references to medical concepts, alongside a measurement of the sentiments of all posts using a supervised learning classifier.

In Chapter 6, we help users to retrieve the most relevant documents when they query a collection of documents annotated by interconnected entities. For this research, we investigated the problem where a query specifies one or more graph entities, which are related to each other through association graphs, in addition to the keywords. For example social networks, where every post is annotated with the author's ID, with all the users connected through a friendship graph. Existing research has incorporated semi-structured data, such as controlled vocabularies and knowledge bases, as a means of improving the quality of ranking by expanding the queries' entities [14]–[16]. However, in our problem the entities are provided by the users as a query component. Consequently, we proposed an original query framework, 'keyword queries on graph-annotated documents', which balances the importance of text relevance and semantic relevance.

Bibliography

[1] S. Kemp, “Digital In 2017: Global Overview,” *Hootsuite*, p. 107, 2017.

[2] S. Fox, “The social life of health information | Pew Research Center,” 2014.

[3] “Health Disparities.” [Online]. Available: http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Data-and-Systems/Health-Disparities.html. [Accessed: 07-Mar-2014].

[4] M. C. Gibbons, L. Fleisher, R. E. Slamon, S. Bass, V. Kandadai, and J. R. Beck, “Exploring the potential of Web 2.0 to address health disparities.,” *J. Health Commun.*, vol. 16, no. March 2014, pp. 77–89, Jan. 2011.

[5] F. J. Grajales, S. Sheps, K. Ho, H. Novak-Lauscher, and G. Eysenbach, “Social media: a review and tutorial of applications in medicine and health care.,” *J. Med. Internet Res.*, vol. 16, no. 2, p. e13, Jan. 2014.

[6] K. Denecke and W. Nejdl, “How valuable is medical social media data? Content analysis of the medical web,” *Inf. Sci. (Ny).*, vol. 179, no. 12, pp. 1870–1880, May 2009.

[7] Y. Lu, P. Zhang, J. Liu, J. Li, and S. Deng, “Health-related hot topic detection in online communities using text clustering.,” *PLoS One*, vol. 8, no. 2, p. e56221, Jan. 2013.

[8] M. T. Wiley, C. Jin, V. Hristidis, and K. M. Esterling, “Pharmaceutical drugs chatter on Online Social Networks.,” *J. Biomed. Inform.*, vol. 49, no. June 2014, pp. 245–254, Mar. 2014.

[9] R. D. Ravert, M. D. Hancock, and G. M. Ingersoll, “Online Forum Messages Posted by Adolescents With Type 1 Diabetes,” *Diabetes Educ.*, vol. 30, no. 5, pp. 827–834, Sep. 2004.

[10] A. D. Farmer, C. E. M. Bruckner Holt, M. J. Cook, and S. D. Hearing, “Social networking sites: a novel portal for communication,” *Postgrad. Med. J.*, vol. 85, no. 1007, pp. 455–459, Sep. 2009.

[11] C. Hawn, “Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care.,” *Health Aff. (Millwood).*, vol. 28, no. 2, pp. 361–8, 2009.

[12] Z. Cahn and M. Siegel, “Electronic cigarettes as a harm reduction strategy for tobacco control : A step forward or a repeat of past mistakes ?,” *J. Public Health Policy*, vol. 32, no. 1, pp. 16–31, 2011.

[13] a R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” *Proc. AMIA Symp.*, pp. 17–21, 2001.

[14] J. Dalton, L. Dietz, and J. Allan, “Entity Query Feature Expansion Using Knowledge Base Links,” *Proc. 37th Int. ACM SIGIR Conf. Res. & Dev. Inf. Retr.*, pp. 365–374, 2014.

[15] C. Xiong and J. Callan, “EsdRank: Connecting Query and Documents through External Semi-Structured Data,” *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag.*, pp. 951–960, 2015.

[16] A. Tonon, G. Demartini, and P. Cudré-mauroux, “Combining Inverted Indices and Structured Search for Ad-hoc Object Retrieval,” 2012.

[17] ECRI, “Social Media in Healthcare,” vol. 1, no. 610, 2011.

[18] “Giving benefits the YouTube treatment - Workplace Benefits Association,” 2008. [Online]. Available: http://www.workplacebenefits.org/news/giving-benefits-youtube-treatment-711171-1.html. [Accessed: 07-Mar-2014].

[19] T. R. Frieden, “CDC Health Disparities and Inequalities Report - United States, 2011.,” *MMWR. Surveill. Summ.*, vol. 60, pp. 1–114, Jan. 2011.

[20] B. Y. G. C. Kane, R. G. Fichman, J. Gallaugher, and J. Glaser, “Community Relations 2.0,” *Harvard Business Review*, no. November, pp. 45–50, 2009.

[21] B. A. Hackworth and M. B. Kunz, “HEALTH CARE AND SOCIAL MEDIA : BULDING RELATIONSHIPS VIA SOCIAL NETWORKS,” *Acad. Heal. Care J.*, vol. 6, no. 1, pp. 55–69, 2010.

[22] “Social Networking Fact Sheet | Pew Research Center’s Internet & American Life Project.” [Online]. Available: http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/. [Accessed: 07-Mar-2014].

[23] “Hispanic Shoppers Bring Social, Mobile Habits to the Grocery Aisle - eMarketer,” 2013. [Online]. Available: http://www.emarketer.com/Article/Hispanic-Shoppers-Bring-Social-Mobile-Habits-Grocery-Aisle/1009839. [Accessed: 07-Mar-2014].

[24] A. Mislove, S. Lehmann, and Y. Ahn, “Understanding the Demographics of Twitter Users.,” in *AAAI*, 2011, pp. 554–557.

[25] B. Mandel, A. Culotta, and J. Boulahanis, “A demographic analysis of online sentiment during hurricane irene,” in *Language in Social Media*, 2012, no. Lsm, pp. 27–36.

[26] “The Literacy Problem.” Harvard.

[27] “Twitter.” [Online]. Available: https://twitter.com/. [Accessed: 23-Apr-2014].

[28] “Google+.” [Online]. Available: https://plus.google.com/. [Accessed: 23-Apr-2014].

[29] “Drugs.com | Prescription Drug Information, Interactions & Side Effects.” [Online]. Available: http://www.drugs.com/. [Accessed: 23-Apr-2014].

[30] “Treatments: reviews of drugs, therapies and remedies by everyday people - DailyStrength.” [Online]. Available: http://www.dailystrength.org/treatments. [Accessed: 23-Apr-2014].

[31] “WebMD Drugs & Treatments - Medical Information and user ratings on prescription drugs and over-the-counter (OTC) medications.” [Online]. Available: http://www.webmd.com/drugs/index-drugs.aspx. [Accessed: 23-Apr-2014].

[32] “Medical Questions Answered - Drugs.com.” [Online]. Available: http://www.drugs.com/answers/. [Accessed: 07-Mar-2014].

[33] “Online Support Groups - DailyStrength.” [Online]. Available: http://www.dailystrength.org/support-groups. [Accessed: 23-Apr-2014].

[34] “WebMD - Better information. Better health.” [Online]. Available: http://www.webmd.com/. [Accessed: 07-Mar-2014].

[35] “THE FLESCH GRADE LEVEL READABILITY FORMULA.” [Online]. Available: http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php. [Accessed: 30-Apr-2014].

[36] “RxList - The Internet Drug Index for prescription drugs, medications and pill identifier.” [Online]. Available: http://www.rxlist.com/script/main/hp.asp. [Accessed: 07-Mar-2014].

[37] “The Streaming APIs | Twitter Developers.” [Online]. Available: https://dev.twitter.com/docs/streaming-apis. [Accessed: 07-Mar-2014].

[38] “Google+ API - Google+ Platform — Google Developers.” [Online]. Available: https://developers.google.com/+/api/. [Accessed: 07-Mar-2014].

[39] J. Hedley, “jsoup Java HTML Parser, with best of DOM, CSS, and jquery.” [Online]. Available: https://jsoup.org/. [Accessed: 11-Feb-2017].

[40] “Top Names Over the Last 100 Years.” [Online]. Available: http://www.ssa.gov/oact/babynames/decades/century.html. [Accessed: 07-Mar-2014].

[41] U.S. Census Bureau Demographic Internet Staff, “Age and Sex Composition in the United States: 2012.”

[42] U.S. Census Bureau Demographic Internet Staff, “Computer and Internet Access in the United States: 2012.”

[43] US Census Bureau Data Integration Division, “Genealogy Data - Frequently Occurring Surnames from Census 2000 - U.S. Census Bureau.”

[44] C. Smith, “User Demographics of the Major Social Networks [Infographic],” 2012. [Online]. Available: http://expandedramblings.com/index.php/user-demographics-of-the-major-social-networks-infographic/#.U0T1CvldVsK. [Accessed: 09-Apr-2014].

[45] “April 2011: The Days of Double-Digit Growth in Social Network Users Are Over.” [Online]. Available: http://www.iab.net/research/industry_data_and_landscape/1675/1644724. [Accessed: 09-Apr-2014].

[46] J. Ashley, “BBC News - Women ‘more likely to report ill health than men,’” 2010. [Online]. Available: http://news.bbc.co.uk/2/hi/health/8588686.stm. [Accessed: 09-Apr-2014].

[47] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors,” in *WWW*, 2010, pp. 591–600.

[48] Q. Gu, C. F. Dillon, and V. L. Burt, “Prescription Drug Use Continues to Increase: U.S. Prescription Drug Data for 2007-2008,” *NCHS Data Brief*, vol. 42, no. September, 2010.

[49] “The comprehensive resource for Google+ trends and statistics.” [Online]. Available: http://www.gplusdata.com/. [Accessed: 07-Jun-2014].

[50] U S Census Bureau, “Statistical Abstract of the United States: 2012,” 2012.

[51] S. Bennett, “Twitter, Facebook And LinkedIn: Age, Ethnicity And Gender Of The Major Social Networks [STUDY] - AllTwitter,” 2011. [Online]. Available: http://www.mediabistro.com/alltwitter/pew-social-network-age-ethnicity-gender_b11681. [Accessed: 07-Mar-2014].

[52] A. Young, H. J. Chaudhry, J. V Thomas, and M. Dugan, “A Census of Actively Licensed Physicians in the United States , 2012,” *J. Med. Regul.*, vol. 99, pp. 11–24, 2012.

[53] C. P. Cooper, C. a Gelb, S. H. Rim, N. a Hawkins, J. L. Rodriguez, and L. Polonec, “Physicians who use social media and other internet-based communication technologies.,” *J. Am. Med. Inform. Assoc.*, vol. 19, no. 6, pp. 960–4, 2012.

[54] K. N. Hampton, L. S. Goulet, L. Rainie, and K. Purcell, “Social networking sites

and our lives," 2011.

[55] B. C. Denavas-walt, B. D. Proctor, and J. C. Smith, "Income, Poverty, and Health Insurance Coverage in the United States: 2012, U.S. Census Bureau, Current Population Reports," Washington, DC, 2013.

[56] US Census Bureau Data Integration Division, "Income."

[57] US Census Bureau, "Statistical Abstract of the United States: 2012," 2012.

[58] P. Wicks *et al.*, "Perceived benefits of sharing health data between people with epilepsy on an online platform.," *Epilepsy Behav. E&B*, vol. 23, no. 1, pp. 16–23, Jan. 2012.

[59] "Disparities in Healthcare Quality Among Racial and Ethnic Minority Groups | Agency for Healthcare Research & Quality (AHRQ)." [Online]. Available: http://www.ahrq.gov/research/findings/nhqrdr/nhqrdr10/minority.html. [Accessed: 14-Jun-2015].

[60] N. E. Adler and K. Newman, "Socioeconomic Disparities In Health: Pathways And Policies," *Health Aff.*, vol. 21, no. 2, pp. 60–76, Mar. 2002.

[61] "Technology use by different income groups | Pew Research Center." [Online]. Available: http://www.pewinternet.org/2013/05/29/technology-use-by-different-income-groups/. [Accessed: 14-Jun-2015].

[62] C. Ryan, "Language Use in the United States: 2011, U.S. Census Bureau," 2013.

[63] S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving, "A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication.," *J. Med. Internet Res.*, vol. 15, no. 4, p. e85, Jan. 2013.

[64] S. Fox and S. Jones, "The Social Life of Health Information | Pew Research Center." [Online]. Available: http://www.pewinternet.org/2009/06/11/the-social-life-of-health-information/. [Accessed: 26-Aug-2015].

[65] W. S. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse, "Social media use in the United States: implications for health communication.," *J. Med. Internet Res.*, vol. 11, no. 4, p. e48, Jan. 2009.

[66] E. Z. Kontos, K. M. Emmons, E. Puleo, and K. Viswanath, "Communication inequalities and public health implications of adult social networking site use in the United States.," *J. Health Commun.*, vol. 15 Suppl 3, pp. 216–35, Jan. 2010.

[67] "Unified Medical Language System (UMLS) - Home." U.S. National Library of Medicine.

[68] J. S. Lou and B. N. Smith, “Social Networking, Health 2.0, and Beyond,” in *Information Technology Essentials for Behavioral Health Clinicians*, pp. 119–131.

[69] P. M. Krueger, M. K. Tran, R. A. Hummer, and V. W. Chang, “Mortality Attributable to Low Levels of Education in the United States.,” *PLoS One*, vol. 10, no. 7, p. e0131809, Jan. 2015.

[70] Y. Liu, S. Xu, H.-J. Yoon, and G. Tourassi, “Extracting patient demographics and personal medical information from online health forums.,” *AMIA Annu. Symp. Proc.*, vol. 2014, pp. 1825–34, Jan. 2014.

[71] Y. Liu, S. Xu, H. Yoon, and G. Tourassi, “Extracting Patient Demographics and Personal Medical Information from Online Health Forums,” pp. 1825–1834, 1825.

[72] E. S. Anderson-Bill, R. A. Winett, and J. R. Wojcik, “Social cognitive determinants of nutrition and physical activity among web-health users enrolling in an online intervention: the influence of social support, self-efficacy, outcome expectations, and self-regulation.,” *J. Med. Internet Res.*, vol. 13, no. 1, p. e28, Jan. 2011.

[73] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” *Proc. Seventh Int. Conf. Lang. Resour. Eval.*, vol. 0, pp. 2200–2204, 2010.

[74] S. M. Mohammad and P. D. Turney, “Crowdsourcing a Word – Emotion Association Lexicon,” *Comput. Intell.*, vol. 59, no. 0, pp. 1–24, 2011.

[75] S. Mohammad and T. Yang, “Tracking Sentiment in Mail: How Genders Differ on Emotional Axes,” *Proc. ACL 2011 Work. Comput. Approaches to Subj. Sentim. Anal.*, pp. 70–79, 2011.

[76] B. Ohana and B. Tierney, “Sentiment classification of reviews using SentiWordNet,” *9th. IT T Conf.*, 2009.

[77] K. Toutanova, D. Klein, and C. D. Manning, “Feature-rich part-of-speech tagging with a cyclic dependency network,” *Proc. 2003 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Vol. 1 (NAACL ’03),* no. June, pp. 252–259, 2003.

[78] “Porter Stemming Algorithm.” [Online]. Available: http://tartarus.org/martin/PorterStemmer/. [Accessed: 11-Feb-2015].

[79] O. Bodenreider, “The Unified Medical Language System (UMLS): integrating biomedical terminology.,” *Nucleic Acids Res.*, vol. 32, no. July 2003, pp. D267–D270, 2004.

[80] K. Denecke and N. Soltani, “Where humans meet machines: Innovative solutions

for knotty natural-language problems," *Where Humans Meet Mach. Innov. Solut. Knotty Nat. Probl.*, pp. 1–315, 2013.

[81] L. J. Akinbami *et al.*, "Attention Deficit Hyperactivity Disorder Among Children Aged 5 – 17 Years in the United States , 1998 – 2009," *NCHS Data Brief*, no. 70, pp. 1–8, 2011.

[82] "Multiple Sclerosis Statistics | Statistic Brain." [Online]. Available: http://www.statisticbrain.com/multiple-sclerosis-statistics/. [Accessed: 28-Jan-2015].

[83] "Low back pain U.S. adults by gender 1997-2013 | Statistic." [Online]. Available: http://www.statista.com/statistics/188858/adults-in-the-us-with-low-back-pain-by-gender-since-1997/. [Accessed: 17-Aug-2015].

[84] "Men | Gender | HIV by Group | HIV/AIDS | CDC." [Online]. Available: http://www.cdc.gov/hiv/group/gender/men/index.html. [Accessed: 29-Feb-2016].

[85] E. Rawes, "10 States with the Highest Marijuana Use," *Money & Career CheatSheet*. [Online]. Available: http://www.cheatsheet.com/personal-finance/10-states-with-the-highest-marijuana-use.html/. [Accessed: 14-Sep-2015].

[86] "Adult Obesity in the United States: The State of Obesity." [Online]. Available: http://stateofobesity.org/adult-obesity/. [Accessed: 21-Sep-2015].

[87] J. C. Smith and C. Medalia, "Health Insurance Coverage in the United States: 2014," *U.S. Census Bur.*, vol. Current Po, no. September, pp. 60–253, 2015.

[88] R. Whittaker, "Smoking cessation intervention for young adults using multimedia mobile phones: development and effectiveness." ResearchSpace@Auckland, 2011.

[89] D. J. Opel, D. S. Diekema, N. R. Lee, and E. K. Marcuse, "Social marketing as a strategy to increase immunization rates.," *Arch. Pediatr. Adolesc. Med.*, vol. 163, no. 5, pp. 432–7, May 2009.

[90] R. Patel, T. Chang, S. R. Greysen, and V. Chopra, "Social Media Use in Chronic Disease: A Systematic Review and Novel Taxonomy," *Am. J. Med.*, vol. 128, no. 12, pp. 1335–1350, Dec. 2015.

[91] C. G. Valle, D. F. Tate, D. K. Mayer, M. Allicock, and J. Cai, "A randomized trial of a Facebook-based physical activity intervention for young adult cancer survivors," *J. Cancer Surviv.*, vol. 7, no. 3, pp. 355–368, Sep. 2013.

[92] C. A. Maher, L. K. Lewis, K. Ferrar, S. Marshall, I. De Bourdeaudhuij, and C. Vandelanotte, "Are health behavior change interventions that use online social networks effective? A systematic review.," *J. Med. Internet Res.*, vol. 16, no. 2, p. e40, Jan. 2014.

[93] A. Sadilek, H. Kautz, and V. Silenzio, “Modeling Spread of Disease from Social Interactions,” *Proc. Sixth Int. AAAI Conf. Weblogs Soc. Media*, pp. 322–329, 2012.

[94] A. Nikfarjam, A. Sarker, K. O’Connor, R. Ginn, and G. Gonzalez, “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features,” *J. Am. Med. Informatics Assoc.*, Mar. 2015.

[95] J. C. Eichstaedt *et al.*, “Psychological Language on Twitter Predicts County-Level Heart Disease Mortality,” *Psychol. Sci.*, vol. 26, no. 2, pp. 159–169, Feb. 2015.

[96] S. A. Sadah, M. Shahbazi, M. T. Wiley, and V. Hristidis, “A Study of the Demographics of Web-Based Health-Related Social Media Users.,” *J. Med. Internet Res.*, vol. 17, no. 8, p. e194, Jan. 2015.

[97] E. Frank, M. A. Hall, and I. H. Witten, “The WEKA Workbench,” in *Data Mining: Practical Machine Learning Tools and Techniques*, no. Fourth Edition, 2016.

[98] M. Williams, A. Villarreal, B. Davis, and P. Talbot, “Comparison of the Performance of Cartomizer Style Electronic Cigarettes from Major Tobacco and Independent Manufacturers,” *PLoS One*, vol. 11, no. 2, p. e0149251, Feb. 2016.

[99] M. Williams, A. To, K. Bozhilov, and P. Talbot, “Strategies to Reduce Tin and Other Metals in Electronic Cigarette Aerosol,” *PLoS One*, vol. 10, no. 9, p. e0138933, Sep. 2015.

[100] M. Williams, S. Ghai, and P. Talbot, “Disposable Electronic Cigarettes and Electronic Hookahs: Evaluation of Performance,” *Nicotine Tob. Res.*, vol. 17, no. 2, pp. 201–208, Feb. 2015.

[101] M. Williams, A. Villarreal, K. Bozhilov, S. Lin, and P. Talbot, “Metal and Silicate Particles Including Nanoparticles Are Present in Electronic Cigarette Cartomizer Fluid and Aerosol,” *PLoS One*, vol. 8, no. 3, p. e57987, Mar. 2013.

[102] M. Williams and P. Talbot, “Variability Among Electronic Cigarettes in the Pressure Drop, Airflow Rate, and Aerosol Production,” *Nicotine Tob. Res.*, vol. 13, no. 12, pp. 1276–1283, Dec. 2011.

[103] M. Hua, M. Alfi, and P. Talbot, “Health-Related Effects Reported by Electronic Cigarette Users in Online Forums,” *J. Med. Internet Res.*, vol. 15, no. 4, p. e59, Apr. 2013.

[104] R. Polosa *et al.*, “Effectiveness and tolerability of electronic cigarette in real-life: a 24-month prospective observational study,” *Intern. Emerg. Med.*, vol. 9, no. 5, pp. 537–546, Aug. 2014.

[105] C. I. Vardavas, N. Anagnostopoulos, M. Kougias, V. Evangelopoulou, G. N. Connolly, and P. K. Behrakis, "Short-term Pulmonary Effects of Using an Electronic Cigarette," *Chest*, vol. 141, no. 6, pp. 1400–1406, Jun. 2012.

[106] V. Bahl, S. Lin, N. Xu, B. Davis, Y. Wang, and P. Talbot, "Comparison of electronic cigarette refill fluid cytotoxicity using embryonic and adult models," *Reprod. Toxicol.*, vol. 34, no. 4, pp. 529–537, Dec. 2012.

[107] R. Z. Behar, B. Davis, Y. Wang, V. Bahl, S. Lin, and P. Talbot, "Identification of toxicants in cinnamon-flavored electronic cigarette refill fluids," *Toxicol. Vitr.*, vol. 28, no. 2, pp. 198–208, Mar. 2014.

[108] R. Z. Behar, B. Davis, V. Bahl, S. Lin, and P. Talbot, "Commentary in response to the letter from Farsalinos et al. regarding our publication entitled: 'Identification of toxicants in cinnamon-flavored electronic cigarette refill fluids,'" *Toxicol. Vitr.*, vol. 28, no. 8, pp. 1521–1522, Dec. 2014.

[109] V. Yu *et al.*, "Electronic cigarettes induce DNA strand breaks and cell death independently of nicotine in cell lines," *Oral Oncol.*, vol. 52, pp. 58–65, Jan. 2016.

[110] C. A. Lerner *et al.*, "Vapors Produced by Electronic Cigarettes and E-Juices with Flavorings Induce Toxicity, Oxidative Stress, and Inflammatory Response in Lung Epithelial Cells and in Mouse Lung," *PLoS One*, vol. 10, no. 2, p. e0116732, Feb. 2015.

[111] T. E. Sussan *et al.*, "Exposure to Electronic Cigarettes Impairs Pulmonary Anti-Bacterial and Anti-Viral Defenses in a Mouse Model," *PLoS One*, vol. 10, no. 2, p. e0116861, Feb. 2015.

[112] K. S. Schweitzer *et al.*, "Endothelial disruptive proinflammatory effects of nicotine and e-cigarette vapor exposures," *Am. J. Physiol. - Lung Cell. Mol. Physiol.*, vol. 309, no. 2, pp. L175–L187, Jul. 2015.

[113] G. Eysenbach, "Infodemiology: The epidemiology of (mis)information.," *Am. J. Med.*, vol. 113, no. 9, pp. 763–5, Dec. 2002.

[114] G. Eysenbach, "Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet," *J. Med. Internet Res.*, vol. 11, no. 1, p. e11, Mar. 2009.

[115] G. Eysenbach, "Infodemiology and Infoveillance," *Am. J. Prev. Med.*, vol. 40, no. 5, pp. S154–S158, May 2011.

[116] M. Hua, H. Yip, and P. Talbot, "Mining data on usage of electronic nicotine delivery systems (ENDS) from YouTube videos," *Tob. Control*, vol. 22, no. 2, pp. 103–106, Mar. 2013.

[117] R. Z. Behar, M. Hua, and P. Talbot, “Puffing Topography and Nicotine Intake of Electronic Cigarette Users,” *PLoS One*, vol. 10, no. 2, p. e0117222, Feb. 2015.

[118] M. Hua, M. Alfi, and P. Talbot, “Health-related effects reported by electronic cigarette users in online forums.,” *J. Med. Internet Res.*, vol. 15, no. 4, p. e59, Apr. 2013.

[119] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” *Proceedings. AMIA Symp.*, pp. 17–21, 2001.

[120] O. Bodenreider, “The Unified Medical Language System (UMLS): integrating biomedical terminology,” *Nucleic Acids Res.*, vol. 32, no. 90001, p. 267D–270, Jan. 2004.

[121] O. Bodenreider and A. T. McCray, “Exploring semantic groups through visual approaches.,” *J. Biomed. Inform.*, vol. 36, no. 6, pp. 414–32, Dec. 2003.

[122] D. Freelon, “ReCal: reliability calculation for the masses - dfreelon.org.” [Online]. Available: http://dfreelon.org/utils/recalfront/. [Accessed: 11-Feb-2017].

[123] Y. Sun, A. K. C. Wong, M. Kamel, and S., “Classification of Imbalanced Data: a Review,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[124] C. Pisinger and M. Døssing, “A systematic review of health effects of electronic cigarettes,” *Prev. Med. (Baltim).*, vol. 69, pp. 248–260, Dec. 2014.

[125] M. Hua and P. Talbot, “Potential health effects of electronic cigarettes: A systematic review of case reports,” *Prev. Med. Reports*, vol. 4, pp. 169–178, Dec. 2016.

[126] R. S. Moheimani *et al.*, “Increased Cardiac Sympathetic Activity and Oxidative Stress in Habitual Electronic Cigarette Users,” *JAMA Cardiol.*, vol. 2, no. 3, p. 278, Mar. 2017.

[127] A. Bhatnagar, “Are Electronic Cigarette Users at Increased Risk for Cardiovascular Disease?,” *JAMA Cardiol.*, vol. 2, no. 3, p. 237, Mar. 2017.

[128] T. Kivrak, M. Sunbul, E. Durmus, R. Dervisova, I. Sari, and O. Yesildag, “Acute myocardial infarction due to liquid nicotine in a young man,” *Ther. Adv. Cardiovasc. Dis.*, vol. 8, no. 1, pp. 32–34, Feb. 2014.

[129] “E-Cigarette Use Among Youth and Young Adults. A Report of the Surgeon General,” 2016.

[130] K. Chatham-Stephens *et al.*, “Notes from the field: calls to poison centers for exposures to electronic cigarettes--United States, September 2010-February

2014.," *MMWR. Morb. Mortal. Wkly. Rep.*, vol. 63, no. 13, pp. 292–3, Apr. 2014.

[131] S. Lee, S. Taleban, S. Targan, and G. Melmed, "E-cigarettes as Salvage Therapy for Medically Refractory Ulcerative Colitis," *Inflamm. Bowel Dis.*, vol. 19, p. S99, Dec. 2013.

[132] M. Camus, G. Gallois, and P. Marteau, "Ulcerative Colitis and Electronic Cigarette: What's the Matter?," *Am. J. Gastroenterol.*, vol. 109, no. 4, pp. 608–609, Apr. 2014.

[133] J. M. Rogér, M. Abayon, S. Elad, and A. Kolokythas, "Oral Trauma and Tooth Avulsion Following Explosion of E-Cigarette," *J. Oral Maxillofac. Surg.*, vol. 74, no. 6, pp. 1181–1185, Jun. 2016.

[134] "Evaluation of e-cigarettes," St. Louis, MO 63101, 2009.

[135] V. Varlet, K. Farsalinos, M. Augsburger, A. Thomas, and J.-F. Etter, "Toxicity Assessment of Refill Liquids for Electronic Cigarettes," *Int. J. Environ. Res. Public Health*, vol. 12, no. 5, pp. 4796–4815, Apr. 2015.

[136] J. G. Allen *et al.*, "Flavoring Chemicals in E-Cigarettes: Diacetyl, 2,3-Pentanedione, and Acetoin in a Sample of 51 Products, Including Fruit-, Candy-, and Cocktail-Flavored E-Cigarettes," *Environ. Health Perspect.*, vol. 124, no. 6, Dec. 2015.

[137] R. P. Jensen, W. Luo, J. F. Pankow, R. M. Strongin, and D. H. Peyton, "Hidden Formaldehyde in E-Cigarette Aerosols," *N. Engl. J. Med.*, vol. 372, no. 4, pp. 392–394, Jan. 2015.

[138] G. Cervellin, M. Luci, C. Bellini, and G. Lippi, "Bad news about an old poison. A case of nicotine poisoning due to both ingestion and injection of the content of an electronic cigarette refill," *Emerg. Care J.*, vol. 9, no. 2, p. 18, Oct. 2013.

[139] C. K. Eberlein, H. Frieling, T. Köhnlein, T. Hillemacher, and S. Bleich, "Suicide Attempt by Poisoning Using Nicotine Liquid For Use in Electronic Cigarettes," *Am. J. Psychiatry*, vol. 171, no. 8, pp. 891–891, Aug. 2014.

[140] E. M. Schipper *et al.*, "A new challenge: suicide attempt using nicotine fillings for electronic cigarettes," *Br. J. Clin. Pharmacol.*, vol. 78, no. 6, pp. 1469–1471, Dec. 2014.

[141] S. L. Thornton, L. Oller, and T. Sawyer, "Fatal Intravenous Injection of Electronic Nicotine Delivery System Refilling Solution," *J. Med. Toxicol.*, vol. 10, no. 2, pp. 202–204, Jun. 2014.

[142] S. Bartschat, K. Mercer-Chalmers-Bender, J. Beike, M. A. Rothschild, and M. Jübner, "Not only smoking is deadly: fatal ingestion of e-juice—a case report," *Int.*

J. Legal Med., vol. 129, no. 3, pp. 481–486, May 2015.

[143] L. M Jablow and R. J Sexton, "Spontaneous Electronic Cigarette Explosion: A Case Report," *Am. J. Med. Case Reports*, vol. 3, no. 4, pp. 93–94, Feb. 2015.

[144] L. McCauley, C. Markin, and D. Hosmer, "An Unexpected Consequence of Electronic Cigarette Use," *Chest*, vol. 141, no. 4, pp. 1110–1113, Apr. 2012.

[145] J. Hureaux, M. Drouet, and T. Urban, "A case report of subacute bronchial toxicity induced by an electronic cigarette: Table 1," *Thorax*, vol. 69, no. 6, pp. 596–597, Jun. 2014.

[146] D. Thota and E. Latham, "Case Report of Electronic Cigarettes Possibly Associated with Eosinophilic Pneumonitis in a Previously Healthy Active-duty Sailor," *J. Emerg. Med.*, vol. 47, no. 1, pp. 15–17, Jul. 2014.

[147] G. Atkins and F. Drescher, "Acute Inhalational Lung Injury Related to the Use of Electronic Nicotine Delivery System (ENDS)," *Chest*, vol. 148, no. 4, p. 83A, Oct. 2015.

[148] S. Modi, R. Sangani, and A. Alhajhusain, "Acute Lipoid Pneumonia Secondary to E-Cigarettes Use: An Unlikely Replacement for Cigarettes," *Chest*, vol. 148, no. 4, p. 382A–382B, Oct. 2015.

[149] K. Moore, H. Young II, and M. F. Ryan, "Bilateral Pneumonia and Pleural Effusions Subsequent to Electronic Cigarette Use," *Open J. Emerg. Med.*, vol. 3, no. 3, pp. 18–22, 2015.

[150] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet Web Inf. Syst.*, vol. 54, no. 1999–66, pp. 1–17, 1998.

[151] V. Hristidis and Y. Papakonstantinou, "DISCOVER: Keyword Search in Relational Databases," *VLDB*, vol. pages, pp. 670–681, 2002.

[152] A. Balmin, H. Vagelis, and P. Yannis, "ObjectRank : Authority-Based Keyword Search in Databases," *Proc. Thirtieth Int. Conf. Very large data bases (VLDB '04)*, vol. 30, pp. 564–575, 2004.

[153] S. Agrawal, S. Chaudhuri, and G. Das, "DBXplorer: a system for keyword-based search over relational databases," *Proc. 18th Int. Conf. Data Eng.*, pp. 5–16, 2002.

[154] D. Carmel *et al.*, "XRANK: Ranked Keyword Search over XML Documents," ... *Work. XML Inf. Retr.*, vol. 3, no. 1, pp. 1–9, 2005.

[155] H. Wasserman and J. Wang, "An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list.," *AMIA Annu. Symp.*

Proc., pp. 699–703, 2003.

[156] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," *SIGIR '96 Proc. ACM SIGIR Conf.*, vol. 19, p. 4, 1996.

[157] F. Farfán, V. Hristidis, A. Ranganathan, and M. Weiner, "XOntoRank: Ontology-aware search of electronic medical records," *Proc. - Int. Conf. Data Eng.*, pp. 820–831, 2009.

[158] A. Arvanitis, M. Wiley, and V. Hristidis, "Efficient Concept-based Document Ranking," *17th Int. Conf. Extending Database Technol.*, no. c, pp. 403–414, 2014.

[159] F. Nikolaev, A. Kotov, and N. Zhiltsov, "Parameterized Fielded Term Dependence Models for Ad-hoc Entity Retrieval from Knowledge Graph," *Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 435–444, 2016.

[160] I. F. Ilyas, G. Beskales, and M. a. Soliman, "A survey of top-k query processing techniques in relational database systems," *ACM Comput. Surv.*, vol. 40, no. 4, p. 11:1--11:58, 2008.

[161] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," *J. Comput. Syst. Sci.*, vol. 66, no. 4, pp. 614–656, 2003.

[162] R. Bonaque, B. Cautis, F. Goasdoué, and I. Manolescu, "Toward social, structured and semantic search," *CEUR Workshop Proc.*, vol. 1310, 2014.

[163] S. Maniu and B. Cautis, "Taagle : Efficient, Personalized Search in Collaborative Tagging Networks Categories and Subject Descriptors," pp. 661–664.

[164] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 1, pp. 17–30, 1989.

[165] W. B. Croft, D. Metzler, and T. Strohman, "Search Engines: Information Retrieval in Practice," *Comput. J.*, pp. 1–542, 2009.

[166] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[167] T. Akiba, Y. Iwata, and Y. Yoshida, "Fast exact shortest-path distance queries on large networks by pruned landmark labeling," *Proc. 2013 Int. Conf. Manag. data - SIGMOD '13*, p. 349, 2013.