Summer 2017

Detecting user demographics in twitter to inform health trends in social media

Christopher R. Markson
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

DETECTING USER DEMOGRAPHICS IN TWITTER TO INFORM HEALTH TRENDS IN SOCIAL MEDIA

by
Christopher R. Markson

The widespread and popular use of social media and social networking applications offer a promising opportunity for gaining knowledge and insights regarding population health conditions thanks to the diversity and abundance of online user-generated information (UGHI) relating to healthcare and well-being. However, users on social media and social networking sites often do not supply their complete demographic information, which greatly undermines the value of the aforementioned information for health 2.0 research, e.g., for discerning disparities across population groups in certain health conditions. To recover the missing user demographic information, existing methods observe a limited scope of user behaviors, such as word frequencies exhibited in a user's messages, leading to sub-optimal results.

To address the above limitation and improve the performance of inferring missing user demographic information for health 2.0 research, this work proposes a new algorithmic method for extracting a social media user's gender by exploring and exploiting a comprehensive set of a user's behaviors on Twitter, including the user's conversational topic choices, account profile information, and personal information. In addition, this work explores the usage of synonym expansion for detecting social media users' ethnicities. To better capture a user's conversational topic choices using standardized hashtags for consistent comparison, this work additionally introduces a new method that automatically generates standardized hashtags for tweets. Even though Twitter is selected as the experimental platform in this study due to its leading position among today's social networking sites, the proposed method is in principle

generically applicable to other social media sites and applications as long as there is a way to access user-generated content on those platforms.

When comparing the multi-perspective learning method with the state-of-the-art approaches for gender classification, a gender classification accuracy is observed of 88.6% for the proposed approach compared with 63.4% performance for bag-of-words and 61.4% for the peer method. Additionally, the topical approach introduced in this work outperforms vocabulary-based approach with a smaller dimensionality at 69.4% accuracy.

Furthermore, observable usage patterns of the cancer terms are analyzed across the ethnic groups inferred by the proposed algorithmic approaches. Variations among demographic groups are seen in the frequency of term usage during months known to be labeled as cancer awareness months. This work introduces methods that have the potential to serve as a very powerful and important tool in disseminating critical prevention, screening, and treatment messages to the community in real time. Study findings highlight the potential benefits of social media as a tool for detecting demographic differences in cancer-related discussions on social media.

DETECTING USER DEMOGRAPHICS IN TWITTER TO INFORM HEALTH TRENDS IN SOCIAL MEDIA

by
Christopher R. Markson

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Information Systems

Department of Informatics

August 2017

APPROVAL PAGE

DETECTING USER DEMOGRAPHICS IN TWITTER TO INFORM HEALTH TRENDS IN SOCIAL MEDIA

Christopher R. Markson

Dr. Songhua Xu, Dissertation Advisor — Date
Assistant Professor of Information Systems, New Jersey Institute of Technology

---

Dr. Yi-fang Brook Wu, Committee Member — Date
Associate Professor of Information Systems, New Jersey Institute of Technology

---

Dr. Michael Bieber, Committee Member — Date
Professor of Information Systems, New Jersey Institute of Technology

---

Dr. NhatHai Phan, Committee Member — Date
Assistant Professor of Information Systems, New Jersey Institute of Technology

---

Dr. Linda Kaufman, Committee Member — Date
Retired Professor, William Paterson University

Dr. Reethi Iyengar, Committee Member — Date
Associate Director of Global Health Economics and Outcomes Research - Oncology, Abbvie, Inc.

BIOGRAPHICAL SKETCH

Author: Christopher R. Markson

Degree: Doctor of Philosophy

Date: August 2017

Undergraduate and Graduate Education:

- Doctor of Philosophy in Information Systems,
  New Jersey Institute of Technology, Newark, NJ, 2017
- Master of Science in Information Systems,
  New Jersey Institute of Technology, Newark, NJ, 2012
- Bachelor of Science in Computer Science,
  William Paterson University, Wayne, NJ, 2006

Major: Information Systems

Presentations and Publications:

Xu, S., Markson, C., Costello, K. L., Xing, C. Y., Demissie, K., and Llanos, A. A., "Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter," *Journal of Medical Internet Research, Public Health and Surveillance*, (2)1, e17, 2016

Markson, C. and Xu, S., "A New Visualization Tool for Exploring Biomedical Patent Literature," *American Medical Informatics Association Annual Symposium*, 2014

Markson, C. and Song, M., "A Technique for Suggesting Related Wikipedia Articles using Link Analysis," *12th Association of Computing Machinery/Institute of Electrical and Electronics Engineers-Computer Science joint conference on Digital Libraries*, ACM, 2012

I would like to dedicate this work to my loving husband, Ilias Siafakas. Without you, I would be lost.

ACKNOWLEDGMENT

First and foremost, I would like to thank my dissertation advisor, Dr. Songhua Xu, for his knowledge and guidance as I pursued this research. The lessons he has taught me will remain with me for a lifetime.

I would also like to sincerely thank Dr. Yi-fang Brook Wu for her constant encouragement and sage advice. Her active engagement in my education, from the earliest parts of my graduate studies in 2012, will never be forgotten. She remained a constant source of academic advice, domain knowledge, and unique perspectives throughout this process. In addition to helping me throughout the dissertation process, she has taught me how to be an educator, how to work with students, and how to deliver effective lectures. I hold her opinion in the highest regard, as a great academic scholar, a friend, and an educator.

A few sentences could not possibly express my gratitude to Dr. Linda Kaufman for her assistance in my education. She first introduced me to research as an undergraduate student in 2003, a project that I will never forget. Without her involvement in my education, I might not have considered graduate school as an option. She a never-ending source of knowledge, a brilliant researcher, and always willing to help her students in any way possible.

Dr. Reethi Iyengar has been a fantastic colleague to whom I am forever grateful. She welcomed me to a research group during my first internship where she provided business savvy guidance and constant reassurance. I thank her for putting in the time and effort to participate in my dissertation work.

Thank you to Dr. Michael Bieber for also serving on this committee. He too has provided academic guidance throughout the pursuit of my degree. Lastly, I would like to thank Dr. NhatHai Phan, who recently joined the Information System department. I sincerely appreciate the expertise he has contributed to this work.

There are countless individuals who have helped me throughout this process, all too many to list in this acknowledgement section. Regina Collins, Ye Xiong, Chao Xu, Mingzhu Zhu, Julia Mayer, and Rich Schuler are all wonderful friends who often helped to take the sting out of those difficult research days. I would specifically like to thank Regina Collins. Without her, this already difficult process would have been nearly impossible.

Finally and most importantly, I could not have accomplished this without the support of my husband and family. Ilias, you have been with me from the first moment of this endeavor. Your unwavering support and love can not be expressed in words. Mom and Dad, you taught me to always dream big and encouraged me every step of the way. It goes without saying, but family is everything.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF SYMBOLS

A	Adjective
API	Application Programming Interface
BCAM	Breast Cancer Awareness Month
$\boldsymbol{\mathcal{BOW}}$	Bag-of-words representation of a user's timeline
$\boldsymbol{\mathcal{BUR}}$	Binary document-term matrix comprised of text from the *i-th* user's posts, profile description, and name in a 5-character gram expansion
$\{\mathcal{CH}_1, ..., \mathcal{CH}_k\}$	Candidate hashtag set
$\boldsymbol{\mathcal{CH}}(u_i)$	Computed values over $\mathcal{CH}$ for u_i
CRCAM	Colorectal Cancer Awareness Month
D	Number of tweets in the collection
d_k	k-th tweet in the collection
GPS	Global Positioning System
HG	Hashtag-gram matrix indicating the relationship between a given hashtag label and the tweet-extracted grams
HT	Hashtag
ID	Identification
LCAM	Lung Cancer Awareness Month
LDA	Latent Dirichlet Allocation
$M_{i,j}$	Hashtag co-occurrence matrix
$M_{i,j}^{\text{norm}}$	Normalized hashtag co-occurrence matrix
N	Lexical noun

LIST OF SYMBOLS

(Continued)

N	Number of hashtags appearing in multi-hashtag tweets
$\boldsymbol{\mathcal{NM}}(u_i)$	Computed gender probability for u_i's name
P	Preposition
PCAM	Prostate Cancer Awareness Month
PMI	Point-wise mutual information
$\boldsymbol{\mathcal{PV}}(u_i)$	Extracted color profile for u_i
$\boldsymbol{\mathcal{S}}(i)$	i-th seed topic
SES	Socioeconomic Status
SVM	Support Vector Machine
t_i	i-th most frequently appearing hashtag in the collection
$\boldsymbol{\mathcal{TV}}_{\mathrm{R}}(u_i)$	The distribution of a user's annotated hashtags over $\boldsymbol{\mathcal{S}}(i)$ according to the occurrence counts of these hashtags in the user's timeline
$\boldsymbol{\mathcal{TV}}(u_i)$	Generated hashtags over $\mathcal{TV}$ for u_i
$TW(u_i)$	The set of tweets by user u_i
u_i	i-th user in the collection
α	Propagation dampening parameter
β	Thresholding parameter
γ	Weighting parameter
$\theta()$	Noise reduction function

LIST OF SYMBOLS

(Continued)

$\biguplus$	Vector sum of two topic distributions
$\|\|$	Concatenation

CHAPTER 1

INTRODUCTION

The rapid development and vast progress of social media technologies and their applications in the recent years have empowered the relatively recent technological phenomenon as an emerging force that profoundly influences and revolutionizes people's daily life and societies' operations at many aspects. In 2016, Twitter received more than 500 million tweets and is visited by over a quarter of a million active users every day. Social media has played a particularly eminent role in enabling and promoting the free expression, sharing, and communication of individuals' thoughts and ideas with the general public or a selected group. Facilitated by the new technological means, users become increasingly willing to share personal information in the online space, which leads to a massive amount of publicly accessible user-generated content (UGC). Such UGC introduces an unprecedented opportunity for researchers to conduct various large-scale user studies for gaining comprehensive insights regarding a large group of users' health-related behaviors. It is noted that user demographics information is essential for studying the disparities regarding population health conditions, e.g., [1, 2, 3], and their health literacy levels, e.g., [4, 5], across various demographic groups. For this purpose, data scientists have designed various algorithms for categorizing authors of UGC based on their personal characteristics, e.g., [6, 7, 8]. However, the majority of the existing algorithms are only able to identify a user's gender information; the detection accuracy is also quite modest with the highest accuracy for gender detection merely 61% with a comparable training set size.

1.1 Motivation

The significance of this work stems from the mass adoption of social media and limited research into the extraction of demographic information of users [9, 10, 11, 12]. As is often the case with the Health 2.0 phenomenon, the availability of health related information (i.e., health-related tweets) strongly benefits from robust methods for understanding context (i.e., user demographics). The extraction of a vast user demography in social media can provide health workers with new ways of understanding immediate health issues, disease propagation, and health literacy among given groups within society. Understanding the social media's perspective on health can provide health workers with real-time data and reduce the reliance on time consuming surveys.

The ability to extract demographic components based on text and language generated by social media activity has implications in many areas, including medical demography and health recommendation systems. In this work, we focus primarily on detection of gender as a demographic component for the purpose of understanding health trends within each gender group. Gender, as a demographic component, was selected due to its strong influence on many aspects of health. Gender has been a predominant focus in disease risk [13, 14, 15, 16], is widely considered essential to individualized health strategies [17, 18], and is fundamental to productive health education [19, 20]. Understanding gender over social media can inform these research domains by 1) understanding the frequency with which gender groups participate in at-risk disease discussions via social media, and 2) utilizing the gender-specific findings in health education and health strategies research to appropriately target educational messages. In addition, this work explores the extraction of ethnicity information from Twitter profiles.

1.2 Impact on Health

Cancer is a major public health problem, impacting more than 14 million men and women in the U.S. as of January 2014, with an estimated 1.6 million additional new cancer cases being diagnosed among Americans in 2015 [21]. African Americans have experienced higher age-adjusted mortality rates when compared with Caucasians [21, 22]. Many factors contribute to these disparities. Socioeconomic status (SES) as a whole, along with its primary components, including education, income, employment status, and neighborhood appear to be obvious correlates of cancer mortality disparities [23, 24, 25]; however, other factors that are not clearly understood may also play a role [22, 26, 27]. One important factor that could particularly contribute to improved cancer prevention and thereby possibly reduce cancer disparities is knowledge and awareness about cancer.

Knowledge and awareness about the four cancers with the highest incidence and mortality among adults in the U.S., namely lung, breast, prostate, and colorectal cancer, has been shown to differ by race/ethnicity [28, 29, 30, 31, 32, 33, 34, 35, 36, 37]. Lung cancer is a good example of these differences. It is widely known that cancer of the lung is the leading cause of cancer death in the U.S. among both men and women and that tobacco smoking is the most significant and preventable cause of the disease. However, findings from one study [31] suggested that two-thirds of U.S. women could not correctly identify lung cancer as the leading cause of cancer death, and this lack of knowledge was greatest among African American women [31]. In terms of breast cancer, evidence has shown that breast cancer knowledge also greatly varies by racial/ethnic group. One study [33] showed that African American women were generally unaware of disparities in breast cancer mortality. Furthermore, one study found that South Asian women tend to have better knowledge of age-related breast cancer risk when compared with Black and White women [34]. Knowledge and awareness about both prostate and colorectal cancers have been shown to be

low among U.S. adults overall and particularly among low SES groups [35, 36, 37, 32]. These examples highlight the importance of promoting knowledge about cancer among some segments of the U.S. population, particularly among groups with the highest cancer burden.

Social media outlets including Twitter, Facebook, and Instagram, are popular online platforms that engage in communication about any and everything, and many studies [38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51] have begun examining the importance of social media in reaching larger audiences for promotion of public health knowledge and patient advocacy. Twitter has become a very popular site and application for the exchange of health-related information. Twitter allows users (individual users and organizations) to exchange information with other users around the world in real-time, through short messages called "tweets" ($\leq$ 140 characters) posted on a given user's timeline (i.e., the chronologically ordered collection of tweets posted by a given user). Twitter also allows users to re-tweet (repost) other users' tweets, which promotes the exchange of messages to a very large number of individuals. Many healthcare agencies and public health organizations (i.e., local and national organizations and private companies) [52, 53, 41, 54, 46] use Twitter as a major online platform for health education and promotion because the majority of Twitter content is publicly available and may provide a novel source of health-related information.

Recent studies [55, 53] have touted the numerous epidemiological advantages of coupling machine learning techniques with social media mining. Marathe et al. [56] discuss the real-time possibilities of understanding disease outbreaks using social media data. Dredze et al. [55] state that geo-specific data coupled with the public forum nature of social media (which encourages the sharing of detailed information) creates new public health capabilities not previously seen. Simultaneously, advances in demographic extraction techniques and computational linguistics have allowed for

a deeper understanding of user demographics [57, 6]. In these studies, Beretta and Burger connected age and gender to linguistic patterns (often word usage). In the case of Beretta [57], user profile images manually labeled by human experts helped to verify the experimental results. Much of the demographic extraction studies have built upon studies originating in the field of psychology, connecting linguistic patterns to demographic elements of participants [58, 59]. In Colley's work [58], participants' inboxes were examined for linguistic differences differentiating the genders.

In this study, we aimed to explore differences in cancer-related tweeting by race/ethnicity, basing our work on Rickford's assertion of unique vernacular patterns amongst African-Americans [59]. Findings from this study will ultimately contribute to the development and implementation of cost-effective, prevention and dissemination strategies, delivered through social media messaging, targeting specific subgroups that would benefit from increased cancer knowledge and awareness.

1.3 Contributions

First this work proposes two approaches for using social media language patterns to infer a user's ethnicity by extending the existing bag-of-words models. Namely, 1) a text classification with synonym expansion approach, and 2) topic-based classification approach. Having validated the accuracy of the text classification with synonym expansion approach against a baseline bag-of-words method, we examine the frequency of cancer discussions online between ethnic groups.

Second, this work also introduces a novel multi-perspective approach for extracting a user demographic information, specifically examining gender. These views consist of 1) the distribution of tweeted topics as inferred by hashtag mining, 2) name information combined with external data sources, and 3) user profile information. To accommodate this approach, we also present a method for automatically proliferating hashtags across all tweets using a new hashtag clustering

approach coupled with a statistical language model. Finally, we introduce a method for combining perspectives to make a final gender assignment to a given user.

CHAPTER 2

LITERATURE REVIEW

The main objective of this dissertation is to improve our understanding of how user demographics can be inferred from the choices users make when expressing themselves online and specifically in this work, on the social media platform, Twitter. The outward expression of users comes in many forms online. This is particularly evident in social media, where the intended purpose of the platform is designed to encourage public sharing of thoughts, ideas, and opinions. Exploiting the fact that the choices users make indirectly, and sometimes unintentionally, embody the user itself, we aim to improve the current methods used to extract demographic information from online social media activity. Section 2.1 discusses the main constructs of Twitter as a social media platform. Section 2.2 introduces existing approaches using language-based features for demographic extraction. Section 2.2 presents work using profile and image-based features for demographic extraction. The remaining sections introduce the natural language processing and machine learning techniques used in this dissertation.

2.1 Key Concepts in Twitter

Social media platforms have evolved into many formats since their introduction. Facebook has seen growing success as it positions itself as a platform for connecting with friends, sharing pictures and videos, and coordinating events [60]. Other platforms have also emerged, choosing to focus on particular elements of social experiences. For example, Instagram chooses to focus on picture sharing, Vine provides users the ability to share short video clips totaling six seconds or less, while Periscope and Twitch have chosen to focus on live-broadcasting of user activities.

This dissertation focuses on Twitter as a social media platform. Similar to the social media platforms mentioned previously, Twitter has introduced characteristics unique to its platform aimed at enabling social activity between users. Twitter is known to have pioneered the micro-blogging platform, providing a new medium for users to broadcast information to their friends that they would otherwise be unlikely to share via traditional communication platforms such as email, text messaging, phone, etc. [61]. These tweets have quickly migrated from desktop generated content to mobile device generated content, often representing a immediate representation of a user's thoughts. This is contrary to traditional blog posts, which are heavily edited prior to posting. In addition, Twitter encourages an open platform for users to publicly share their tweets to the wider Twitter audience. This is evident in accounts set to publicly share content by default. Additionally, Twitter has increasingly served as a news aggregator for users to gather current event information. The combination of information-seeking activities, live broadcasting, and public sharing of data makes Twitter an ideal platform for exploring health trends. Similar benefits are not seen on other platforms, such as Facebook, where content is generally shared with a circle of friends.

2.1.1 Twitter User Profile

Twitter, as a platform, provides users with several customizable choices for the display of their public profile to other Twitter users. The user profile is a place where users can make personalized decisions about the appearance of their Twitter site. Profile data elements can be broken into two main categories: 1) Personal Information and 2) Outward Appearance.

First, users can elect to share personal information via their Profile page publicly shared with other Twitter users without the ability to control access. Twitter provides users the ability to publish their Name, as a free-text field, where users can choose to

Figure 2.1 A sample of personal information provided by President Obama's Twitter profile. *Source: https://twitter.com/barackobama (accessed on May 28, 2014).*

provide their first name, surname, or both. In addition to providing their name, users are provided a field labeled Description, which is again a free-text field, intended to provide users an additional space to summarize information about themselves. Often, the Description field is chosen to be used to provide a self-described explanation of oneself in the form of interests, background, or other pertinent information.

Finally, users have the option of choosing multiple colors to customize the appearance of their account for themselves and other viewers of their profile. The customization options provided include: Background color, Sidebar color, Text color, Link color, and Sidebar Border color.

2.1.2 Example Tweet

Tweets are limited to 140 characters or less. The content within a tweet can contain text and/or an HTML link to an image or external website. In the example provided in Figure 2.2, the user "@JohnElway" is mentioned in the body of the tweet.

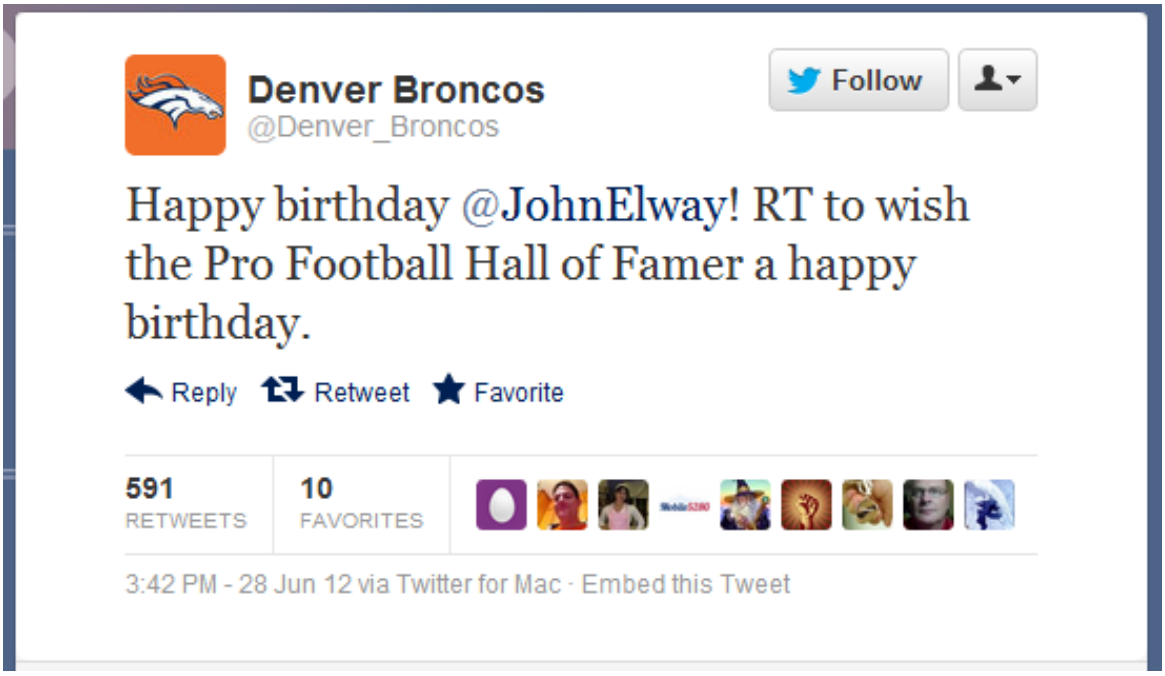

Figure 2.2 A sample tweet containing a message directed at a particular user. *Source: https://twitter.com/denver_broncos (accessed on June 28, 2012).*

2.1.3 Twitter Keyword Summary

Twitter is a social media site that allows users to post content to a set of *followers*, who have chosen to subscribe to the user's feed. Twitter has been instrumental in introducing several new constructs for sharing, posting, and directing information over social media. A basic understanding of these foundational elements are introduced below. An understanding of them is central to the ideas introduced in this paper.

- **Hashtag** Appended to tweets by users to attribute a topic summarization of the posted tweet. As the popularity of a topic grows, the hashtag can be elevated to *trending* status, where other users may choose to adopt the hashtag in their tweets. Hashtag are prefixed with a #-symbol [62]. For example one user posted, *'Employers pay penalties in pay or play either way! #obamacare.'*

- **Mention** A method of directing a tweet at a given user by including a user's handle (in the form of @username) within the posted tweet. Typically, this can be used in three ways: 1) to direct a message at a particular user as a form of communication, 2) to reply to a tweet posted by another user, thus creating a chain of events, and 3) to bring another user's attention to a given tweet by copying the original tweet text and including the user's handle as part of the reposting. Mentions act as one of the primarily methods of inter-user communication on the social media site. For example, one user posted, *'I wish @[username withheld] MP a speedy recovery n good health.'*

- **Retweet** Retweets are a mechanism used by users to share content with their set of followers. This content is often not original and often originates from another user's timeline [63]. For example, one user chose to share the following information: *'RT @[username withheld]: Can we build a health system for our BOOMERS? [http web link].'* The original Tweet's content includes all information following "RT @[username retracted]:."

- **Timeline** A timeline is a chronological collection of tweets by a given user. Timelines serve as a historical record of a given user's activity on the social media site and can be accessed by any user (if the user has elected to be public) or by *followers* only (if the user has elected to be private).

- **Favorite** A method where a given user marks another user's tweet as favorable. These favorite posts are accumulated in a aggregated list of favorite tweets for the user to retrieve at any time.

- **Profile** A place where users can make personalized decisions about the appearance of their Twitter site. Additional personal information can be provided here, such as Name and Description. Both the Name and Description field are free-text areas. Notably, the Description field can be used to provide a self-described explanation of oneself in the form of interests, background, or other pertinent information. A user's location can also be provided in the user profile, as well as an http link to an external website.

- **Tweet** The tweet is the fundamental element in the Twitter social network platform. Tweets are limited to less than 140 characters but can provide links to external content. While tweets are text-based elements, they can have links to images or external websites embedded within them, thus providing a richer set of content beyond exclusively text.

2.2 Social Media Demographics Extraction

Social media demographics extraction research follows a similar theme throughout the existing work. These works often rely on a narrowly focused feature set used to inform classification models of a user's demographics. Predominately, statistical language models derived from word frequencies (bag-of-words) of user posts are used in demographic extraction research. As is the case with [11, 10, 64] which aims to extract a user's age and [6, 7, 65] which focuses on gender, all use an analysis of word usage patterns to generate statistical models. Still other researchers make small modifications to the standard bag-of-words techniques by extending bag-of-words to bi-grams or tri-grams, as is the case in work by Peersman et al. [65].

Language has been an important tool used by researchers to classify users into particular groups, often by demographics. In work by Akram et al. [66], specific personality traits were detected using language as well as gender as a demographic element. Akram et al. premise was that psychologists believe language (i.e., the

phrases and text that people use) is the key to human thoughts [67]. To study this, these researchers attempted to connect a user's Twitter posting activity with the "Big Five Model", a psychology model that provides broad classification of personality traits by dividing the human personality into five separate categories [68]. These categories are openness, conscientiousness, extraversion, agreeableness, and neuroticism [69]. The first part of their work looks to detect the five personality elements using the language in tweets. Secondly, they looked at mental health prediction (specifically psychopathic tendencies). They hypothesized that a tweet of a psychopathic user would contain more cause and effect statements, more primitive needs, and less emotional discussion when compared to those of healthy individuals. Finally, this work looked at gender prediction working under the notion that research has shown women generally discuss relationships more than men. Conversely, men often discuss objects more than women [70]. For this study, the researchers collected tweets from 345 Twitter users and aggregated each of the user's history into a single file, representing the posting history of each user individually as a document. Linguistic Inquiry and Word Count (LIWC) [71] and Natural Language Toolkit (NLTK) [72] were selected by the researchers to extract meaningful features that have close relationships with the psychology-related classification goals (e.g., Big Five Personality, psychopathic behavior, and gender derived from psychology tendencies). While this study did not examine the accuracy of the classifications it made, it was used to assess the overall landscape of Twitter users, stating that a small percentage of users in fact had psychopathic tendencies. The gender-related study has shown that the gender makeup of the researchers' classifications matched that of the reported gender makeup of Twitter as a platform.

While language has been used by many researchers to classify users into groups, other derived features have also proven to be important beyond language. Specifically, Alowibdi et al. looked at using color choices and name extraction

for detecting a Twitter user's gender [73]. This work [73] criticized much of the language-based classification models as having too many features, a problem caused by using bag-of-words models to represent user's posting activity. They state that the increasingly large number of features used to train models increases complexity, both in terms of computational complexity as well as complexity that reduces the ability to interpret the classification models. [73] disregarded the language contained in the posting activity of users and instead focused on three components for feature derivation: 1) first name, 2) user name, and 3) profile colors. This work looked at 194,293 Twitter profiles by extracting the ground truth data (Gender: Male/Female) according to links provided in a given user's Twitter profile. For example, if a Twitter user provided a link to a Facebook profile, [73] would scrape the gender information from the user's Facebook profile and subsequently consider it has ground truth. Following the collection of the ground truth labels of gender, each of the users' profiles are collected for feature extraction. First name and usernames are converted to phonemes using the LOGIOS lexicon tool. Using this tool, names such as "Mary" would be represented as "M EH R IY", breaking down the words to their phonetic components. The individual phonemes were then expanded to n-gram phonemes (where n ranged from 1-5). As a result, "Mary" would ultimately be represented as ("M", "M EH", "EH", "EH R", "R", "R IY", "IY") for a 2-gram scenario. Along similar lines of thought, each of the individualized colors from the user's profile (background, text, link, sidebar fill, and sidebar border colors) were converted to a spectrum of 512 colors from their HTML color codes. Color as a feature set was examined in a previous study by Alowibdi et al. [74]. Alowibdi et al. considered three classification models for classifying users into male and female groups, namely Naive Bayes, Decision Trees, and a Naive Bayes-Decision Tree Hybrid approach. They showed that both name and color choices did have a positive predictive impact on the gender classification results. Specifically and unsurprisingly, they found that first

names made a stronger contribution to gender classification than color and usernames. Finally, Sayyadiharikandeh et al. examined the classification of users' genders based on tweets, screen name, and profile picture by using a combination of classifiers, image recognition tools, and unigram representations of tweets [75].

Researchers have also explored interactive methods for inferring demographics of Twitter users. Specifically, Beretta et al. [57] looked at a two-step process for learning the age and gender of Twitter users. This group of researchers considered the name extracted from Twitter profiles as a feature for classifying gender. To do this, names were compared against a database of names as related to gender and converted into one of five enumerations (Female, mostly Female, Male, mostly Male, Unisex). Age information was collected by looking at user tweets which contained birthday-related text (e.g., "Today is my 25th birthday."). Age was then discretized into two categories, users below 30 and users above 30. Uni-grams (bag-of-words approach) were extracted from user tweets and coupled with part-of-speech tags to incorporate stylistic features in the model. Using Support Vector Machines (SVM) and Naive Bayes, each tweet is then classified as originating from a younger/older or male/female user. The probabilities of each age/gender classification are then aggregated to form a final classification of the user. SVM was shown to outperform Naive Bayes in terms of performance. Ultimately, this classification model was used as an initial guess for a larger system that was used to refine the predictions made by the classifier. The researchers designed a tool which provided users with an initial guess (using the features and classification approach described above). Human users were then provided with an interface that displayed additional information about the users (e.g., pictures, profile images, etc.) that they could incorporate into their final consideration of gender and age. This approach was intended to incorporate human refinement into the classification results, ultimately resulting in a semi-automatic approach to classifying the age and gender of Twitter users.

The Burger et al. approach [6] to gender and age classification is considered one of the leading approaches in terms of accuracy . This work collected a large sample of Twitter users along with their gender by gathering age and gender information from linked blog sites in the user's Twitter profile. This was done because of the structured nature of the blog sites specifically labeling date of birth and gender information. To construct a feature set for gender and age classification, Burger et al. collect tweets, full name, username, and user description information. Using a character-level n-gram approach, aggregated tweets, name, username, and descriptions are converted to a vector space representation, ultimately resulting in very high dimensional data on the order of 15+ million features. Coupling all four feature sets together provided the highest accuracy using an SVM classifier. Burger et al. provides an analysis of how the number of users included in the training set impacts the overall accuracy of the approach. In addition, Burger et al. looks at how increasing the number of tweets collected for each user also increases the accuracy of the classification (age/gender). Finally, Burger et al. compared the automated classification approach to human classifications (using Amazon's Mechanical Turk). Human classifiers were asked to classify 100 or more profiles into male and female categories. Accuracy scores for each human were calculated and compared against the automated classification results. Burger et al. showed that the machine learning approach outperformed all but the top 5% of human classifiers. This approach showed a wide range of accuracy scores, directly related to the size of the training set used when training the model. Having approximately 184,000 Twitter users with a labeled gender, he tested his approach using various sizes of training data to ultimately show that using less than 1000 users in the training data produces less than 70% classification accuracy. This leaves obvious room for improvement, both in terms of a smaller number of required users for training as well as the number of features used in the classification model by reducing the dimensionality from 15+ million. While 1,000 users in a training set

might be considered reasonable, Twitter's throttled API access can at times make data collection difficult for groups of demographically labeled users.

Demographics have been considered from multiple perspectives on Twitter. Chen et al. [76] aimed to improve previous work on demographic extraction by including profile self-descriptions and profile images as features in a classification model. These researchers collected hand-annotated accounts using Amazon's Mechanical Turk. Mechanical Turk users were presented with account information for each Twitter user (including their name, profile image, self-description, and one sample tweet) and asked to infer the user's gender (either Male, Female, or Unsure). Hand annotations were compared, with labels obtaining more than two out of three agreements among annotators retained for the study. In their work, the researchers take a new approach to feature generation by considering the neighboring users' tweets and self-descriptions for feature extraction. Additionally, n-grams are generated from the 10,000 most frequently appearing uni-grams. Combined with this, an additional representation of language is generated by creating a 100-topic LDA model. This model is used to assign latent topics to each tweet and self-description. Largely, the problem with this approach stems from the fact that Latent Dirichlet Allocation (LDA) [77] has been shown in previous studies to perform poorly when applied to short-text, or limited-content text, such as in Twitter [78, 79]. The final derived feature comes from the profile image, where the researchers quantize images into specific visual words by using SIFT (Scale-invariant Feature Transformation) [80] for object recognition. SIFT is an algorithm for detecting objects in an image. This approach is known for its ability to detect objects, which the model was trained to recognize, in environments with various lighting conditions, cluttered images with multiple objects, and varied size/scaling of the trained object. Ultimately, these features were tested individually and as combinations using SVM to generate the final classifications. Still other researchers have considered deriving feature sets from

other types of information provided by users. Specifically, Ma et al. [81] looked at automatically annotating posted images into various content categories also using SIFT. They then constructed models using crowd-sourced labeling of users according to their gender. A slight variation of the image categorization technique comes from Merler et al. [82]. These researchers trained a classification model to detect the types of faces which appeared in profile pictures to gender.

Still others have taken a different approach to extracting demographics from online activity of users. Culotta et al. [83] constructed a regression model based on the website viewing activity of users and their neighbors. These researchers constructed a training set of labeled data by assuming the Twitter followers fit the common demographic of a given website provided in their profile. In other words, if the common demographic of www.ign.com is 18-24 males and the user provided a link to that website in their Twitter profile, this user was then assumed to have these demographic characteristics. Using this information, a rich network of connected users on Twitter was constructed. By using the characteristics of 10 connected friends from a given user, a regression model was able to predict the user's demographics. Similarly, Ito et al. [84] looked at Twitter user attributes and the content of their neighbors to infer user demographics.

Other forms of online communications have been considered when attempting to extract user demographics. Filippova [85] considered the comment posting activity of Youtube users. Filippova looked at using features extracted from the language of user posts to infer their demographics (age and gender). He found that features such as use of pronouns, determiners, and function words helped indicate a given user's gender.

Profiling users beyond demographic components has also been actively researched in recent years, using more diverse feature sets to classify users into groups by personality type [86, 12], account type [8], and political affiliation [87, 88]. These

researchers use descriptive feature sets including network statistics (in-degree and out-degree centralities), posting activity, interaction with other users, the presence of named entities, and topic distributions of the user's timelines to inform their models. While the classification outcomes of these research do not relate directly to demographics, the use of diverse feature sets has informed our research.

As shown above, researchers have considered various methods for extracting user demographics from online activity. Demography in medical studies has long served as important cornerstone of understanding disease susceptibility and propagation [89, 90, 91, 92]. This research extends our understanding of how to extract user demographic information and simultaneously presents new opportunities for health demography studies via social media through the application of the proposed user mining approaches.

2.3 Twitter Text Mining

Much of the existing user classification and demographic extraction work focuses on word statistics and language models. This lead us to explore the existing literature describing methods of transforming the raw text contained in tweets into new features. Topic modeling, statistical models that infer latent topic representations of text content, remains an important research area in social media text mining [93]. However, multiple studies have shown [78, 79] that standard LDA topic modeling [77] approaches struggle to generate topics due to the limited text content ($\leq$ 140 characters) available in tweets. As a result, extensive work has been conducted to overcome challenges posed by limited-content. Various tweet pooling schemes have been developed to aggregate tweets and, therefore, increase text content available to latent topic model approaches, such as LDA. Additionally, researchers have taken advantage of *hashtags* to discover sentiments among broad topics [94]. Hashtags have remained an important linguistic structure within social media, particularly for topic

detection. Mehrotra et al. [95] explored the concept of pooling tweets by hashtag, reducing the impact of limited content and exploiting the user-generated topic assignments. Ramage et al. [96] also examined the application of supervised topic modeling approaches in combination with inferred labels using hashtag and trending topic detection, showing improvements in topic modeling over existing approaches. In these works [95, 96], while considered the most accurate labeling techniques, still come with limitations. In the Ramage et al. approach, a new topic label for each hashtag in the collection leads to specific and often redundant labels. Similarly, [95] pools tweets by shared hashtags yet fails to account for multi-hashtagged tweets. Both of these approaches share a similar limitation in that models are trained using the language of a single hashtag. For example, in each of these approaches, different labels would be generated for *#photo* and *#photography*, which ultimately represent the same conceptual hashtag. Consequently, the language model, as a result of the vocabulary it associates with each hashtag, is likely to experience difficulty in distinguishing between a *#photo* tweet and a *#photography* tweet, thus producing topic classification with low confidence. We propose a new method for assigning labels to tweets by first clustering similar hashtags, constructing a new feature set using part-of-speech (POS) tagging, and further refining results by using probability scores.

Many researchers have looked at classifying tweets into specific topics, often choosing to focus on a particular set of topics for detection. Batool et al. looked at classifying tweets into one of eight categories, namely diabetes, food, diet, medication, education, dengue, parkinson, and movies [97]. This work initially focused on detecting a classification of diabetes-related tweets, however, was expanded to include non-diabetic related topics (i.e., movies and dengue) to showcase the ability of categorization of broad and specific topics. By searching for specific topics across the publicly available tweets, these researchers refined the text in tweets by applying several filtering layers before classification. Namely, entity extraction was

performed on tweets by looking at the natural language components of the text, i.e., part-of-speech tags. Following that, synonyms of words were detected to normalize features by converting terms to both their singular form and appending addition terms found in the WordNet dictionary, e.g., *calories* to *calorie* and *exercises* to *exercise, workout*. Although the research did not specify the classification algorithm adopted, the preprocessing, filtering, and expansion of text was cited as improving classification accuracy from 0.1% to 55%, demonstrating the importance of transforming raw text to meaningful features. Still other researchers looked to classify tweets into categories by overcoming the limited content problem ($<$ 140 characters) using expansion techniques augmenting original tweets with web content (specifically from Yahoo! Answers) [98]. Content expansion techniques were shown to be promising methods for improving the accuracy of tweet classification. Still others chose to expand tweets with other data sources, such as Wikipedia, to improve categorization results [99].

2.4 Classification and Language

Many researchers have drawn connections between writing styles and choices and the demographics (e.g., gender, ethnicity, etc.) of the authors. Colley and Todd [58] conducted an experiment on male and female participants by monitoring their email activity. While finding little difference in the use of email, Colley discovered that there was a measurable difference in the types of content conveyed over email when comparing male and female-authored messages. Topics commonly considered to be female-centric commonly appeared in the writing of email, such as intimacy, shopping, and nightlife. Men often included location descriptions and people-related content in their emails.

In addition to detecting demographic elements of users, researchers have considered building models to detect specific authors based on writing styles. Authorship attribution models, such as those in [100], rely on standardized word

frequencies of select terms and punctuation; ultimately developing models with features precisely tuned to detect the author under consideration. This targeted approach to vocabulary has been shown to improve classification results by only including features, which are of importance to authorship attribution.

We have identified opportunities in the literature for enhancing the performance of language models designed for Twitter. Specifically, we aim to extend the ability to create accurate supervised learning models for topic extraction from limited-content tweets. We have also observed limited exploration into demographic extraction of Twitter users. As a result, we explore the application of language models to user content combined with the extraction of user information for the purposes of inferring user demographics.

CHAPTER 3

LANGUAGE AND CANCER TRENDS IN SOCIAL MEDIA

3.1 Introduction

In this chapter, we explore the relationship between language use on Twitter and a user's ethnicity. Previous studies have shown that language usage patterns have strong connections to one's ethnicity [59]. With the lack of publicly available ethnicity information at a Twitter user level, we introduce and test two new methods for extracting ethnicity information from a user's timeline. This work first considers the expansion of user timelines through the addition of noun and verb synonyms. Conversely, we test a reduction method that represents a user's timeline with the probability distribution over a set of latent topics. The proposed approaches are compared against a bag-of-words baseline model.

Public health studies have shown that cancer mortality rates vary greatly among ethnic groups [21, 22]. In this work, we also explore the application of ethnicity extraction models to Twitter data for the purpose of exploring health discussion trends among ethnic groups. We adopt the highest performing ethnicity extraction model for assigning an inferred ethnicity to an unlabeled user. Having labeled all users in the collection, we observe the trends with which each ethnicity discusses cancer, breast cancer, lung cancer, colon cancer, and prostate cancer. Cancer is a particularly important disease for analysis on social media because of widespread impact on U.S. patients [21] as well as the large disparity in terms of awareness [31, 33] and also mortality rates among minorities and non-minorities. In addition, cancer awareness campaigns have become increasingly popular public health devices for increasing awareness of specific types of cancer. These awareness campaigns often run annual month-long events for driving patient discussions around cancers. As

a result, we examine the rate of discussion of these diseases between April 2014 to January 2015 for Caucasian and African-American users to observe the impact awareness campaigns have as well as the overall discussion rates of each ethnic group.

3.2 Method

3.2.1 Data Collection and Preprocessing

Tweets were collected from April 1, 2014 through January 21, 2015 using the Twitter public streaming Application Programming Interface (API) to collect 1% of public tweets, yielding 281,276,343 tweets submitted by 40,403,529 unique users. We are aware that there are publicly available datasets intended for Twitter analysis, however, these datasets are often small (in terms of total counts of tweets collected), short (in terms of the duration of the collection period), and often topically focused (collecting only tweets which contain a given keyword). For this reason, we elected to create our own Twitter dataset using the free streaming API Twitter provides, consisting of 1% of public tweets, made freely available to researchers for researcher purposes. For this study, we restricted our collection to English-only tweets. We provided no restriction on GPS values for each tweet due to the sparsely available GPS data and instead focus our tweet location to U.S.-only accounts using an approach introduced later in this paper. Due to a technical issue with our collection system, tweets from May 13, 2014 through July 24, 2014 were not retained. These tweets were not able to be recollected because of the nature of the streaming API. Free tweets are provided to researchers who maintain an active connection to the Twitter servers. The power outage which occurred on July 24, 2014 which disrupted the connection to the Twitter servers caused the loss of data during that period. The lost tweets during this time period were irretrievable via the streaming API. During the uninterrupted data collection period, the Twitter-provided unique user ID number, tweet, Data/Time, profile-identified location, and GPS latitude and longitude values

were collected (when available). Following the collection of tweets, user timelines were re-constructed by grouping tweets using the unique user ID number. Figure 3.1 and 3.2 show the distribution of character lengths for tweets in the collection.

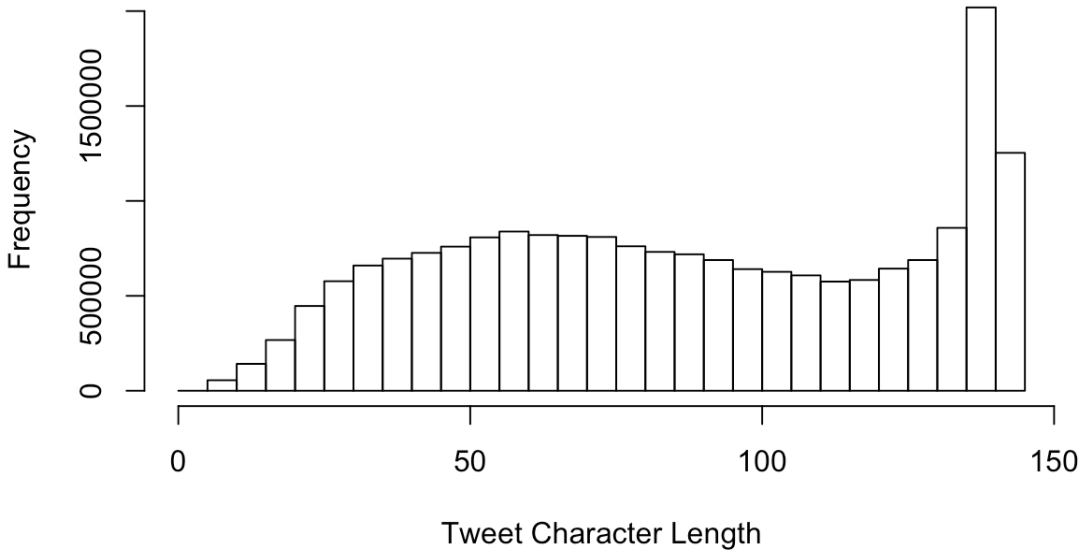

Figure 3.1 A histogram of tweet character lengths.

The preprocessing procedure for cleaning tweets followed a consistent approach across all collected timelines. Given that the focus was on the predictive power of text, tweets containing linking information outside of the self-contained tweet, predominately non-language elements (i.e., URLs, usernames, and re-tweet information) were systematically removed. For example a tweet containing elements such as, "www.t.co", "cnn.com", "@username", and "RT @username" would be removed from the collection. While re-tweeted text may provide information about individuals and/or organizations a user interacts with via Twitter, at this scale we were unable to include all re-tweets using the provided Twitter API due to rate limitations (i.e., restrictions imposed by Twitter limiting the number of searches we could conduct in a 15-minute period). User timelines (tweets aggregated by the user) that contained little information were removed by systematically eliminating those that were shorter than

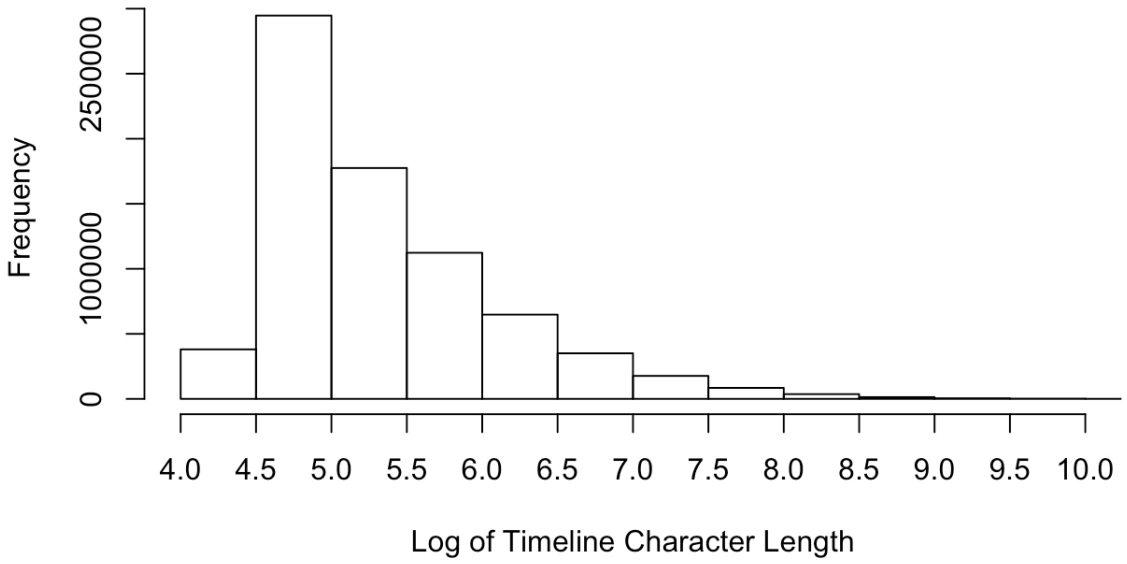

Figure 3.2 A histogram of the log of timeline lengths of users.

eighty-five characters from the study. To select this character threshold, we observed timelines of increasingly longer lengths from 10 characters upwards. During this cleaning process, it was apparent that timelines shorter than eighty-five characters generally contained fewer than fifteen words, which provided little information to make accurate classifications. These preprocessing methods left us with a final tweet count of 19,818,236 belonging to 779,653 unique users' timelines for analysis.

3.2.2 Label Extraction and Identification of Race/Ethnicity

The approach to classifying users' ethnicities presented in this paper relies on Vapnik's supervised learning classification approach, Support Vector Machine (SVM)[101], which requires accurate training data to inform the classification model and also a reliable set of testing data for assessing the accuracy of classifications. To acquire training data indicating the ethnicity of Twitter users, we looked for specific declarative statements within each user's timeline (i.e., statements where users explicitly defined an element of their personal identity). Timelines that contained

such declarative statements were labeled accordingly, receiving one of four enumerated keys. These keys indicated the types of ethnicity explored by this study, taking the values of: *Caucasian*, *African-American*, *Asian*, and *Hispanic*. Searches were conducted across all timelines to detect the following terms: "black," "african american," "africanamerican," "african-american," "white," "caucasian," "asian," "hispanic." These searches significantly limited the timelines that needed to be considered for labeling. As a result, this smaller subset of timelines was manually labeled. Timelines were labeled by hand because of varying forms of declarative statements. However, declarative statements were required to have a first person pronoun as the sentence subject followed by clearly stated race-related comment. Examples of declarative statements include: "I am African-American", "I'm Asian", or "I'm a black man." Additionally, more complicated declarative statements were race-labeled. For example, "I'm a hard working individual and a black man." Although we are aware of the sociological differences in race and ethnicity, with race associated with biology and ethnicity associating with culture, this study does not make distinctions between the two types of declarative statements since the end Twitter users who contributed to such statements are not always consistent.

3.2.3 Classification Approach — Classification of Race/Ethnicity

Individual tweets are short, often uninformative messages providing little classification potential for identification of user profile information. This led us to examine users' timelines, rather than individual tweets, to enhance the accuracy of our classification approach by extracting features consisting of deeper information around users' activities. Users' tweets were aggregated into timelines containing the chronological order of their submitted tweets for the 10-month data collection period. This was done by grouping tweets by a given user into a single document representing the entire

user's posting history. This provided a larger text source for identifying descriptive elements indicative of a given user's ethnicity.

Baseline classification models described in previous work [6] adopt document-term matrices for representing the frequency with which terms appear in a given timeline. In the context of this work, classification algorithms are trained to detect vocabulary usage patterns among a group of ethnic users. Recall that the known ethnicity of a user is acquired according to their self-identified ethnic background (e.g., "I'm African-American," appearing within their timeline). The vocabulary usage patterns detected among the self-identified users are then applied to users who chose not to explicitly define their ethnicity. This is done to infer previously unknown demographic features of a wider population of users. Demographic information is often used to summarize health trends across a population and, therefore, the extraction of demographic information in a wide social media population could provide useful information for public health studies. Two opposing approaches were examined in this study: 1) how can timeline synonym expansion enhance predictive ability? and 2) can a topical representation of user posting activity produce an equivalent accuracy score? These scenarios were born from two ideas. The first is that users often express similar thoughts on social media with varying lexical choices. Secondly, can these similar thoughts be equivalently represented using a topic modeling approach.

Connecting the users' vocabulary choices to specific ethnicities proved to be difficult. This was discovered while building the baseline classifier [102, 103, 104] using a bag-of-words approach similar to the baseline in previous works [88]. For example, one ethnic group may often use terms such as wife, spouse, and marriage, consistently appearing as some of the most identifying terms for that group. Having identified that Twitter users often used varying terms to describe the same concept, we expanded tweets with additional vocabulary in an attempt to increase lexical overlapping of

group member term usage to easily segment profile types. Using part-of-speech tagging, we identified nouns and verbs within tweets. Then for each tweet, using WordNet (a lexical database where nouns, verbs and adjectives are collected into sets of cognitive synonyms) [105, 106], the top five synonyms according to the WordNet ranking of frequency of synonym usage relative to the target term, when available, for each noun and verb were appended to the tweets, resulting in expanded tweets while retaining their original meanings. This allowed for more frequent overlap between tweet term usage among racial and ethnic groups and a more accurate classification algorithm. To the best of our knowledge, using synonym expansion of tweets to enhance the bag-of-words feature set has not been explored in detecting the ethnicity of Twitter users.

Latent Dirichlet allocation (LDA) [77] is a statistical method for computing abstract topics of a given document using the co-occurrences of terms within the documents of a corpus. LDA assumes that documents are distributions over topics and topics are distributions over words. Based on the occurrences of words within a larger corpus, a new document's distribution of topics can be inferred based on the vocabulary that is present. Our second ethnic classification approach used LDA to detect patterns among topics rather than vocabulary usage by first converting tweets into topics. We acknowledge that LDA is typically used for topic detection in long documents and its limitation when applied to topic detection from short text. Nevertheless, by our study design, all tweet text contributed by a Twitter user were first aggregated to generate the user's total writing record on Twitter, after which LDA is applied onto the aggregated writing record of a user (averaging 324 characters). Although this writing record remains relatively short in length, the duration of the collection period and size is consistent with the collection in the original work that proposes the author-based aggregation technique [95]. In Figures 3.1 and 3.2, we summarize tweet and total tweet writing record (user timeline)

length of the collection of tweets examined herein. This author-based aggregation step greatly mitigates the sparsity issue of short input text to the LDA model. It is noted that the above pre-processing step is also popularly adopted when topic modeling is applied to Twitter data [95, 79, 107]. Using LDA topic distributions to represent timelines resulted in a reduction of features (variables used for classifying the ethnicity of a user; for example, these variables consisted of frequency counts of stemmed-words such as "togeth", "damnnn," and "sharp," which generally indicated an African-American user, and "newyork," "lifetime," and "whatchya," which were strongly associated with Caucasian users) by 99.7% while improving classification accuracy for some ethnic groups. The number of abstract topics, and thus the number of features representing Twitter timelines, was decided on by iteratively building classification algorithms with increasing larger topic sizes. LDA models with topic sizes ranging from 10 to 100 were constructed for the user timelines. Accuracy of the model within this corpus of timelines peaked at approximately forty-five abstract topics, which was then adopted for each testing set. In this approach, we aimed to reduce the number of features representing the activities of each Twitter user. Having reduced users' timelines to representation comprised of LDA topic distributions, we then adopted a Support Vector Machine (SVM) classification approach with a radial basis function kernel for our classification algorithm. This method was chosen for its demonstrated ability to perform well with text data and is consistently considered the best approach in text classification studies [108]. SVM is a supervised classification algorithm that attempts to maximize the margin between classes in the training dataset. One benefit of the SVM classification algorithm is the kernel trick, which allows for the direct transformation of data points using kernels. In this work, we considered several kernels including linear, radial basis function, and sigmoid with the highest results observed when using the radial basis kernel.

We used ten-fold cross validation to test the accuracy of the models. The labeled dataset was divided into ten, equally sized bins. Nine of the ten bins were used to train the model, while the remaining bin was used for testing. We iterated over the bins ten times, reserving a new bin for testing with each additional iteration. Due to the unbalanced nature of our dataset, we chose two evaluation metrics. First, for each ethnicity, we computed the Balanced Accuracy (Equation 3.1), a performance metric intended for unbalanced classes [109]. Second, we provided the overall accuracy for all ethnicities (Equation 3.2), as well as the accuracy for Caucasians and African Americans (the two groups focused on in the second part of this study).

$$BalancedAccuracy = \frac{1}{2}\left(\frac{t_p}{t_p + f_p} + \frac{t_n}{t_n + f_n}\right) \tag{3.1}$$

$$Accuracy = \frac{(t_p + t_n)}{(t_p + f_p + t_n + f_n)} \tag{3.2}$$

Where t_p: true positive, t_n: true negative, f_p: false positive, f_n: false negative.

In addition, we provided a confusion matrix of the classification results in Table 3.4 (results for text classification with synonym expansion) and Table 3.6 (results for the topic-based method) to give further details of the classification performance.

3.2.4 Statistical Analysis

Each statistical analysis for this study was carried out using the R Statistical Software Package [110]. To measure the statistical significance of the observed differences between groups, t-tests were conducted with pairwise comparisons of ethnic groups (i.e., Caucasian vs. African American, Caucasian vs. Hispanic, etc.). We tested the hypothesis that there were no statistically significant pairwise racial and ethnic group differences in cancer term usage during each month of the study period. Because pairs

of ethnic groups were tested independently of one another, no adjustments for multiple comparisons were made. P-values < 0.05 were considered statistically significant.

3.3 Results

To evaluate the success in the classification of race/ethnicity, we compared the accuracy of text classification with synonym expansion against the topic-based method (Tables 3.3 and 3.5). These approaches are compared against the performance reported using the bag-of-words approach (Table 3.1). We found that the accuracy of text classification with synonym expansion outperformed the topic-based approach in most cases. Using the synonym expansion approach, we achieved the following accuracies for correctly identifying user ethnicities: 88.87% among Caucasian users, 81.26% among African-American users, 72.32% among Asian users, and 69.07% among Hispanic users. The overall accuracy for all groups using this approach was 76.07%. Using topic detection, we observed no improvement in overall accuracy at 55.59%. Among the groups we also observed a lower accuracy score (Caucasian, African-American, Asian, and Hispanics resulting in 71.89%, 68.32%, 53.43%, and 54.50% respectively). In addition, we report the confusion matrices of the classification results to illustrate the differences in class sizes among the collection (see Tables 3.2, 3.4, and 3.6).

We suspect topic detection classification produced lower accuracy scores due to the loss of nuanced lexical differences between ethnic groups lost during the feature reduction process. For example, a topic model may identify a timeline containing some topical references to "family." Thus, this would be represented as a feature in the topic-based approach. However, we observed that the explicit usage of terms such as "husband," "girl," "boo," "baby," or "wife" provide a stronger indication of ethnicity than broad topics. These terms are lost in the topic-based feature representation of user timelines and subsequently produce lower accuracy scores.

Table 3.1 Text Classification using a Bag-of-Words (Baseline) Classification Model and Accuracy Results

Race and Ethnicity		%
BalancedAccuracy		
	Caucasian	83.47
	African American	77.26
	Asian	67.28
	Hispanic	69.87
Accuracy		
	All Groups	72.35
	Caucasians and African Americans	84.19

Table 3.2 Confusion Matrix of Bag-of-Words (Baseline) Model Classification Results

Classification	Reference, n			
	Caucasian	African American	Asian	Hispanic
Caucasian	2453	449	137	126
African American	287	1469	318	203
Asian	20	30	274	40
Hispanic	26	40	27	261

Table 3.3 Text Classification with Synonym Expansion Model Classification and Accuracy Results

Race and Ethnicity		%
BalancedAccuracy		
	Caucasian	88.87
	African American	81.26
	Asian	72.32
	Hispanic	69.07
Accuracy		
	All Groups	76.07
	Caucasians and African Americans	88.32

Table 3.4 Confusion Matrix of Synonym Expansion Model Classification Results

Classification	Reference, n			
	Caucasian	African American	Asian	Hispanic
Caucasian	1689	125	26	21
African American	261	1231	183	276
Asian	24	38	211	27
Hispanic	16	26	30	216

Table 3.5 Topic-Based Model Classification and Accuracy Results

Race and Ethnicity		%
BalancedAccuracy		
	Caucasian	71.89
	African American	68.32
	Asian	53.43
	Hispanic	54.50
Accuracy		
	All Groups	55.59
	Caucasians and African Americans	70.03

Table 3.6 Confusion Matrix of Topic-Based Model Classification Results

Classification	Reference, n			
	Caucasian	African American	Asian	Hispanic
Caucasian	1067	117	49	71
African American	890	1286	337	380
Asian	26	10	39	35
Hispanic	7	7	25	54

Given the higher overall accuracy, as well as the high accuracies among Caucasian and African-American users, we selected the synonym expansion approach for classifying the remaining unlabeled users within the collection. Additionally, we elected to exclude users classified as Asian and Hispanic from this study for multiple reasons. First, the population sizes where users declared ethnicities of these types were markedly smaller than populations of Caucasians and African-Americans. Second, we believe we may have excluded some Asian and Hispanic users by limiting the tweet collection to English-only tweets. The combination of these complications (small population sizes and the restriction of English-only tweets) is a likely reason for the reduction in accuracy among these groups and their subsequent exclusion from the study.

In this study, we have established and tested a systematic method for detecting ethnicities among Twitter users. Using the more accurate approach, text classification with synonym expansion, we detected and assigned ethnicities to all users within the collection consisting of 19,818,236 tweets posted by 779,653 unique users. tweets were divided by posting date into nine months, accounting for the ten-month study period with portions of May and July and the entirety of June lost due to system failure. Various descriptive statistics were calculated to describe the health effects extracted from the dataset.

As shown in Table 3.7, the number of unique users varied widely by race and ethnicity. To detect significant differences in term usage between ethnic groups, each term contribution was normalized by the percentage distribution of its population. Additionally, the term frequency for each ethnic group is provided without normalization. The number of unique users from each ethnic group was examined for each month. Caucasian users dominated the dataset (92.32%, 719798/779653), while African-American users often represented 7.12% (55549/779653) of the population, and both Asian and Hispanic users made up a small percentage of the overall

Table 3.7 Distribution of Unique Active Twitter Users during each Month of the Study Period by Race/Ethnicity

Month	Race and Ethnicity				Total
	African American, n (%)	Caucasian, n (%)	Asian, n (%)	Hispanic, n (%)	
April	49104 (9.72)	452924 (89.64)	1289 (0.25)	1935 (0.38)	505252
May*	40956 (12.76)	277169 (86.36)	1177 (0.37)	1646 (0.51)	320948
July*	43349 (9.58)	405185 (89.57)	1661 (0.37)	2191 (0.48)	452386
August	54740 (7.91)	632687 (91.47)	1820 (0.26)	2466 (0.36)	691713
September	52224 (10.16)	457300 (89.02)	1789 (0.35)	2417 (0.47)	513730
October	50120 (11.07)	398440 (88.02)	1763 (0.39)	2371 (0.52)	452694
November	50060 (10.80)	409125 (88.30)	1762 (0.38)	2370 (0.51)	463317
December	48247 (11.20)	378412 (87.86)	1727 (0.40)	2292 (0.53)	430678
January	30707 (15.62)	162682 (82.75)	1435 (0.73)	1780 (0.91)	196604

population (0.55%, 4306/779653). We were less confident in predications of Asian and Hispanic ethnicity among users based on the smaller training set as well as the lower accuracy values among these ethnic groups.

This study focused on the social media attention given to site-specific cancers and differences by race/ethnicity. Specifically, Twitter timelines were examined for the frequency of occurrence of the following terms: "cancer," "breast cancer," "prostate cancer," "colorectal cancer," and "lung cancer." These terms were detected using methods adopted in previous studies examining discussions about specific health topics on Twitter [111]. We are aware of other work [112] that distinguishes between medically-related use of the term 'cancer' and non-medically related uses. However, when examining our own dataset, by sampling 200 randomly chosen tweets containing

the word "cancer" (representing 1% of all "cancer" tweets in the collection), we observed only 8.5% of tweets were used in the context of Zodiac signs and 2% referred to destructive practices (e.g., "He was a cancer to the community."). We suspect the low percentage of non-medically related usage may be a result of the cleaning process performed, where tweets containing URLs were stripped from the collection (i.e., horoscope tweets often contain links to an extended version of the horoscope). Furthermore, we examined samples of each of the bi-gram terms of interest (e.g., "breast cancer," "prostate cancer," "colorectal cancer," and "lung cancer"). We observed no uses of the term "cancer" in a context other than the medical terminology when examining these samples, presumably because of their specificity. We retained the uni-gram term in our study for comparison; however, we focus the discussion on the results related to the bi-gram terms.

First, we examined user activity by ethnicity during each month of the study period to understand seasonal peaks in term usage on Twitter (Table 3.7). Second, we then counted the frequency of cancer terms for each month and by ethnicity. The types of cancer examined in this study include: breast cancer, prostate cancer, colorectal cancer, and lung cancer. For "cancer" related tweets, we counted the detection of the following keywords: benign, cancer(s), cancerous, carcinogen, carcinogenic, chemo, chemotherapy, chemotherapeutic, cyst(s), growths, leukemia, lymphoma, malignant, metastases, metastasis, metastatic, neoplasm, neoplasm, oncologist, oncology, radiation, radiotherapy, recurrence, and tumor(s). These sets of terms were adopted from a previous study [113]. Similar to our study, [113] counted the frequency of cancer-related tweets on Twitter and Facebook. It then compared these results to the frequency of obesity-related tweets and Facebook posts. For specific cancer types, we used the National Institute of Health's website [114] for other disease synonyms. For breast cancer, we searched for: breast cancer, breast carcinoma, cancer of the breast, malignant neoplasm of (the) breast,

malignant tumor of (the) breast, and mammary cancer. For colorectal cancer, we searched for: colorectal cancer and colon cancer. For lung cancer, we searched for: lung cancer, cancer of bronchus, cancer of the lung, lung malignancies, lung malignant tumors, lung neoplasms, malignant lung tumor, malignant neoplasm of lung, malignant tumor of lung, pulmonary cancer, pulmonary carcinoma, pulmonary neoplasms, and respiratory carcinoma. Finally, for prostate cancer, we searched for: prostate cancer, cancer of the prostate, malignant neoplasm of the prostate, prostate carcinoma, prostate neoplasm, prostatic cancer, prostatic carcinoma, and prostatic neoplasm. All searches were conducted within our tweet collection. Observable differences between Caucasian and African American groups were present in almost all of the chosen cancer terms across each month of the study period (Figure 3.3). However, observations of certain terms, namely "colorectal cancer," showed prominently lower frequency counts when compared with other terms and thus were not shown graphically.

Referencing Figure 3.3, it is important to note the sharp decreases seen following cancer awareness months (Prostate Cancer Awareness Month [PCAM, September], Breast Cancer Awareness Month [BCAM, October], and Lung Cancer Awareness Month [LCAM, November]), particularly among African Americans. Both groups are seen returning to lower frequencies following awareness months; however, this observation is more pronounced among African Americans, specifically following BCAM.

Finally, we examined the differences in term usage by race/ethnicity within each month of the study period using t-tests of pairwise differences (Table 3.8). During most months, the Caucasian and African American groups showed statistically significant differences in terms of Twitter activity. However, in terms of colorectal cancer, we observed few months where there was a statistically significant difference between these two groups. Again, we suspect this is a result of the limited number of

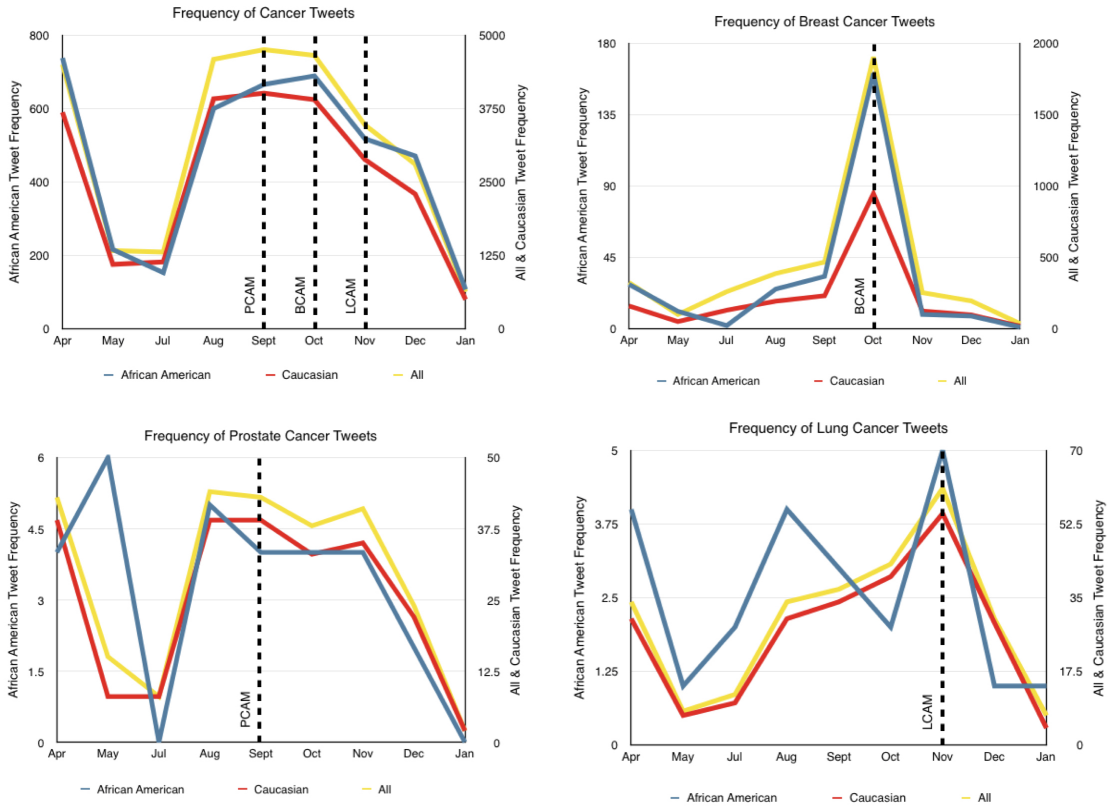

Figure 3.3 Monthly frequency of cancer terms by race/ethnicity (African American, left axis; Caucasian, right axis), and all Twitter users (right axis). Cancer terms are "Cancer" (top left), "Breast Cancer" (top right), "Prostate Cancer" (bottom left), and "Lung Cancer" (bottom right).

users discussing this particular type of cancer via Twitter. Lastly, lung cancer showed a statistically significant difference between Caucasians and African Americans during the months of September through December. For the months of January through August, statistically significant differences between groups were not detected.

3.4 Discussion

3.4.1 Principal Findings

In this study, we observed interesting patterns of media attention given to specific cancer terms among unique Twitter users during a 9-month period in 2014. With a focus on cancer in general, and breast, prostate, and lung cancers specifically, which

Table 3.8 Statistical Significance of Pairwise Differences in Cancer Term usage between African Americans and Caucasians during each Month of the Study Period

	Cancer term, t-test				
Month	Cancer	Breast Cancer	Prostate Cancer	Colorectal Cancer	Lung Cancer
April	0.00003	0.053025	0.014894	0.025347	0.080356
May	0.008194	0.584394	0.122251	0.095581	0.510364
July	0.013599	<0.0001	0.006656	0.157299	0.890133
August	<0.0001	0.001168	0.157209	0.312076	0.165111
September	<0.0001	0.00007	0.017132	0.157299	0.013196
October	<0.0001	<0.0001	0.242175	0.974206	0.000162
November	<0.0001	<0.0001	0.027708	0.014306	0.000631
December	0.000266	0.000001	0.027575	0.317311	0.000067
January	0.241671	0.00945	0.1573	0.083265	0.91944

are the leading cancers among men and women in the United States, we observed some variation in the frequency of term usage during and after specific months known to be cancer awareness months, specifically September (Prostate Cancer Awareness Month [PCAM]), October (Breast Cancer Awareness Month [BCAM], and November (Lung Cancer Awareness Month [LCAM]). Interestingly, colorectal cancer, the third most common cancer in both men and women [21], received the least attention on Twitter among users sampled in this study across the board. We observed differences in frequency of use of each of the cancer terms of interest throughout the duration of the study period by race/ethnicity, which we hypothesize are related to observable cancer disparities in the United States. These findings highlight the necessity for increased cancer awareness in the population and the importance of studying how individuals use social media to spread information about cancer, which could ultimately be utilized in the future for real-time cancer awareness intervention implemented through Twitter (and other social media channels).

Overall, we found that the frequencies of mentions of "cancer" among Caucasian and African American users were similar in terms of seasonal increases or decreases, although it appeared that African Americans maintained a higher percentage of normalized tweet frequency of this broad term compared to the Caucasian group. In terms of the frequencies of mentions of "breast cancer", Caucasian users consistently had a higher percentage of use during all months of the study period. As expected, the frequency of use of this term was highest during BCAM, with a dramatic decrease in the months following, ultimately returning to levels lower than observed leading up to BCAM. This was true among both Caucasians and African Americans; however, there was a steeper decline in the mentions of "breast cancer" on Twitter among African Americans following BCAM.

This may be an area that can be the focus of future interventions aimed at increasing breast cancer awareness throughout the year, which could contribute

to increased knowledge, improved within-guidelines screening rates, and increased preventive activities among groups with a disproportionate disease burden. For example, weekly Twitter chats hosted by the #bcsm ("breast cancer social media") community have been shown to raise awareness and decrease medical anxiety in patients [50]. Identifying individuals who were active during BCAM and inviting them to participate in Twitter chats could be a way to build an engaged, on-going community of active participants in discussions about cancer in groups with a disproportionate disease burden. Chats can be facilitated with the use of a consistent hashtag, which is a convention on Twitter designed for marking tweets about specific topics. Enlisting experts and celebrities to guest host chat sessions may be a way to promote sustained engagement, particularly because people tend to prefer health-related messages on social media that come from sources with high status and credibility [44]. These interventions would leverage Twitter's capabilities to deliver just-in-time information and social support, involving individuals proactively in evidence-based discussions about cancer throughout the year [115]. This intervention method may be appropriate for other types of cancer as well.

During PCAM, there was a substantially higher frequency of discussion of prostate cancer among Caucasians compared to African Americans. In July and January, among Caucasian users, we observed the lowest levels of prostate cancer discussion. Conversely, among African Americans, we observed a steady decrease in prostate cancer discussion from August through January. Following PCAM, we observed a decline in the frequency of use of the term "prostate cancer" among both groups; however, these declines were slower than that observed with other cancer awareness campaigns. For example, when examining the frequency of use of the term "lung cancer," we observed a peak in November (LCAM) and then a dramatic decrease to levels lower than observed in the months prior to LCAM.

The months following cancer awareness month campaigns also presented interesting findings. While awareness month campaigns (e.g., PCAM, BCAM, LCAM) could be considered successful in promoting discussion around various cancer topics, our findings suggest that these campaigns as evidenced by mentions of cancer terms via Twitter during specific cancer awareness months, did not appear to sustain long-term interest and discussion. This phenomenon was particularly evident when examining breast cancer discussion frequency, but was also present in both lung cancer and prostate cancer social media activity. In fact, our findings showed that racial/ethnic groups often returned to a state of lower participation following awareness campaigns when compared with preceding months. Notably, this reduction in discussion frequency appeared to be more prevalent among minority groups. For example, African Americans reduced their participation by 73% in the month following BCAM when compared with months preceding the program. Among Caucasians, we also saw a drop in participation where we observed only a 47% reduction. Similarly for LCAM, we observed a 50% drop among African Americans compared with a 25% drop in the Caucasian cohort. Finally, in terms of discussion of colorectal cancer, we saw poor participation throughout the months of the study. This could be an indication of poor marketing or the taboo nature of the topic among some populations as well as lack of collection of tweets during Colorectal Cancer Awareness Month (CRCAM) due to a technical issue with our data collection system.

These drops in participation are likely related to media exposure and framing, two media effects that are mediated by structural determinants of health (e.g., SES, race, and ethnicity) [116]. Media exposure is the extent to which individuals encounter information about cancer in the mass media rather than specifically seeking it out; framing describes how topics like cancer are discussed in the mass media. This finding points to the need for interventions that use appropriate framing for minority populations. For example, using Twitter to share narratives about cancer could

be particularly fruitful. Digital narratives have been successfully implemented in interventions aimed at raising awareness and improving screening rates in breast cancer, colorectal cancer, and prostate cancer [116, 117, 118]. Although tweets are short, they could be used to share short-form narratives or could be employed in conjunction with other storytelling techniques to provide engaging narratives about cancer with the aim of raising awareness and disseminating credible information about cancer to populations with a disproportionate disease burden [119].

With the growing popularity of social media and the previously unavailable personal insights it offers, social media mining presents new opportunities and methods applicable to epidemiological research. Existing studies have examined the health impacts of social media, as shown in previous work [51] where researchers concluded that Tobacco Control Programs are ineffective in capitalizing on social media platform's potential. In contrast, Thackeray et al. examined the frequency of breast cancer-related tweets during BCAM [44] and concluded that Twitter could be a tool used for increasing health conversations to maximize health marketing. In the present study, we examined how new text-mining techniques can be used to extract a user's race/ethnicity through lexical analysis, thereby providing a new opportunity to inform future studies to potentially address racial/ethnic health disparities. However, this work can be further expanded to examine differences across other demographic characteristics, as well as the investigation of disparities with respect to diseases other than cancer. Finally, understanding a social media user's demographic makeup also presents new opportunities for appropriately targeting health education materials.

3.4.2 Limitations

There were limitations of this study that should be considered. Our findings provide only a glimpse of all tweets, focused on cancer-specific topics, among users without private Twitter accounts, during one year. Thus, there could very well be an

underestimation of the frequency of cancer-focused discussion via Twitter. Relatedly, it is possible that tweets of interest were missed due to our choice of keywords or use of alternate terms and/or spellings of some words among the users. It is possible that we missed tweets of interest based on the keywords we have chosen to examine and, consequently, the true frequencies of cancer-related tweets may be higher than what we currently examined in the analysis. Nevertheless, our large-scale systematic examination of 779,653 unique Twitter users and their tweets contributed during a 9-month period would still provide a meaningful glimpse into users' social media activity related to general or specific cancer topics. We choose to report several representative case studies using the most popular cancer terms used by Twitter users. As demonstrated through these multiple case studies, commonly enabled by the proposed approach, the new method has the promise to be generically applicable for detecting, tracking, and comparing user interests regarding other cancer or disease topics. Additionally, due to technical issues with our collection system, we were unable to retain collected tweets from the middle of May through the end of July 2014, which could have contributed to the very low frequency of use of the term "colorectal cancer." In addition, March, which is CRCAM, was not included in our collection period and could also contribute to the low frequency of the term "colorectal cancer." Another possibility is that not all public tweets were delivered from the Twitter public API; but there is no way to determine the likelihood of this possibility. The collection period excluding winter and post-holiday months (late January to March) could potentially miss important patterns that may emerge through the analysis of this time period. Additionally, colorectal cancer screenings tend to target an older population, an age group that is known to not adopt social media as much.

We tested the rate at which users included the term "cancer" in a non-medical context (e.g., "He was a cancer to the community." or "I just read my cancer horoscope."). Observing a rate of 8.5% related to Zodiac signs and 2% related to

destructive practices, we recognize the potential impact on the study. However, given the relatively low rate of usage in these contexts and the likely similar rate of usage across ethnicities, we believe the impact to the overall study is limited. Additionally, disease-specific terms, such as "breast cancer" or "lung cancer" would not be plagued by such a limitation and, as a result, we focus most of our findings on disease-specific discussion disparities.

LDA has been shown to provide limited topic modeling results when applied to Twitter datasets. We attempted to mediate this limitation by adopting modified approaches suggested in existing literature, such as tweet aggregation. Given the shorter time period of the collection, the average tweet length remained relatively short and, thus, may have negatively impacted the results of the LDA approach. Future studies may look to examine how a longer study collection period might improve the LDA approach proposed in this work.

And finally, because several regional, temporal, and country-specific factors may have some influence on the contents of information shared or communicated via Twitter, we went to considerable lengths to limit our dataset to US-based users. Ideally, we would have liked to filter our dataset by a Twitter-provided variable, distinguishing US-based users from non-US-based users. However, because Twitter does not provide this information, we chose to adopt an alternate method for the extraction of US users by looking at the "Location" portion of a user's profile. This is a free-text area provided by Twitter where users can input information such as New York or San Francisco, California, excluding users with non-US locations in their profile. This method was chosen for the following two reasons: (1) only a small fraction of users provide geo-tagged tweets, and (2) it is difficult to assume that geo-tagged tweets taken internationally do not belong to a US-national. Geo-tagging of tweets varies in location for a given user and, therefore, does not provide an accurate understanding of the location a user defines as home.

3.4.3 Conclusion

This study introduces methods that have the potential to serve as a very powerful and important tool in disseminating critical prevention, screening, and treatment messages to the community in real time. These findings could help improve future social media studies, identify trends within groups of users, and target group-specific health education literature by learning users' characteristics through language differences. This study also introduced and tested a new methodology for identifying race/ethnicity among users of social media, which presents a unique opportunity to study risk profiles, risk factors and behaviors for several conditions by race/ethnicity and has significant implications in reducing disparities through targeted intervention and dissemination of evidence-based information tailored to specific racial and/or ethnic groups.

CHAPTER 4

GENDER INFERENCE IN TWITTER

4.1 Introduction

Earlier in this work, we introduced the importance of demographics in health studies. As a result of this fact, we previously introduced a method for inferring a users' ethnicity based on language usage patterns in Twitter timelines. Having the ability to infer a user's ethnicity allowed for a previously unavailable approach to analyzing cancer-related discussion patterns among ethnic groups on social media. We saw the opportunity to extend this idea by analyzing other demographic groups and their cancer-related tweeting patterns. However, when attempting to apply the synonym expansion approach introduced earlier to other demographic elements, we often were unable to report comparatively high accuracies. As a result, we explored new methods for inferring the gender of Twitter users, which is important demographic elements to medical informatics research. This work introduces the following two major contributions in Twitter text mining for the purposes of user gender inference. Below we introduce these two main contributions.

The first contribution is a new classification-based algorithm for inferring any social media user's gender information through comprehensively analyzing the text content generated and shared publicly by the user in social media. The analysis examines both the distributions of topics involved in the online user contributed content and the general language use and writing patterns exhibited in such content. Additionally, the algorithm also carefully observes the personal information self-revealed by the user in his or her social media account profile to conduct the above inference in a manner aware of the personal context. Due to the scope of this study, the proposed algorithm focuses on inferring the gender of a social media user.

The proposed method collaboratively leverages a wide spectrum of multi-modality information regarding a social media user to infer the person's gender. Such information consists of two kinds of: 1) content posted or shared by the user on social media, and 2) personal information self-revealed in the online social media profile of the user. The first kind of information listed in the above comprises: a) word choices made by a user in his or her tweets and their relationships to a predefined set of hashtags, and b) distributions of user-supplied hashtags associated with a user's tweets. For the hashtag information, it may be either manually tagged by a user during the posting time or automatically generated using the proposed standardized hashtag generation algorithm. The second type of information listed in the above consists of the frequency at which a user favorites other users' tweets, which is considered to capture the interaction of the user with other peer twitter users, a user's self-identified first name in his or her account profile on Twitter, as well as the color choices made by the user respectively concerning the foreground and background text of the profile and other profile setup choices.

Existing user demographics inference methods only examine frequencies of words used in a user's tweets to derive the person's demographics information [10, 64, 6]. As a result of the proposed approach, the number of features used to represent each user is drastically reduced by adopting a topic-based representation of users as compared to the high dimensionality associated with a bag-of-words representation of a user.

To enrich clues available for inferring a user's gender information, a new algorithm is introduced that automatically proliferates hashtags regarding user generated social media content.

In Twitter, hashtags are user-supplied tags for summarizing or highlighting key concepts or themes of content in a tweet. Unfortunately, there is no uniform or standardized ontology or vocabulary set according to which Twitter users select and

assign hashtags to individual tweets. Consequently, tweets dedicated to the same topic or carrying similar content may be assigned distinct hashtags due to the subjective personal choices made by their posting users.

Such subjectivity and diversity in the hashtags-to-tweet mapping relationship, which are independently and personally determined by individual tweet posting users, incur much difficulty and challenge for a computer algorithm to correctly, reliability, and comprehensively understand tweet content and retrieve tweets based on their associated hashtags. To make the matter more computationally difficult, not all tweets are tagged with hashtags when they are initially posted.

Twitter also allows a user to tag a tweet using multiple hashtags to increase the expressiveness of its hashtag mechanism. An example multi-hashtag tweet is as follows: *'Fashion Friday: well dressed for every occasion! #fashion #style.'*

This study exploits multi-hashtags associated with a tweet by examining the co-occurrence frequencies of multiple hashtags anchored onto a common tweet (see Figure 4.1). We look for strong relationships between hashtags by considering the frequency with which the hashtag pairs appear together with tweets. Hashtag pairs with highly frequent co-occurrences are then selected for further processing. The above hashtag clustering procedure allows us to collapse multiple related hashtags into a hashtag group, each of which is represented as a meta-hashtag or hashtag cluster.

To capture the relationship between the raw text of a tweet and its meta-hashtag(s), the proposed method further constructs a term-frequency vector that represents the occurrence frequencies of words and noun-phrases in a tweet.

It shall be noted that the proposed method applies the meta-hashtag generator onto any tweet under analysis regardless of whether it has been initially tagged by its author when it is posted. For the initially unlabeled tweets, the automatically generated meta-hashtags can be used as its meta-labels; for those initially labeled

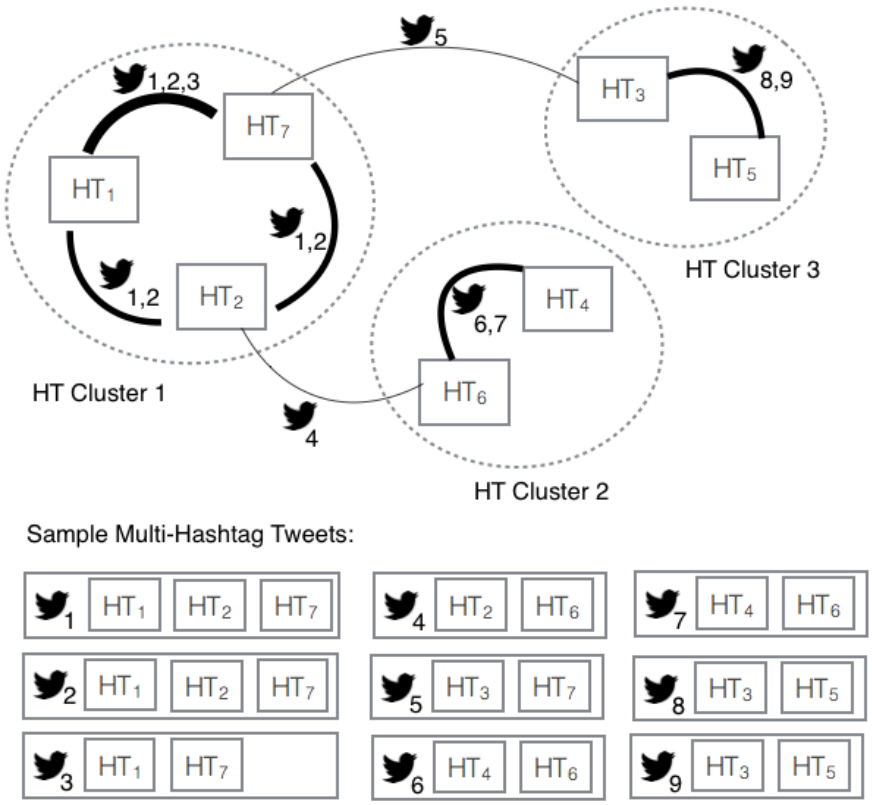

Figure 4.1 A graphical representation of the clustering of hashtags according to co-occurrence relationships in multi-hashtagged Tweets. The weight for an edge connecting two hashtags is determined by the number of tweets containing both hashtags. The more frequent the co-occurrence relationship is, the wider the edge becomes.

tweets, the algorithmically generated meta-hashtags can be nevertheless leveraged to enrich and standardize its meta-labels to mitigate the subjectivity in the tagging choices personally made by its author.

4.2 Method

4.2.1 Collecting and Preprocessing Experimental Data

Collecting Twitter Text and User Data In this study, we collected, from March 27, 2014 through April 18, 2016, 1% of publicly available English tweets using Twitter's public streaming Application Program Interface (API). The obtained collection contains 40,739,997 tweets posted by 875,937 unique users. For each tweet collected, we extract: 1) the user ID of its posting author, 2) the location of the user, if it is voluntarily disclosed in the user's Twitter profile, 3) the tweet's text content, 4) the posting date and time of the tweet, and 5) the GPS latitude and longitude of the tweet's author when the message is posted, if such information is also voluntarily disclosed. The Twitter API restricts that no more than 180 user profiles can be downloaded in any consecutive 15-minute window, which limits the population size of Twitter users studied in this work. Working with this constraint, we randomly selected 105,000 Twitter users from the aforementioned user population to gather their personal profile information on Twitter, including the user's first name, description, color choices, favorite counts, friend counts, and follower counts. Such acquired user profile information is utilized both to directly infer a user's gender and to algorithmically generate standardized hashtags for individual tweets as an additional source of features for enhancing the accuracy of user gender inference.

Data Preprocessing Due to variations in language usage patterns of English across countries, we focus on U.S.-based English speaking Twitter users in this study, even though the proposed method can be easily applied to deal with other user populations. Given this scope of study, only tweets by U.S. Twitter users

are retained in the initially collected data set; other tweets are removed from the data collection before further analysis. To detect a Twitter user's residency, we use the location indicated in the user's *home* field on his/her Twitter profile. This field accepts free-form text input provided by a Twitter user, such as *New York* or *San Francisco, CA*. To process such free-form text input, we compare each user's self-disclosed location information against a publicly available dataset of U.S. location names released by [120]. To deal with potential typos, a user is considered as a US resident if the location information supplied by the user can be matched with a location name from the aforesaid dataset within one Levenshtein character distance [121]. In the above user residency determination process, we do not utilize geo-tags associated with all tweets posted by the user for two reasons: 1) only a small fraction of Twitter users (less than 5%) posts geo-tagged tweets; 2) geo-tags of tweets do not accurately indicate a user's residency since the person can travel to a multitude of places domestically and internationally while posting geo-tagged tweets.

UGC, such as social media data collected in this study, is known to carry abundant noise, spelling errors, and region or group-specific language usage characteristics. To mitigate the impact of these issues on our user gender inference task, a pre-processing step is applied onto the tweet dataset collected before any content analysis is performed. First, special elements embedded in a tweet are extracted, including: 1) header information for a retweet, 2) user name, 3) URLs, 4) hashtags, and 5) emojis, if any of such elements is available. All these elements can be reliably detected using regular expressions applied onto a tweet's text body. Once detected, these elements are removed from the tweet's text and recorded in a separate data structure. Second, we detect and replace any numbers, contractions, and abbreviations with their plain text equivalents to standardize the text content of a tweet. For example, the sentence "We're @ 123 Main Street." is rewritten as "We are at one hundred twenty-three Main Street." Third, we detect and replace consecutive

letters in a word that repeat the same character for more than three times, which usually happens when a tweet author attempts to emphasize his/her opinion, with only two copies of the character. For example, "happyyyyy" is transformed into "happyy." This transformation can both standardize and reduce the vocabulary size of the sample tweet collection, leading to more reliable computational analysis at a downstream analysis step. Such tactic is inspired by the prior work by [122]. Lastly, words identified as common typos or special spellings are replaced with their correct and conventional English spellings following the procedure introduced in [123].

Obtaining Ground Truth User Gender Information To acquire ground truth labels regarding a user's gender for training a supervised learner, we employ the following two strategies, including: 1) looking for specific *declarative statements*, where the user explicitly indicates his or her gender. Such statements often appear in the *description* field of the user's Twitter account profile. Due to the versatile natural language expressions used to formulate these statements, such as "I'm a mother of two" or "25 year old woman," it is challenging to automatically extract these ground truth gender labels reliably and comprehensively. Therefore, we manually examine the description field in the account profile of every user in our experimental dataset to identify the user's gender information whenever it is feasible and reasonable to do so. It is noted that we exclude the description field when deriving features for user gender inference in this study because only a small percentage of users voluntarily discloses their gender information on Twitter through this field, which is 2.8% in our experimental dataset, leading into unnoticeable performance benefit to exploit the field. Testing the classification accuracy using a bag-of-words approach applied to the description field produced classification results no better than random assignment, leading us to exclude this field from the study. 2) We further check the Facebook profile of every Twitter user through an automated process to see

whether the user's Twitter profile points to the person's Facebook profile. From the Facebook profile, we can acquire the user's gender if such information is publicly available. This latter strategy was inspired by the earlier practice introduced in [73] for collecting user demographic data. After deploying both strategies in the above, we collected a labeled data set that displays the following demographic makeup: 1,492 men and 1,508 women. This ground truth dataset size is consistent with other studies conducted in this area of research [10]. These 3,000 users with known gender information collectively posted 265,418 tweets. The mean length of these tweets is 62.69 words and median length is 60 words. Figure 4.2 reports more detail on length distributions of these tweets.

4.2.2 Generating Standardized Hashtags for Tweets

As mentioned earlier, when a tweet is initially posted, its author can freely and subjectively assign one or multiple hashtags to annotate the message. The assignment decision is made according to both the content of the tweet and the individual's personal preferences and language use habits. The lack of an ontology for hashtags further increases the diverse choices available to a tweet author. As a result, the number of unique hashtags in a tweet collection may grow unlimited (See Figure 4.3). Therefore, for two tweets carrying similar or even identical content, their authors may choose different hashtags. To cope with such diversity and inconsistency in hashtag selection for tweets, the proposed method introduces a procedure that automatically assigns each tweet one or multiple standardized hashtags from a controlled vocabulary where each assigned hashtag is associated with a probabilistic value, indicating the confidence in such an assignment. Hashtags frequently used to annotate tweets carrying similar or closely related meanings are grouped into a *hashtag cluster* so that the proposed method can more effectively and reliably understand a tweet's semantics. The reason is because after reducing the large number of distinct hashtags

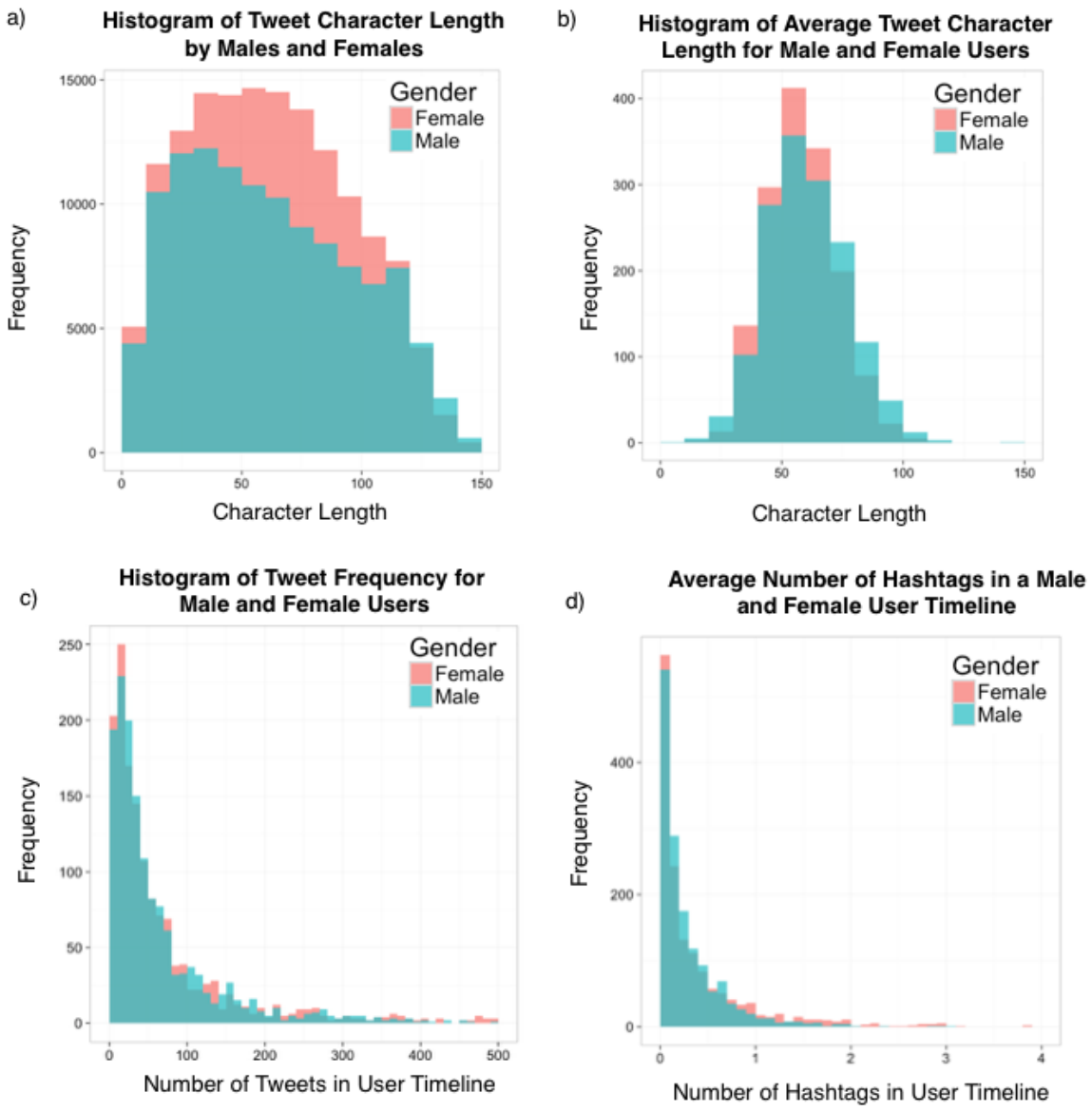

Figure 4.2 a) Distribution of tweet length in characters by male vs. female users, b) Distribution of average tweet length in characters by male vs. female users, c) Distribution of tweet frequency counts by male vs. female users, and d) Distribution of the average number of hashtags used by male vs. female users.

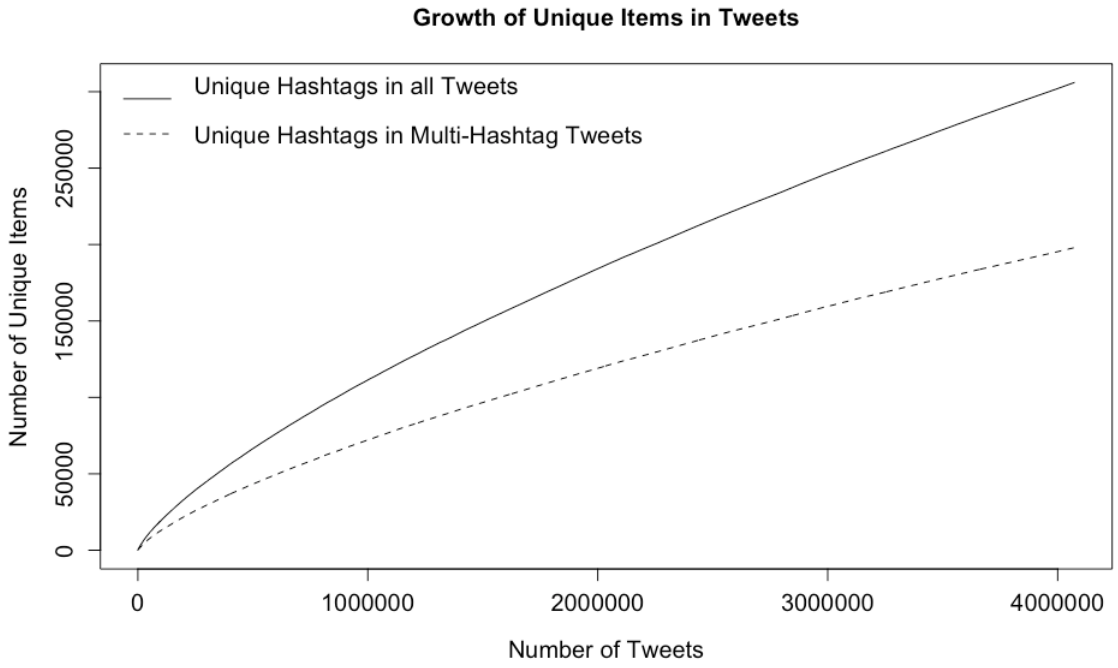

Figure 4.3 Numbers of unique hashtags in our experimental tweet collection.

originally associated with tweets into a much smaller number of hashtag clusters, the problem space for learning to infer a Twitter user's gender is also greatly reduced, enabling more accurate and efficient computational inference.

Extracting Co-Occurring Hashtags The automatic hashtag generation procedure starts with a set of seed hashtags, each of which is later observed as a topic feature in the proposed gender inference model. These seed hashtags are carefully chosen such that every included hashtag would meaningfully contribute to the concerned gender inference task. According to prior studies conducted in online broadcasting and media research by [124], the following 19 topics are selected due to their demonstrated connections with the gender of a message's author, including: *hobby, school, music, shopping, video games, movies, television, sports, society, news, religion, alcohol, sex, depression, loneliness, violence, friends, family,* and *romance*. We use the notation $\mathcal{S}(i)$ to refer to the i-th seed topic listed above. For each seed hashtag, the generation procedure first identifies hashtags frequently co-appearing with the

given seed hashtag. For instance, for the seed hashtag "sports," multiple user elected hashtags, such as "basketball," "NBA," and "NCAA," would frequently co-occur.

In this study, we exploit multiple hashtags commonly associated with a tweet to derive semantic relatedness among hashtags. For example, from a tweet "I just booked a trip to Florida. #vacation #beach," we can detect that the two hashtags "#vacation" and "#beach" may be related. If we witness the co-occurrence of a pair of hashtags repeatedly, it would be reasonable to infer that the two involved hashtags are closely related. The frequency with which a hashtag co-appears with another hashtag implies how closely the two hashtags are related. Figure 4.1 illustrates the main idea behind this hashtag co-occurrence mining process.

To identify hashtags frequently co-occurring with each seed hashtag respectively, a hashtag co-occurrence matrix $M_{i,j}$ is constructed, whose dimensionality is $N \times N$ where $N = 197,958$, which is the number of unique hashtags appearing in a randomly sampled tweet sub-set of our experimental dataset. This sub-set consists of $D = 4,073,999$ tweets, comprising 10% of tweets in the entire experimental dataset. This tweet sampling step is introduced to accelerate the computational analysis. Empirically, we find that increasing this sampling rate does not noticeably affect the experimental results, suggesting the adequacy of the sampling rate for this study. $M_{i,j}$ is initialized according to pair-wise hashtag co-occurrence relationships exhibited in the sub-set. Let d_k be the k-th tweet and t_i be the i-th hashtag appearing in the subset. Given the symmetry of the co-occurrence relationship, $M_{i,j}$ is represented as an upper-triangular matrix as follows:

$$M_{i,j} = \sum_{k=1}^{D} \sum_{i=1}^{N} \sum_{j=1}^{N} Pair(k, i, j), \tag{4.1}$$

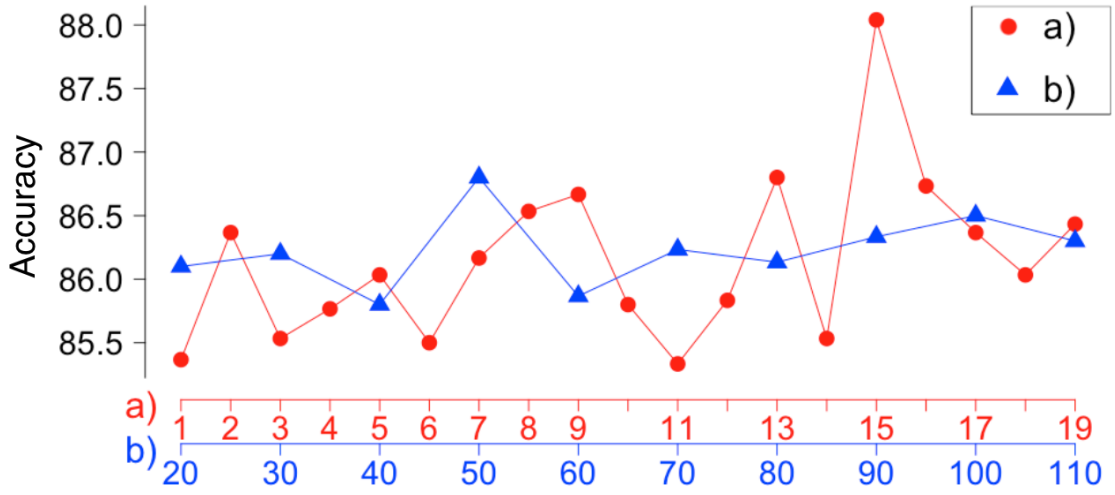

Figure 4.4 a) Gender inference accuracy when different quotas are adopted to select candidate hashtags for each seed hashtag. The quota size of 15 is adopted in our model implementation due to its optimal experimental performance. b) Gender inference accuracy when different pairwise hashtag occurrence counts are used to select candidate hashtags related to a seed hashtag.

where:

$$Pair(k,i,j) = \begin{cases} 0: & \text{if } i \geq j; \\ 1: & \text{if } (t_i \in d_k) \wedge (t_j \in d_k) \wedge (i < j); \\ 0: & \text{otherwise.} \end{cases}$$

The aforesaid automatic hashtag generation procedure consists of two key steps, including: 1) identifying hashtags related to a seed hashtag and 2) generating standardized hashtags based on text analysis, which will be respectively explained below.

Identifying Hashtags Closely Related to a Seed Hashtag Most of hashtags harvested from the above sub-sampled tweet collection are unrelated to any of the seed hashtags $\mathcal{S}(i)$ $(i = 1, \cdots, 19)$. It may be tempting to prune these unrelated hashtags when constructing the co-occurrence matrix, $M_{i,j}$. We choose not to do so to preserve the semantically revealing co-occurrence relationships among hashtags to carry out a propagation-based hashtag clustering procedure (see detail in Section

4.2.3); otherwise, if hashtags beyond the seed ones are pruned in advance, we will forfeit a number of hashtag co-occurrence relationships, resulting in a much more sparse co-occurrence matrix with greatly reduced semantic clues. For each seed hashtag, the top fifteen hashtags attaining the highest co-occurrence counts with the seed hashtag are selected. It is noted that the sets of most frequently occurring hashtags for two distinct seed hashtags may overlap. For example, a hashtag like "#entertainment" may frequently co-appear with both *television* and *sports*; similarly, "#love" frequently co-appears with both *family* and *friends*. As a result, the number of hashtags identified as candidate hashtags is 238 rather than 285 (=19 topics $\times$ top 15 hashtags per topic). We refer to the set of candidate hashtags as $\mathcal{CH}$ from now on. Figure 4.4.(a) shows that the number of frequently co-occurring hashtags considered for each seed hashtag is experimentally optimized in this work.

We also considered an alternative approach to selecting candidate hashtags for each seed hashtag using a threshold based on an absolute count of hashtag co-occurrences. The alternative approach accepts a hashtag if its co-occurrence count with a seed hashtag exceeds the given threshold. In case that a hashtag co-appears with multiple seed hashtags, the highest co-occurrence count with a seed hashtag is considered for making the admission decision.

Among the above two approaches, the first fixed quota-based selection mechanism produces a gender identification accuracy superior to the latter co-occurrence count-based approach, the experimental evidence of which is shown in Figure 4.4. We assume the reason behind the relative performance advantage of the first approach is because the number of hashtags co-appearing with a seed hashtag varies noticeably from one to another, which calls for a floating threshold if the absolute number of co-occurrence count is observed to select candidate hashtags. In contrast, the fixed quota-based selection mechanism copes more suitably with such disparity and adaptive admission need, leading to better experimental performance.

Automatic Generation of Standardized Hashtags based on Text Analysis
In principle, a tweet's author is expected to select the message's hashtags according to the content conveyed through the tweet. To emulate this process, we may computationally generate standardized hashtags based on a tweet's text, which will be free from the aforementioned diversity and subjectivity issues commonly witnessed with the conventional practice of manual hashtag selection for tweets.

Single-word-based features, also known as uni-grams, have been previously used to construct generative language models for tweets, e.g., [125]. However, uni-grams often provide insufficient clues for comprehensive text understanding. To cope with this limitation, in this study, we leverage multi-grams combined with uni-grams as features to construct a hashtag generation model for tweets. Through supervised learning, the model analyzes a tweet's text to derive its likely hashtags. This model design is inspired by the previous work of [126], which adopts noun phrase-based multi-grams for topic detection from tweets.

To carry out the text analysis based approach for standardized hashtag generation, first, the cleaned text of each tweet is tokenized. For every extracted token, a corresponding part-of-speech (POS) tag is assigned. Both steps are performed using the software package developed by [127]. Next, utilizing the sequence of POS-tagged text tokens, the proposed method extracts noun phrases from the sequence. The extracted noun phrases are subsequently used as multi-gram features in the hashtag generation model. Constructing features on the granularity of multi-grams rather than uni-grams enables the model to capture the semantic topics underlying a tweet more accurately and comprehensively. The reason why we utilize noun phrases to construct multi-gram features is that these phrases are frequently employed to specify key semantic entities in a tweet, such as *people*, *places*, *things*, *times*, and *locations*, but not other less essential content in the tweet. The representation is based on the assumption that the above five categories of entities

largely determine the meaning of a tweet, which in turn significantly influence the hashtag selection choice for a tweet. Such a representation is more effective than alternative representation methods reported in prior studies, such as the work by [6] where explicit choices are made regarding the size of n-gram features, where terms are represented using uni-grams and n-grams for n up to five. For instance, using the representation based on noun-phrases of varying lengths, our method would nicely capture semantics for the expression "United States Congress" from the sentence "I made my first trip to the United States Congress.". In contrast, using the traditional fixed length n gram-based representation, a set of uninformative features such as "States Congress," "to the United," "made my first trip to," are introduced, none of which properly captures the semantic concept underlying the tweet text. To identify the aforesaid noun phrases from a tweet, we apply the regular expression-based detection method proposed by [128] onto the POS tagged sequence of text tokens identified from each tweet. The adopted regular expression is as follows:

$$\Big((A|N)^{+}|\big(((A|N)^{*}(NP)^{?})(A|N)\big)\Big)N, \tag{4.2}$$

where A stands for an adjective; N stands for a lexical noun; P stands for a preposition; the superscripts "+," "?," and "*" respectively indicate the cases where a concerned pattern appears one or multiple times, does not appear at all or appears only once, and does not appear or appears one or multiple times; parenthesis indicates grouping among POS tags; | acts as an OR operator.

The number of occurrences of each noun-phrase, which is represented as a multi-gram, in the entire experimental dataset is further recorded in a document-term matrix where uni-grams are also recorded as a special case for representation comprehensiveness. In addition, verbs are further captured in our representation for tweet content as a special type of uni-grams. To construct the hashtag generation model, we select tweets from our experimental collection that have been manually

labelled using one of the 238 candidate hashtags identified by a tweet's author (see Section 4.2.2). In this selection process, we also discard all tweets with length fewer than 30 characters. The reason is because such short tweets usually do not carry sufficient text to properly convey a meaningful signal for learning to capture the relationships between a tweet's text and its hashtags. Such an assumption is also empirically verified through our experiments.

From all tweets selected through the above procedure, a fixed number of tweets is randomly selected for training to generate each of the 238 candidate hashtags for any given tweet. We experimented with using a varying size of these training samples, ranging from 25 tweets per hashtag to 150 in increments of 25 tweets. The lowest overall accuracy for gender inference was observed at 2.39% for the case when 25 tweets are used to train the generation for a hashtag. The highest accuracy was obtained when 100 tweets are used at training to generate a hashtag. As the number of samples in the training set increased, the number of unique vocabulary terms also increased, likely resulting in an over fit model as indicated by the reduced accuracy on the validation set for models developed on a larger training size. Given the above experimental exploration, we choose to use 100 samples to train to generate a hashtag, leading to a training collection of 23,800 tweets in total (=238 $\times$ 100 samples).

For each tweet selected as a training instance, the aforementioned uni-gram and noun-phrase based variable length multi-gram features are first extracted, which produces 26,737 distinct uni-grams and 3,126 distinct multi-grams, with a total of 29,863 text gram-based features. We subsequently construct a hashtag–gram matrix (HG), which is of the dimensionality of 23,800 $\times$ 29,863. For example, for a tweet, "I'm going to see the LA Lakers play against the Chicago Bulls with some friends, tonight. #basketball," a sparse row vector of dimensionality 29,863 can be constructed as a training record, which indicates the presence of multi-grams (*LA Lakers, Chicago Bulls*) and uni-grams (*Bulls, Chicago, Friends, Going, Lakers, Play,*

See, Tonight) associated with the tweet and the absence of other text gram-based features for the tweet. Given the gold standard records supplied by the matrix HG, a multi-classification support vector machine (SVM) model is trained using the one vs. one training scheme [101] to construct our text analysis-based hashtag generator for tweets. We choose the learning model due to its satisfactory ability to work with text data represented by document-term matrices as abundantly reported in the literature, e.g., [129, 94, 130, 131]. This classification model aims to capture the generative relationship between each candidate hashtag in $\mathcal{CH}$ and the vocabulary used in a tweet. The output corresponding to the example above is a vector of 238 dimensions, where the vector component corresponding to the seed hashtag *#basketball* is set to 1 and all other vector components set to 0. This model, which classifies into one of the seed hashtags with 11.38% accuracy, intends to roughly estimate the topic of a tweet for aggregation at a later processing step (see Section 4.2.3). The candidate hashtags generated through this step will be further aggregated into the 19 topics represented in $\mathcal{S}$ through the procedure introduced in Section 4.2.3.

For each tweet, we first extract the aforesaid feature vector based on multi-grams and uni-grams. If an extracted gram is not considered by the generation model, it is simply ignored by the model. The generation model then outputs a 238-dimensional vector, whose j-th component indicates the tweet's strength of association with each candidate hashtag, $\mathcal{CH}_j (j = 1, \cdots, 238)$.

Using the approach proposed by [132] to converting a SVM classification result into a probability distribution over all potential class labels, we produce a probability distribution over $\mathcal{CH}$ for each tweet. In our context, for each pair of candidate hashtags, we compute a pair-wise class probability score $r_{x,y}$ among the $k = 238$

candidate hashtags by solving the following system:

$$\begin{aligned} \forall x, p_{\mathcal{CH}_x} &= \sum_{y:y\neq x} \Big(\frac{p_{\mathcal{CH}_x} + p_{\mathcal{CH}_y}}{k-1}\Big) r_{x,y} \\ \forall x \sum_{x=1}^{k} p_{\mathcal{CH}_x} &= 1, p_{\mathcal{CH}_x} \geq 0 \end{aligned} \tag{4.3}$$

where $p_{\mathcal{CH}_x}$ is the probability of a tweet associated with the x-th candidate hashtag for $x = 1, \cdots, 238$, and the values of $r_{x,y}$ and $p_{\mathcal{CH}_x}$ are determined by minimizing the following equation:

$$\min_{r_{x,y}, p_{\mathcal{CH}_x}} \sum_{x=1}^{k} \Big(\sum_{y:y\neq x} r_{x,y} p_{\mathcal{CH}_x} - \sum_{y:y\neq x} r_{x,y} p_{\mathcal{CH}_y} \Big)^2. \tag{4.4}$$

4.2.3 Deriving Feature Vectors to Characterize Topic Distributions in a User's Timeline

Classifying tweets into a controlled set of pre-defined topics enables the proposed method to more efficiently and effectively examine any potential relationship between topics latent in a user's tweets and the person's gender, the advantage of which will be demonstrated through experimental results reported later in this article. According to previous studies, e.g., [133, 134, 135], topic-based semantic modeling and mining generally perform superiorly to traditional bag-of-words-based modeling practice due to the former approach's representation effectiveness and conciseness.

To derive a feature vector for characterizing latent topic distributions underlying a user's Twitter timeline, this study adopts two alternative approaches, including a method that examines pairwise hashtag co-occurrence relationships and another method that exploits point-wise mutual information between pairs of hashtags.

Constructing a Matrix of Pairwise Hashtag Co-Occurrence Relationships

The initial hashtag co-occurrence matrix $M_{i,j}$ constructed earlier at Section 4.2.2 only represents the direct co-occurrence relationship between a pair of hashtags a and b,

the relationship of which is denoted as $a \sim b$. The matrix however does not explicitly capture any indirect co-occurrence relationships among a group of hashtags. For example, when $a \sim b$ and $b \sim c$, $a \sim c$ is not directly encoded in the matrix. To facilitate the exploitation of such indirect co-occurrence relationships among hashtags for deriving an expressive feature vector on topic distributions of a Twitter user's timeline, the proposed method carries out a co-occurrence relationship propagation process as follows.

Given $M_{i,j}$, we first normalize the matrix through respectively dividing elements in each row of the matrix by the maximum element of the row such that every matrix element is normalized into the range of $[0, 1]$. The resulting matrix is defined as $M_{i,j}^{\text{norm}}$. Starting with $M_{i,j}^{\text{norm}}$, we can propagate explicitly represented direct co-occurrence relationships between hashtags to derive indirect co-occurrence relationships via Equation (4.5):

$$M_{i,j}^{\text{pro}} = \theta\big(M_{i,j}^{\text{norm}} + \alpha\theta([M_{i,j}^{\text{norm}}]^2) + \alpha^2\theta([M_{i,j}^{\text{norm}}]^3)\big). \tag{4.5}$$

This equation carries a parameter $\alpha \in [0, 1]$, which controls the attenuation effect modeled in the propagation process where a smaller value of α dampens the propagated impact on $M_{i,j}^{\text{pro}}$ more significantly. The equation also carries a thresholding function θ, which specifies when a signal of uncertainty shall be filtered. With its aid, all matrix elements smaller than a threshold θ_0 are set to zero to eliminate highly uncertain signals introduced in the propagation process. The above formula only models the effect of propagation up to two rounds. In principle, formulas of higher orders can be deployed to model additional rounds of propagation. However, due to the aforementioned attenuation and uncertain signal elimination effects, we experimentally verify that those formulas of higher orders do not bring noticeable performance benefit. Experimentally, we further find that the proposed method attains its highest user gender inference performance when configured using the

parameters, $\alpha = 0.25$ and $\theta_0 = 0.3$, under the non-linear weighting scheme; For the binary weighting scheme, the optimal parameters used to configure the proposed method are $\alpha = 0.3$ and $\theta_0 = 0.55$ (see detail in Section 4.2.3).

Constructing a Matrix of Point-Wise Mutual Information In the procedure introduced at Section 4.2.3, this work examines the propagation of hashtags via the proposed propagation equation shown in Equation 4.5. Associative relationships between hashtags can be examined as well and thus, the proposed method employs a second approach that examines the point-wise mutual information (PMI) extracted from hashtags in the sample tweet collection. PMI was originally introduced to represent word association norms derived from a corpus [136]. The reason that we adopt this metric is due to the increasing popularity of PMI deployed to discover latent word relationships embedded in social media data, e.g., the work by [137, 138]. The word association score, PMI, is computed using the occurrence probability of two hashtags along with the probability of the joint occurrence of these words in a document, i.e.,:

$$PMI(a, b) = log\left(\frac{P(a \cap b)}{P(a)P(b)}\right). \tag{4.6}$$

To apply Equation (4.6) in this study, we estimate $P(a)$ as the probability that a hashtag "a" appears in the sample tweet collection, i.e.,:

$$P(a) = \frac{\sum_{k=1}^{D} F(a,k)}{D}; F(a,k) = \begin{cases} 1: & a \in \mathcal{d}_k; \\ 0: & \text{otherwise.} \end{cases} \tag{4.7}$$

In Equation (4.7), D is the number of tweets in the collection, i.e., 4,073,999 as discussed earlier in Section 4.2.1, $\mathcal{d}_k$ is the k-th tweet in the collection, and $P(a \cap b)$

is estimated as follows:

$$P(a \cap b) = \frac{\sum_{k=1}^{D} F(a,b,k)}{D}; F(a,b,k) = \begin{cases} 1: & a \in \mathcal{d}_k \wedge b \in \mathcal{d}_k; \\ 0: & \text{otherwise.} \end{cases} \tag{4.8}$$

Finally, we normalize the PMI value derived through the above equations using the method suggested in [139], i.e.,:

$$PMI^{\text{norm}}(a,b) = \frac{PMI(a,b)}{-log[P(a \cap b)]} = \frac{log[P(a)P(b)]}{log[P(a \cap b)]} - 1. \tag{4.9}$$

The normalized PMI score is bounded in the range of $[-1, 1]$, which exhibits the following useful properties: 1) if two terms a and b are mutually exclusive, $PMI^{\text{norm}}(a,b) \rightarrow -1$; 2) if the two terms occur independently, $PMI^{\text{norm}}(a,b) \rightarrow 0$ since $log[P(a)P(b)] = log[P(a \cap b)]$; 3) if the two terms always co-occur, $PMI^{\text{norm}}(a,b) \rightarrow 1$. The above properties of the normalized PMI metric make it well-suited for performing analysis in this study.

Utilizing the normalized PMI metric, we can construct a PMI matrix through Equation (4.10):

$$PMI^{\text{norm}}_{i,j} = \begin{bmatrix} PMI^{\text{norm}}(1,1) & \dots & PMI^{\text{norm}}(1,j) \\ \vdots & \ddots & \vdots \\ PMI^{\text{norm}}(i,1) & \dots & PMI^{\text{norm}}(i,j) \end{bmatrix}, \tag{4.10}$$

where i and j respectively correspond to the i-th and j-th unique hashtags in the collection D.

Deriving Feature Vectors to Characterize Topic Distributions in User Timelines Assume a user, u_i, contributes k tweets to the collection, which are assumed to be $\mathcal{d}_j$ $(j = 1, \cdots, k)$ without the loss of generality. We first construct a matrix $\boldsymbol{\mathcal{CH}}(u_i)$ to represent the topic distributions of these k tweets over the set of

topics respectively represented by the 238 candidate hashtags, $\mathcal{CH}_i$ $(i = 1, \cdots, 238)$, as follows:

$$\boldsymbol{\mathcal{CH}}(u_i) = \begin{bmatrix} P(\mathcal{CH}_1|\mathcal{d}_1) & P(\mathcal{CH}_2|\mathcal{d}_1) & \dots & P(\mathcal{CH}_{238}|\mathcal{d}_1) \\ P(\mathcal{CH}_1|\mathcal{d}_2) & P(\mathcal{CH}_2|\mathcal{d}_2) & \dots & P(\mathcal{CH}_{238}|\mathcal{d}_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(\mathcal{CH}_1|\mathcal{d}_k) & P(\mathcal{CH}_1|\mathcal{d}_k) & \dots & P(\mathcal{CH}_{238}|\mathcal{d}_k) \end{bmatrix}. \tag{4.11}$$

Each element $P(\mathcal{CH}_l|\mathcal{d}_k)$ recorded in $\boldsymbol{\mathcal{CH}}(u_i)$ indicates the likelihood that the tweet $\mathcal{d}_k$ is associated with the candidate hashtag $\mathcal{CH}_l$, which can also be interpreted as the strength of the topic represented by $\mathcal{CH}_l$ embodied in $\mathcal{d}_k$. In this work, each matrix element of $\boldsymbol{\mathcal{CH}}(u_i)$ is estimated using the text analysis-based hashtag generation model introduced in Section 4.2.2.

Once $\boldsymbol{\mathcal{CH}}(u_i)$ is constructed, we can now derive feature vectors to characterize topic distributions underlying a Twitter user's timeline through leveraging either matrix prepared in Sections 4.2.3 and 4.2.3. Let $TW(u_i)$ be the full set of tweets posted by user u_i in our experimental data set, M be either the matrix $M^{\text{pro}}_{i,j}$ constructed in Section 4.2.3 or the matrix $PMI^{\text{norm}}_{i,j}$ constructed in Section 4.2.3. Let $M_{\ell,k}$ be the element retrieved from M that corresponds to the ℓ-th seed hashtag, $\mathcal{S}_\ell$, and the k-th candidate hashtag, $\mathcal{CH}_k$. Note that either version of M is an $N \times N$ matrix containing pairwise information between all hashtags in the experimental collection. We then introduce a filtering function $\delta(\mathcal{CH}_k, \mathcal{TV}_\ell, \beta)$, in which $\beta \in [0, 1]$ is a filtering threshold value indicating whether a matrix element $M_{\ell,k}$ carries a non-trivial number in a binary way, i.e.,:

$$\delta(\mathcal{CH}_k, \mathcal{TV}_\ell, \beta) = \begin{cases} 1: & \text{if } M_{\ell,k} > \beta; \\ 0: & \text{otherwise.} \end{cases} \tag{4.12}$$

Based upon the above notations, we can construct a feature vector to characterize topic distributions in user u_i's Twitter timeline $TW(u_i)$ as follows:

$$\boldsymbol{\mathcal{TV}}_{\text{binary}}(u_i, \beta, n, c) = \begin{bmatrix} \sum_{k=1}^{238} \delta(\mathcal{CH}_k, \mathcal{TV}_1, \beta) \sum_{\mathcal{d}_j \in TW(u_i)} P(\mathcal{CH}_k | \mathcal{d}_j) \\ \sum_{k=1}^{238} \delta(\mathcal{CH}_k, \mathcal{TV}_2, \beta) \sum_{\mathcal{d}_j \in TW(u_i)} P(\mathcal{CH}_k | \mathcal{d}_j) \\ \vdots \\ \sum_{k=1}^{238} \delta(\mathcal{CH}_k, \mathcal{TV}_n, \beta) \sum_{\mathcal{d}_j \in TW(u_i)} P(\mathcal{CH}_k | \mathcal{d}_j) \end{bmatrix}. \tag{4.13}$$

In (4.13), a binary weighting scheme is deployed to derive the topic distribution in a user's Twitter timeline by aggregating the density of each candidate hashtag in each tweet of a user, which is captured by $P(\mathcal{CH}_k|\mathcal{d}_j)$ where the aggregation weighting factor $\delta(\mathcal{CH}_k, \mathcal{TV}_\ell, \beta)$ controls, in a binary fashion, whether the relatedness of a given candidate hashtag, $\mathcal{CH}_k$, with a topic $\mathcal{TV}_\ell$ shall be considered during the feature vector derivation process. Finally, c can take the values of "pro" or "PMI" depending on the matrix from which $\boldsymbol{\mathcal{TV}}$ is derived, the hashtag propagation or PMI matrix, respectively.

It is noted that topics discussed on Twitter can be vague or ambiguous, e.g., "Today is a great day." The scope of potential topics present can also be vast. Due to the limited volume of labeled training data available, we cannot generate an exhaustive list of meta-hashtags, one for each possible topic mentioned by a user. For this reason, we create a special category named "other." If the aforementioned candidate hashtag is not successfully assigned to any topic in $\boldsymbol{\mathcal{TV}}$ in the above feature vector derivation process, the candidate hashtag label will be attributed to the "other" category, which is represented as the 20-th dimension of $\boldsymbol{\mathcal{TV}}(\cdot, \cdot, 20)$. The counterpart topic distribution vector that does not consider the "other" category is noted as

$\boldsymbol{\mathcal{TV}}(\cdot,\cdot,19)$. We use a similar parameter for controlling the inclusion or exclusion of the "other" category later in Equation (4.15).

To cope with the possible non-linear relationship between the matrix element $M_{\ell,k}$ and a proper measure regarding the relatedness between a candidate hashtag $\mathcal{CH}_k$ and a topic $\mathcal{TV}_\ell$, we additionally adopt a sigmoid-shaped weighting function, ψ, which has been shown generally effective in tackling classification tasks [140], to derive the feature vector on topic distributions. ψ is defined as:

$$\psi(\mathcal{CH}_k, \mathcal{TV}_\ell, \gamma) = \frac{1}{1 + e^{-\gamma M_{\ell,k}}}, \tag{4.14}$$

where $\gamma \in (0, \infty)$ is a parameter controlling the shape of the weighting function. Using the notation of $\psi(\mathcal{CH}_k, \mathcal{TV}_\ell, \gamma)$, we can construct a non-linearly weighted version of the topic distribution vector for user u_i's Twitter timeline as follows:

$$\boldsymbol{\mathcal{TV}}_{\text{non-linear}}(u_i, \gamma, n, c) = \begin{bmatrix} \sum_{k=1}^{238} \psi(\mathcal{CH}_k, \mathcal{TV}_1, \gamma) \sum_{d_j \in TW(u_i)} P(\mathcal{CH}_k | d_j) \\ \sum_{k=1}^{238} \psi(\mathcal{CH}_k, \mathcal{TV}_2, \gamma) \sum_{d_j \in TW(u_i)} P(\mathcal{CH}_k | d_j) \\ \vdots \\ \sum_{k=1}^{238} \psi(\mathcal{CH}_k, \mathcal{TV}_n, \gamma) \sum_{d_j \in TW(u_i)} P(\mathcal{CH}_k | d_j) \end{bmatrix}. \tag{4.15}$$

The parameters $\beta \in [0, 1]$ in the δ function and $\gamma \in [0, \infty]$ in the ψ function are used for thresholding purposes, which control how many candidate hashtags are considered when deriving a user's topic distribution feature vector. Like Equation (4.13), the computing procedure defined in Equation (4.15) may consider a candidate hashtag multiple times for deriving the density distribution over multiple topics where each topic is represented by a candidate hashtag. This design choice was elected to allow fuzzy considerations for topic modeling. For example, consider a candidate hashtag (*#love*), which is related to both seed hashtags (*#romance*) and

(*#family*) according to the co-occurrence relationship matrix or the PMI matrix respectively constructed in Sections 4.2.3 and 4.2.3. The design of both Eqs. (4.13) and (4.15) is able to capture such a relationship, leading to more comprehensive and reliable extraction and modeling of topic distributions in a user's timeline. In Section 4.3.4, we report that when $\gamma = 1.0$ and $\beta = 0.25$, the proposed method attains its highest accuracy in user gender inference. Both the binary and non-linear approach to weighting $\boldsymbol{\mathcal{TV}}$ produces a similar maximum accuracy (88.3%), albeit with a different set of optimized parameters. However, in general, the non-linear model outperformed the linear model when looking at classification accuracies across experimental conditions, particularly when combining hashtag propagation and PMI features together, thus producing an accuracy of 88.6%. To produce an accuracy of 88.6%, the following parameters were used: $\boldsymbol{\mathcal{TV}}^{\text{norm}}_{\text{non-linear}}(u_i, 19, pro) + \boldsymbol{\mathcal{TV}}^{\text{norm}}_{\text{non-linear}}(u_i, 19, PMI) + \boldsymbol{\mathcal{NM}}(u_i) + \boldsymbol{\mathcal{PV}}(u_i)$ with $\alpha = 0.25$, $\theta = 0.05$, $\gamma = 3.0$ for the non-linear weighting scheme.

Considering the disparity in the intensity of individuals' Twitter posting activities, to make our topic distribution vectors more comparable when dealing with users of varying Twitter posting intensities, we further derive a normalized version of the above topic distribution vector, $\boldsymbol{\mathcal{TV}}$, by dividing each set of row element in $\boldsymbol{\mathcal{TV}}$ by its corresponding row-maximum. Going forward, we will refer to the normalized version of $\boldsymbol{\mathcal{TV}}$ as $\boldsymbol{\mathcal{TV}}^{\text{norm}}$.

Two versions of $\boldsymbol{\mathcal{TV}}^{\text{norm}}$ are derived, one for $\boldsymbol{\mathcal{TV}}_{\text{binary}}$ and another for $\boldsymbol{\mathcal{TV}}_{\text{non-linear}}$, the result of which are respectively denoted as $\boldsymbol{\mathcal{TV}}^{\text{norm}}_{\text{binary}}$ and $\boldsymbol{\mathcal{TV}}^{\text{norm}}_{\text{non-linear}}$.

4.2.4 Deriving Additional Features for User Gender Inference

To accurately infer the gender of a Twitter user, the proposed method leverages additional features besides the aforementioned topic distribution features as follows.

Table 4.1 An Overview of Features Used in This Study

Feature	**Description**	**Dims/ Data Type/ Range**
$\mathcal{NM}$	The probability of a user being male.	1/ Float/ [0,1]
$\mathcal{TV}(u_i, 20, c)$	The thematic sum of scores automatically generated hashtags for a given user concerning the 19 key topics and an "other" category.	20/ Float/ $[0,\infty]$
$\mathcal{TV}_{\mathrm{R}}(u_i)$	The distribution of a user's annotated hashtags over the 19 key topics according to the occurrence counts of these hashtags in the user's timeline. For addition to $\mathcal{TV}(u_i, 20, c)$, the 20-th element of $\mathcal{TV}_{\mathrm{R}}$ is set to zero.	19 or 20/ Integer /$[0,\infty]$
$\mathcal{TV}(u_i, n, c)$ $\biguplus$ $\mathcal{TV}_{\mathrm{R}}$	The vector sum of $\mathcal{TV}$ and $\mathcal{TV}_{\mathrm{R}}$.	19 or 20/ Float/ $[0,\infty]$

Table 4.1 (Continued) An Overview of Features Used in This Study

Feature	**Description**	**Dims/ Data Type/ Range**
$\mathcal{TV}(u_i, 20, c)$ $\|\|$ $\mathcal{TV}_{\mathrm{R}}$	$\mathcal{TV}$ concatenated with $\mathcal{TV}_{\mathrm{R}}$. (Dimensionality: 19+1+19).	39/ Float/ [0,∞]
$\mathcal{PV}$	Feature set extracted according to a user's profile.	8/ Integer/ [1,8]
$\mathcal{BOW}$	Bag-of-words representation of a user's timeline.	3,000 by 125,951 / Integer/ [0,∞]
$\mathcal{BUR}$	Binary document-term matrix comprised of text from the *i-th* user's posts, profile description, and name in a 5-character gram expansion.	3,000 by 998,095 / Integer/ [0,∞]

Features Derived from User Profiles The proposed method derives another set of features, represented as a feature vector $\boldsymbol{\mathcal{PV}}$, to characterize a user's account activities and inter-user activities on Twitter. $\boldsymbol{\mathcal{PV}}$ comprises a Twitter user's relative frequency of: number of followers, number of friends, and number of favorited tweets, which are encoded as the first three components of $\boldsymbol{\mathcal{PV}}$. Each of the feature components is computed over the entire period covered by the experimental data collection. In addition, the proposed method extracts features regarding a user's personal choices in setting up his/her Twitter account profile. Specifically, the color choices for the foreground, background, sidebar, sidebar border, and links on a user's Twitter profile are extracted according to the corresponding HTML color codes used in the account profile page. These colors are then discretized into 14 broad color classes through a color wheel-based approximation method introduced in [141], i.e., (1-Red, 2-Flush Orange, 3-Yellow, 4-Chartreuse, 5-Green, 6-Spring Green, 7-Cyan, 8-Azure Radiance, 9-Blue, 10-Electric Violet, 11-Magenta, 12-Rose, 13-White, 14-Black). Therefore, for each user u_i, a five dimensional color choice vector is extracted in the form of $\boldsymbol{\mathcal{CL}}(u_i) = \big(Foreground(u_i), Background(u_i), Sidebar(u_i), SidebarBorder(u_i), Link(u_i)\big)$. $\boldsymbol{\mathcal{CL}}(u_i)$ is used to define the last five feature components of $\boldsymbol{\mathcal{PV}}$. Overall, $\boldsymbol{\mathcal{PV}}$ has eight dimensions.

Deriving Personal Features for a User The method further extracts the self-disclosed first name of a user u_i by retrieving the first string listed in the *Name* field of the person's Twitter profile. Such first name information is subsequently compared with the Social Security Administration's (SSA) Name-Gender Frequency Dataset [142]. This dataset provides the top 1,000 names and the frequency of each such name adopted each year for male versus female newborns during the aggregated period of 1950–2013. This data was used to compute the likelihood of a popular first name given to a user of a specific gender. Specifically, the method represents the

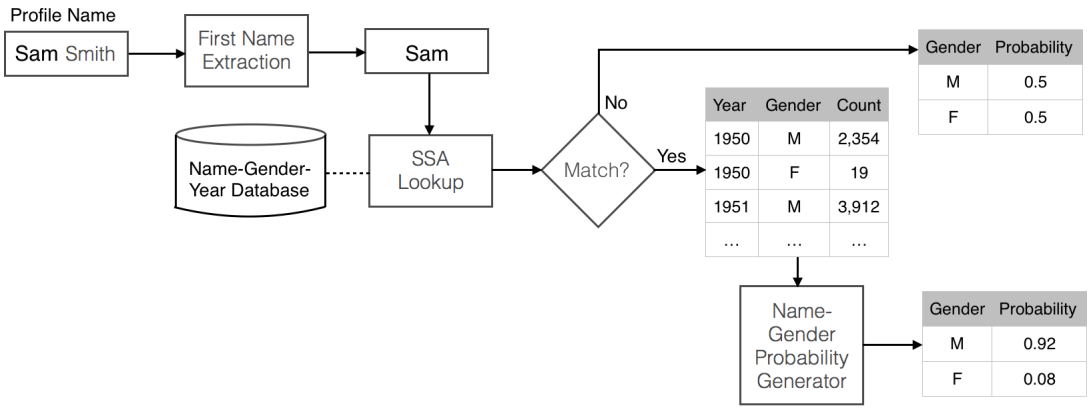

Figure 4.5 Processing a first name extracted from a user profile into a gender-probability score.

probability of a given name assigned to a male, $\boldsymbol{\mathcal{NM}}(u_i) = p(male)$. If a first name outside this list of popular first names is encountered, the method assumes an equal probability for the name to be given to either gender. An overview of these processes is shown in Figure 4.5.

In summary, features introduced in the above can be broadly classified into three categories, including topic-related features, $\boldsymbol{\mathcal{TV}}(u_i, \cdots)$, profile features $\boldsymbol{\mathcal{PV}}(u_i)$, and a gender distribution score, $\boldsymbol{\mathcal{NM}}(u_i)$, based on a user's self-declared first name. All these features collectively and comprehensively characterize a user's activities and personal preferences on Twitter, which are exploited by the proposed method for user gender inference.

4.2.5 User Classification

Utilizing the aforesaid sets of features, including $\boldsymbol{\mathcal{TV}}(u_i, \cdots)$, $\boldsymbol{\mathcal{PV}}(u_i)$, and $\boldsymbol{\mathcal{NM}}(u_i)$, we leverage a SVM-based classification method [101] for gender determination. We choose to employ the SVM-based classification technique because of the plentiful success of the technique in extracting Twitter users' demographic information as abundantly reported in the literature, e.g., [6, 143]. In addition, the decision

boundaries generated by the SVM model allow for the interpretation and analysis of the proposed model.

4.3 Experimental Results

In this section, we experimentally explore the capability of the proposed method in inferring a Twitter user's gender through comparing the performance of the proposed method with that of several state-of-the-art peer methods.

4.3.1 Peer Methods

For benchmarking purposes, two well-known peer methods for gender inference are considered in our experiment. The first method is a bag-of-words-based approach discussed in [143] where a user's aggregated timeline is represented through uni-grams. The relationship between a user's vocabulary usage and gender is then examined by the model for gender determination. We refer to this approach as $\mathcal{BOW}(u_i)$. The second peer method is proposed by [6], which examines 5-gram character expansions of a user's posts, profile description, and user-provided name for gender inference. For example, the name "John", through 5-gram character expansion, would be expanded to J, O, H, N, JO, OH, HN, , OHN, JOHN. We refer to this feature set as $\mathcal{BUR}(u_i)$. Both methods, which adopted SVM as a classification algorithm, are tested using the experimental dataset presented in this work under a 10-fold cross validation theme.

When we compare the accuracy of the best-performing model, $\mathcal{TV}_{\text{non-linear}}(u_i, 20, pro)$, we observed that the performance of the $\mathcal{BOW}(u_i)$ and the $\mathcal{BUR}(u_i)$ models was exceeded (respective accuracies, $\mathcal{TV} = 0.694$, $\mathcal{BOW} = 0.634$, $\mathcal{BUR} = 0.614$). Interestingly, the reduction in the dimensionality from the gram-based models ($\mathcal{BUR}$ has 998,095 and $\mathcal{BOW}$ has 125,951 dimensions) to just 20 dimensions introduced by this work improved accuracy. We expect this enhancement in accuracy comes from two primary factors. First, the reduction in dimensionality reduces the

overfitting problem by filtering out noisy features used in the gram-based approaches. Secondly, we are able to target our approach to learning the specific 20 topics previously shown capable of revealing user gender.

4.3.2 Feature Combinations

The feature set of $\boldsymbol{\mathcal{TV}}(u_i, n, c)$ characterizes a user through the person's topic interest evolving across the timeline; the feature set of $\boldsymbol{\mathcal{PV}}(u_i)$ characterizes a user's profile choices, as well as his/her interactions with peers; the feature of $\boldsymbol{\mathcal{NM}}(u_i)$ represents a user's personal information, which comes from the user's first name in this work. Each of these feature sets carries its own characterization of a given user. Experimentally, we test the accuracy produced by training the model using the above feature set separately and collectively.

4.3.3 Modifications Applied to $\boldsymbol{\mathcal{TV}}(u_i, n, c)$

Recall that $\boldsymbol{\mathcal{TV}}(u_i, n, c)$ represents the distribution of topics discussed by a given user via the UGC contained in their timeline. Also, recall that the aggregated values of $\mathcal{TV}(u_i, n, c)_j$ are accumulated based on the assignment of probabilities for topic j from the language model described in Section 4.2.2. However, users may outwardly choose to append such hashtags contained in the candidate hashtag set ($\mathcal{CH}$) to tweets as part of their UGC. In this modification to $\boldsymbol{\mathcal{TV}}(u_i, n, c)$, we test the inclusion, exclusion, and separation of user generated candidate hashtags.

First, the inclusion of user generated hashtags (HT) is computed by generating the candidate hashtag distributions using the trained language model. Since this information is user provided, we know with certainty that a given tweet's content has a relationship with the specified hashtag. Therefore, we use a probability score for such a tweet of 1.0 for the specified candidate hashtag. All other probabilities for

Table 4.2 An Overview of the Accuracy Scores Produced by the Various Testing Conditions and Proposed Approach, Hashtag Propagation and PMI

Hashtag Propagation (c = pro)				
	$\mathcal{TV}$	$\mathcal{TV}$+ $\mathcal{NM}$	$\mathcal{TV}$+ $\mathcal{PV}$	$\mathcal{TV}$+ $\mathcal{NM}$+ $\mathcal{PV}$
	Binary Weighting–$\mathcal{TV}_{\text{binary}}$			
$\mathcal{TV}_{\text{binary}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.652±0.019	0.852±0.013	0.685±0.013	0.863±0.016
$\mathcal{TV}_{\text{binary}}(u_i, 20, c)$	0.658±0.021	0.867±0.019	0.671±0.018	0.867±0.011
$\mathcal{TV}_{\text{binary}}(u_i, 20, c)\|\|\mathcal{TV}_R$	0.647±0.022	0.859±0.018	0.674±0.014	0.857±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.674±0.019	0.869±0.013	0.681±0.007	0.87±0.02
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, c)$	0.692±0.017	0.88±0.016	**0.703±0.011**	0.881±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, c)\|\|\mathcal{TV}_R$	0.677±0.021	0.859±0.018	0.674±0.014	0.861±0.015
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 19, c) \uplus \mathcal{TV}_R$	0.68±0.022	0.878±0.015	0.687±0.019	0.882±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 19, c)$	0.693±0.02	0.88±0.013	0.703±0.013	**0.883±0.015**
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 19, c)\|\|\mathcal{TV}_R$	0.683±0.015	0.861±0.016	0.675±0.014	0.861±0.01

Table 4.2 (Continued) An Overview of the Accuracy Scores Produced by the Various Testing Conditions and Proposed Approach, Hashtag Propagation and PMI

Hashtag Propagation (c = pro)				
	$\mathcal{TV}$	$\mathcal{TV}+$ $\mathcal{NM}$	$\mathcal{TV}+$ $\mathcal{PV}$	$\mathcal{TV}+$ $\mathcal{NM}+$ $\mathcal{PV}$
	Non-Linear Weighting–$\mathcal{TV}_{\text{non-linear}}$			
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.651±0.022	0.852±0.013	0.685±0.01	0.863±0.01
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c)$	0.659±0.02	0.867±0.019	0.671±0.014	0.865±0.014
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c) \| \mathcal{TV}_R$	0.645±0.025	0.854±0.016	0.673±0.022	0.861±0.014
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.677±0.017	0.871±0.013	0.684±0.02	0.863±0.01
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c)$	**0.694±0.012**	0.88±0.017	0.702±0.017	0.882±0.007
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c) \| \mathcal{TV}_R$	0.675±0.016	0.853±0.017	0.673±0.028	0.86±0.008
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c) + \mathcal{TV}_R$	0.681±0.021	0.878±0.016	0.69±0.016	0.873±0.015
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c)$	0.69±0.016	**0.881±0.014**	0.702±0.02	**0.883±0.014**
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c) \| \mathcal{TV}_R$	0.687±0.013	0.859±0.021	0.674±0.021	0.86±0.013

Table 4.2 (Continued) An Overview of the Accuracy Scores Produced by the Various Testing Conditions and Proposed Approach, Hashtag Propagation and PMI

Point-wise Mutual Information (c = PMI)				
	$\mathcal{TV}$	$\mathcal{TV}+$ $\mathcal{NM}$	$\mathcal{TV}+$ $\mathcal{PV}$	$\mathcal{TV}+$ $\mathcal{NM}+$ $\mathcal{PV}$
	Binary Weighting–$\mathcal{TV}_{\text{binary}}$			
$\mathcal{TV}_{\text{binary}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.599±0.017	0.846±0.014	0.682±0.018	0.861±0.015
$\mathcal{TV}_{\text{binary}}(u_i, 20, c)$	0.613±0.019	0.856±0.011	0.662±0.014	0.86±0.012
$\mathcal{TV}_{\text{binary}}(u_i, 20, c) \| \mathcal{TV}_R$	0.612±0.021	0.842±0.009	0.667±0.012	0.857±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.649±0.022	0.857±0.014	0.663±0.014	0.862±0.011
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, c)$	0.658±0.018	0.866±0.017	0.671±0.018	0.872±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, c) \| \mathcal{TV}_R$	0.658±0.018	0.866±0.017	0.671±0.018	0.872±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 19, c) \uplus \mathcal{TV}_R$	0.645±0.014	0.864±0.014	0.68±0.02	0.868±0.012
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 19, c)$	0.655±0.02	0.865±0.011	0.675±0.016	0.873±0.014
$\mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 19, c) \| \mathcal{TV}_R$	0.612±0.021	0.846±0.012	0.667±0.012	0.857±0.014

Table 4.2 (Continued) An Overview of the Accuracy Scores Produced by the Various Testing Conditions and Proposed Approach, Hashtag Propagation and PMI

Point-wise Mutual Information (c = PMI)				
	$\mathcal{TV}$	$\mathcal{TV}+$ $\mathcal{NM}$	$\mathcal{TV}+$ $\mathcal{PV}$	$\mathcal{TV}+$ $\mathcal{NM}+$ $\mathcal{PV}$
	Non-Linear Weighting–$\mathcal{TV}_{\text{non-linear}}$			
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.574±0.023	0.841±0.016	0.68±0.021	0.861±0.015
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c)$	0.583±0.017	0.844±0.009	0.666±0.019	0.856±0.016
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c)\|\|\mathcal{TV}_R$	0.581±0.024	0.844±0.017	0.671±0.02	0.858±0.01
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.644±0.023	0.858±0.016	0.657±0.014	0.864±0.011
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c)$	0.658±0.014	0.868±0.019	0.678±0.019	0.872±0.011
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c)\|\|\mathcal{TV}_R$	0.581±0.024	0.844±0.017	0.671±0.02	0.858±0.01
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c) + \mathcal{TV}_R$	0.644±0.014	0.856±0.013	0.658±0.026	0.864±0.008
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c)$	0.658±0.025	0.869±0.014	0.677±0.015	0.873±0.015
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c)\|\|\mathcal{TV}_R$	0.583±0.02	0.844±0.017	0.671±0.02	0.858±0.01

Table 4.2 (Continued) An Overview of the Accuracy Scores Produced by the Various Testing Conditions and Proposed Approach, Hashtag Propagation and PMI

Hashtag Propagation and Point-wise Mutual Information		
	$\mathcal{TV}_{\text{binary}}(c = pro)+$ $\mathcal{TV}_{\text{binary}}(c = PMI)+$ $\mathcal{NM}+$ $\mathcal{PV}$	$\mathcal{TV}_{\text{non-linear}}(c = pro)+$ $\mathcal{TV}_{\text{non-linear}}(c = PMI)+$ $\mathcal{NM}+$ $\mathcal{PV}$
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.863±0.012	0.864±0.011
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c)$	0.866±0.13	0.865±0.009
$\mathcal{TV}_{\text{non-linear}}(u_i, 20, c)\|\|\mathcal{TV}_R$	0.859±0.016	0.861±0.012
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c) \uplus \mathcal{TV}_R$	0.871±0.008	0.871±0.011
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c)$	0.882±0.008	0.882±0.013
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, c)\|\|\mathcal{TV}_R$	0.872±0.010	0.869±0.013
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c) + \mathcal{TV}_R$	0.878±0.010	0.878±0.014
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c)$	0.884±0.010	**0.886±0.012**
$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, c)\|\|\mathcal{TV}_R$	0.872±0.006	0.870±0.010

such a tweet are assigned to 0.0. These additional probability scores are aggregated into the final version of $\boldsymbol{\mathcal{TV}}(u_i, n, c)$.

Secondly, we consider the exclusion of user generated HTs by computing the $\boldsymbol{\mathcal{TV}}(u_i, 20, c)$ according to the language model only. In this condition, the user generated HTs are not represented in the model.

Finally, we consider the separation of user generated HTs into independent features from the language model generated HTs. For this approach, language model and user annotated distributions of hashtag clusters are aggregated independently. One feature is added to represent "other" hashtags not identified as members of the hashtag clusters. In Table 4.1, we outline each of these modifications and their respective attributes such as dimensionality and data types.

4.3.4 Parameter Optimization

Hashtag Propagation There are several parameters introduced in Section 4.2.3. In this section, we introduce the ranges of parameters considered and their impact on the gender classification accuracy scores.

The Hashtag Propagation method introduced in Section 4.2.3 proposes two parameters, namely α and θ. Recall that α is the attenuation coefficient, which varies the rate at which hashtags are propagated in M^{norm}. In this experiment, we vary the α parameter from 0.0, no propagation, to 1.0, strong propagation, in 0.05 increments.

The Hashtag Propagation approach also introduce a function, $\theta()$, which controls the noise introduced by the propagation function. Recall that the θ function drops values below a specified threshold. In this experiment, we vary $\theta()$ from 0.0, no noise control, to 1.0, strong noise control, in 0.05 increments.

In addition to these parameters, we test the β parameter from 0.0 to 1.0 in 0.25 increments (as seen in Equation (4.12)). The γ parameter is also tested using the

weighted $\boldsymbol{\mathcal{TV}}$ approach (as defined in Equation (4.14)) by varying γ from 1.0 to 5.0 in increments of 1.0.

Finally, we apply quantile limit ($\mathcal{Q} = \{1, 2, 3, 4\}$) to the retention of values in $\boldsymbol{\mathcal{TV}}(u_i, n, c)$ before hashtag aggregation is applied. Again, this consideration is intended to control the noise introduced by the language model. For example, if a probability score produced by the language model is below the first quantile value ($\mathcal{Q} = 1$ of the topic's distribution of probabilities, we exclude the probability. In addition to testing the first quantile, we also consider the second ($\mathcal{Q} = 2$), third ($\mathcal{Q} = 3$), and fourth quantile ($\mathcal{Q} = 4$; essentially representing the maximum values in the range of probability for a given topic). The filtered values, having removed elements which did not exceed the specified quantile limits, are subsequently used in the hashtag propagation method.

The best performing overall model was produced using all three proposed feature sets ($\boldsymbol{\mathcal{TV}}^{\text{norm}}_{\text{non-linear}}(u_i, 19, pro)+\boldsymbol{\mathcal{NM}}(u_i)+\boldsymbol{\mathcal{PV}}(u_i)$). In general, $\boldsymbol{\mathcal{TV}}^{\text{norm}}(u_i, n, c)$ improved the accuracy of the model as compared to the non-normalized implementations. In addition, removing the "other" category in $\boldsymbol{\mathcal{TV}}(u_i, 19, c)$ also increased the accuracy of the overall model. It is likely that the gathering of remaining probabilities into a single feature ("other") provides little information regarding the gender of a user and, therefore, should be removed from the classification model.

PMI The Point-wise Mutual Information (PMI) method introduced in Section 4.2.3 introduces similar considerations for parameters as in the Hashtag Propagation approach. In this work, we consider a parameter, β, which controls the minimum PMI value required for which a candidate hashtag to be given membership to a topic in $\mathcal{S}(i)$. In other words, a low β value requires a weaker association between hashtags for membership. Conversely, a higher value of β indicates a much stronger association requirement between a given candidate hashtag $\mathcal{CH}$ and $\mathcal{S}(i)$. Recall that

Table 4.3 Accuracies of Non-Topic Features and Peer Methods

	Non-Topic Features
$\mathcal{NM}$	0.831±0.019
$\mathcal{PV}$	0.619±0.017
	Peer Methods
$\mathcal{BUR}$	0.614±0.032
$\mathcal{BOW}$	0.634±0.033

this work adopts normalized PMI, which bounds PMI scores from -1.0 to 1.0. First, we consider values greater than β ($> \beta$) as a requirement for a candidate hashtag, $\mathcal{CH}$, for inclusion in a topic, $\mathcal{S}(i)$. This condition assumes a strong association is required between hashtag topic and candidate hashtag. This parameter, similar to the Hashtag Propagation testing conditions, is varied for 0.0 to 1.0 in 0.25 increments. Secondly, we consider the weighted approach to $\mathcal{TV}$ by varying the γ parameter from 1.0 to 5.0 in increments of 1.0.

Finally, we adopt a quantile limit to the retention of values in $\mathcal{TV}(u_i, n, c)$ similar to the quantile filtering method introduced in the Hashtag Propagation parameter optimization section above. For PMI, we again consider the 1st, 2nd, 3rd, and 4th quantile limits ($\mathcal{Q} = \{1, 2, 3, 4\}$) for each topic in $\mathcal{TV}$.

In order to optimize the accuracy of the gender inference model involving the various parameters introduced above, we iteratively test each parameter for its impact on the accuracy score. All combinations of parameters are exhaustively tested using the proposed model via a brute force method of optimization.

4.3.5 Results

All tests were conducted using 10-fold cross validation. To evaluate the performance of the proposed method in this study, we define accuracy as:

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

where T_p, T_n, F_p and F_n are respectively true positive, true negative, false positive, and false negative rates. Table 4.2 compares the accuracy attained by the proposed method and the peer methods. The leftmost column in the table lists the specific version of the topic distribution vectors used in each run of the experiment. We have chosen accuracy as our metric of choice after careful consideration. First, accuracy is the metric which most closely represents measurement of the problem we are attempting to solve in this work, i.e., classifying gender appropriately given a user's social media activity. Secondly, we have a well-balanced dataset, which limits the bias that an accuracy score might introduce had the number of men and women in the collection not been approximately equal.

One of the highest accuracy model where hashtag propagation and PMI methods are treated independently is attained by the aggregated feature set of $\boldsymbol{\mathcal{TV}}^{\text{norm}}(u_i, 19, pro) + \boldsymbol{\mathcal{NM}}(u_i) + \boldsymbol{\mathcal{PV}}(u_i)$. The proposed model produced the highest accuracy at 88.3% with $\alpha = 0.25$, $\theta = 0.3$, $\gamma = 1.0$ for the non-linear weighting scheme, and $\mathcal{Q} = 3$. A similar accuracy level of 88.3% was attained using the binary weighting approach with the model parameters configured as $\alpha = 0.3$, $\theta = 0.55$, $\beta = 0.25$, and $\mathcal{Q} = 3$. This indicates that the level of importance for the propagation parameter α is noticeably affected by the weighting scheme adopted.

However, the highest accuracy overall was produced when combining the hashtag propagation and PMI features with the non-linear approach. To produce an accuracy of 88.6%, the following parameters were used: $\boldsymbol{\mathcal{TV}}^{\text{norm}}_{\text{non-linear}}(u_i, 19, pro) +$

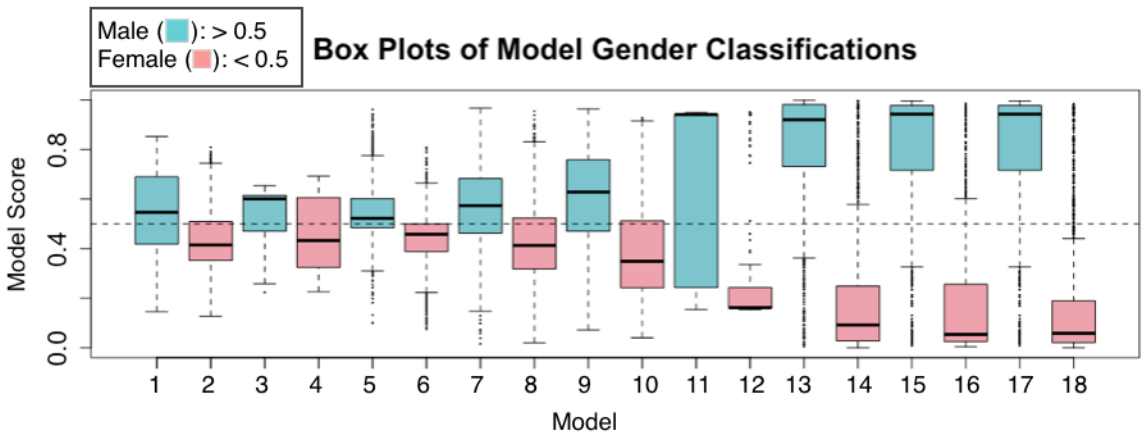

Figure 4.6 A box plot of the distribution of weighted user scores (according to the SVM model weights) for male and female-centric topics in TV. Model weights greater than 0.5 indicate a male classification, model weights less than 0.5 indicate a female classification. Models: 1&2) $\mathcal{BUR}$, 3&4) $\mathcal{PV}$, 5&6) $\mathcal{BOW}$, 7&8) $\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, pro)$, 9&10) $\mathcal{PV} + \mathcal{TV}^{\text{norm}}_{\text{binary}}(u_i, 20, pro)$, 11&12) $\mathcal{NM}$, 13&14) $\mathcal{NM} + \mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, pro)$, 15&16) $\mathcal{PV} + \mathcal{NM} + \mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, pro)$, 17&18) $\mathcal{PV} + \mathcal{NM} + \mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, pro) + \mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, PMI)$

$\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 19, PMI) + \mathcal{NM}(u_i) + \mathcal{PV}(u_i)$ with $\alpha = 0.25$, $\theta = 0.05$, $\gamma = 3.0$ for the non-linear weighting scheme, and $\mathcal{Q} = 3$.

It is also worth noting that using user topic-based features alone, the proposed method attained its highest accuracy of 69.4%,which outperformed both peer methods (respective accuracies, $\mathcal{BOW} = 0.634$, $\mathcal{BUR} = 0.614$). The model which relied solely on the derived topic features performed best under the following conditions: exclusion of user annotated HTs, normalization using $\mathcal{TV}^{\text{norm}}_{\text{non-linear}}(u_i, 20, pro)$, and the hashtag propagation aggregation non-linear weighting approach. The topic-based model produced the highest accuracy at when $\alpha = 0.7$, $\theta = 0.05$, $\gamma = 5.0$, and $\mathcal{Q} = 3$.

4.4 Discussion

In this section, we will highlight the important and interesting results from this work. Specifically, we will discuss the best performing model, including potential reasons for increased accuracy by the exclusion of user generated HTs, decision boundaries, and feature weights generated by the topic-based model. Finally, we will highlight

the performance of the proposed model in comparison to the peer methods, as well as introduce some applications of the proposed work.

First, the Hashtag Propagation aggregation method outperformed the PMI-based aggregation method under almost all conditions. This could be a result of the flexibility introduced by the Hashtag Propagation model. Recall that this aggregation approach allows for the bridge of hashtags relationships with intermediately shared hashtags. Such a method relaxes the condition of strictly requiring the co-occurrence of associated hashtags. Also, the non-linear weighting scheme generally produced higher accuracy scores when compared to the binary weighting scheme , contributing three of the five maximum accuracies for each feature set. However, when examining the overall average difference between scores, the binary model appeared to be more generalizable (on average producing an accuracy increase of 0.004). Most of this increase is a result of the PMI approach, which may be more amenable to the binary weighting scheme (averaging a 0.007 increase in accuracy), compared to the Propagation approach, which averaged a negligible increase of 0.0006.

Secondly, the highest overall accuracy of the proposed model and the most accurate version of the model that only leverages topic-based user features both neglect user annotated hashtags. This may seem a surprising result since the user annotated hashtags offer authentic human labels. We assume the reason why ignoring this type of authentic input is because of the aforementioned inconsistency and lack of comprehensiveness in user labeling practice. In contrast, hashtags automatically generated by our trained language models (see Section 4.2.2) are both standardized and systematic, providing a more informative source of reference than the original user manual labels. A secondary factor that contributes to the above performance implication is because of the frequent practice by some users to assign popular but unrelated tags to tweets simply to promote the visibility of their tweets [144, 145]. Such a distorted tagging practice for self-promotion on social media would lead to a

reduced accuracy in understanding a tweet's true content based on its associated tags. We believe the misrepresentation of gender identity online does not suffer from the same concern, as there is no obvious benefit to such a widespread misrepresentation across the community. Whereas the misrepresentation of content labels potentially provides the benefit of a larger audience to a user's social media content. If even our automatic hashtag generation model was trained using such a set of unreliable tags, due to the law of large numbers and the self-canceling effects among the imperfect data, the derived model is capable of generating more accurate annotation data than the original user assigned hashtags. Finally, because the original hashtags are retained but not utilized in the gender inference model, future comparisons could be made against the likelihood of misrepresentation of hashtags between genders.

When considering the decision boundaries and weights learned by the topic-based classifier, we see interesting results confirming relationships discovered in the language and psychology literature that guided this work, e.g., [146, 58, 124]. When observing the w vector produced by the SVM model generated using the **topic** feature set, we see that high weights (negative weights indicating female, positive weights indicating male users) as follows: *hobby* -0.595, *music* -0.416, *shopping* -0.328, *religion* -0.133, *alcohol* -0.262, *depression* -0.186, *violence* -0.15, and *loneliness* -0.122 generally indicate female users. On the other hand, topics most strongly associated with male users include: *school* 0.239, *video games* 0.066 (to a lesser extent), *television* 0.193, *sports* 0.180, *romance* 0.258, and *other* 0.240. These findings reaffirm results in prior literature [58] stating that women often exhibited a higher likelihood of sharing emotion-based content when compared to men. The analysis of the model introduced by this work also confirms similar trends according to the algorithmically-derived topics among the users of social media.

4.5 Limitations

We recognize the limitations of this work. First, users may falsely indicate their genders on social media account profiles, either due to sarcasm or intentional deceit, which is an outstanding issue with nearly all social media data sources. We also feel this issue is not of great magnitude, given the lack of benefit to a user for such a misrepresentation online. However, to mitigate this problem, we cross-check a user's gender information indicated outside Twitter, i.e., Facebook, if feasible.

This approach to gender inference on Twitter, while potentially applicable to other demographics, is limited by the availability of topics with a known connection to the demographic of interest. In other words, we have leveraged a list of known topics related to gender for the derivation of user topic distributions. Such topics may not exist for all demographic elements, which thereby limits the scope of demographics able to be inferred by this approach.

Additionally, time variant demographics, such as age, might require additional considerations before applying the approach proposed in this work. While there may exist topics related to age (e.g., retirement, high school, child rearing, etc.), the timescale at which these topics are related to the user's age is constantly changing. A user's high school-related tweet collected two years prior might have less impact on inferring a user's current age in the present. Static demographics, such as age or ethnicity, do not share the same concern, as the distribution over topics is time independent.

Finally, our approach to optimizing the parameters introduced by this approach is not ideal. The optimized parameters for the collection presented in this dissertation may not be the same optimized parameters for other datasets. The majority of the computational effort is devoted to selecting the correct parameters for gender inference. This work would benefit from an approach that automatically optimizes

the parameters, thus reducing the computational complexity and eliminating the brute force method adopted in this work.

4.6 Conclusion

In this work, we propose a new method for inferring a user's gender from all available information about the individual on Twitter. A secondary contribution of the work is the introduction of a new method for automatically generating standardized hashtags for tweets. Through comprehensive benchmarked experiments in comparison with peer methods, we demonstrated how this new approach can more reliably identify user gender information than the state-of-the-practice. Social media data, as exemplified by our Twitter collection, is sparse and messy. One of the benefits of this work is the derivation of multiple perspectives characterizing a single user, including features according to a user's online social media profile, language choices and topic distributions, and names. Such an aggregated approach of gender modeling and inference enable the proposed classifier to make a more informed and reliable decision concerning a user's gender.

The gender information automatically inferred by the proposed research can help researchers utilizing social media data to gain more demographic insight into the underlying user base, e.g., understanding trends pertinent to a specific gender as reflected in a social media dataset or obtaining any other useful health 2.0 information in Internet-based public health research. Such information will also enable gender-targeted message delivery and promotion, such as distributing gender-specific health messages. It is also noted that when the amount of training data is sufficient, we can apply the proposed method for inferring other demographic attributes of social media users, such as their age and ethnicity. Due to the scope of efforts in ground truth label acquisition and sample gathering, we will pursue this extension work in our immediate future research.

Social media data, as well as our Twitter collection, is sparse and messy. One of the important benefits of this work is the derivation of multiple perspectives of a single user. Viewing a user from a profile perspective, language/topic perspective, and personal perspective allows the aggregate feature set and, thus, the classifier to make a final classification using only those perspectives which are informative, thus reducing the impact of sparse and messy data.

Additionally, one benefit of the research presented in this paper is the ability to automatically label social media users according to their gender. Standard approaches taken by the Pew Research Center [147] include the manually intensive and prohibitively expensive surveying of individual social media users. By instituting an automatic approach to labeling users can reduce effort and provide more up-to-date statistics on social media demographics. Similarly, targeted messaging to specific genders is made easier by the work presented here. A specific example of targeted information distribution or message could be gender-specific health-related messages delivered to the appropriate users via a Twitter direct message or user-mention in a public post. For example, female breast cancer information could be targeted for delivery to female users by using this approach to detect the gender of a user. Finally, the purpose of this work is to aid in the rapid study of health trends on social media. Understanding the users' genders coupled with extracted health information can improve the area of internet-based public health research.

In this work we have introduce two contributions, a method for automatically proliferating hashtags to un-tagged Tweets and a new method to extract a user's gender from Twitter activity. We have demonstrated how this approach, while using fewer features than existing methods to represent users, can produce higher accuracy scores in terms of gender extraction.

CHAPTER 5

CONTRIBUTIONS AND FUTURE WORK

This work produces several contributions, specifically introducing new data mining methods with applications in social media mining as well as the use of these approaches for social media-based public health studies. As a result, health 2.0 research can be extended to larger groups of users by inferring user demographic information which is not explicitly provided by the majority of users. In this section, the important contributions of this work will be outlined and discussed.

5.1 Contributions

5.1.1 Inferring Ethnicity using Language on Social Media

This work introduced a new method for inferring a given user's ethnicity based on the user's language usage patterns. Many users on social media choose not to provide their ethnicity information publicly or are not provided the appropriate fields for disclosing this information by the social media platform. As a result, it is difficult for public health researchers to easily identify health trends among ethnic groups online. Two approaches for detecting language patterns are examined and compared with a baseline bag-of-words approach. In both approaches, only the user's Twitter timeline is considered by collecting ten months worth of posting activity. The first approach used synonym expansion to increase the number of terms when training the ethnicity classification model. This approach expands verbs, nouns, and adjectives with their synonyms using WordNet. As a result, a user who chooses to use the term "car" would also be represented with the feature "vehicle," "automobile," etc. The second approach generated latent topic distributions for each user using Latent Dirichlet Allocation topic modeling. Topic distributions were then used as features

for inferring a user's ethnicity. Comparing with bag-of-words (BOW) as a baseline, synonym expansion proved to be the highest accuracy approach.

5.1.2 An Analysis of Cancer-Related Discussions among Ethnic Groups

Having established an approach for inferring user ethnicity using synonym expansion with high accuracy, this work also examined the application of ethnicity inference in social media for health trend detection. Synonym expansion was used to infer the ethnicity of a large population of users over a ten month period from March 2014 to January of 2015. During the same period, the number of occurrences of the cancer, breast cancer, lung cancer, colorectal cancer, prostate cancer-related tweets were counted. This approach to analyzing health discussion trends in social media indicated that a statistically significant difference between African Americans and Caucasians was observed in almost every month throughout the ten-month study period. Additionally, this work observed a measurable difference between African Americans and Caucasians in the month following breast cancer awareness month (October). This finding results in an important implication, namely, that awareness campaigns are potentially ineffective toward the groups proportionately most impacted by the disease. This finding brings to light an opportunity for change in awareness campaign messaging or targeting techniques to better meet the awareness needs of the groups which are most impacted by diseases.

This work also provides a broader contribution in the form of new methods for identifying health trends across a large population. Using the language-based methods introduced earlier, demography studies can be conducted on larger scales with minimal effort, paving the way for understanding public health in new ways.

5.1.3 Automatically Assigning Meta-Hashtags to Untagged Tweets

Twitter users often choose to post untagged content lacking hashtags. Just as HTTP links are an important construct developing a rich network of connections between webpages, ultimately connecting information; hashtags create common links between social media content. The application of hashtags to content makes searching and retrieving relevant social media content easy, studying social media trends and identifying trending topics fast, and building profiles of interest for social media users more accurate. This work introduces and validates a new method for automatically assigning meta-hashtags to untagged tweets using a new supervised learning approach. The approach is able to learn a set of manually defined topics for the purposes of classifying unlabeled tweets. To do this, a hashtag cluster is generated for each manual topic by analyzing the co-occurrences of hashtags in multi-hashtag tweets across the collection. Training set tweets (i.e., tweets which have been labeled by Twitter users to contain one of the labels identified in each of the hashtag clusters) are parsed using a part-of-speech tagger. The parts-of-speech are used to identify noun-phrases within each of the tweets. Finally, this work has analyzed the differences in performance among various classification algorithms and successfully identified Deep Learning Neural Networks as the ideal algorithm for solving this classification problem. As a result of this work, a new method for assigning manually identified hashtags to an unlabeled set of tweets has been introduced.

5.1.4 Inferring Gender Demographics of Twitter Users

One of the main contributions of this work is a new method for inferring the gender of Twitter users using a combination of features intended to describe a user from multiple perspectives. This approach combines information from the user's profile, discussion topics, and personal information to generate an inferred gender. Features range in values derived from the name of the user, to the frequency with which they

post a given gender-related topic. Specifically, this work looks at deriving features from a user's first name, their color choices within their profile, and the frequency with which users post gender-related topics (e.g., video games, sports, shopping, etc.) automatically derived from the proposed approach. Feature sets are used to generate individual user demographic inferences by learning the patterns from labeled users. This approach to demographic extraction for gender has been shown to outperform other approaches in the literature.

5.1.5 General Applicability of Work

This work can be applied to various fields outside of public health research, including other fields of research, commercial, and political opportunities.

Other fields of research may benefit from the approaches introduced in this dissertation. The fields of demography and anthropology could benefit from this work by reducing the amount of effort required to run large-scale gender or ethnic studies across communities by inferring such information from social media. While this approach may not be as rigorous as a traditionally run demographic study, it could be used as a rough estimator or piloting approach to discover if an area of interest requires more detailed examination.

There are also obvious commercial applications of this work. Advertisers often target specific segments of the population for the delivery of their ads. Using this work, ads could be tailored to specific genders and automatically be delivered to the appropriate users. Additionally, products that are developed specifically for a gender could be advertised to that gender only, thus reducing the amount of wasted advertising spend.

Finally, this work may have applications in the segmentation of users for political purposes. Political action and awareness campaigns have become increasingly prominent on social media platforms. The ability to segment users according to their

political affiliation or specific demographics could help target users for information delivery according to specified political campaigns. Additionally, the wide availability of users' demographic information could help campaigns better understand their voting base or potential opportunities for growing it.

5.2 Future Work

There are two primary areas of work where potential extensions to these studies could be examined. The areas of opportunity for future work, beyond this dissertation, revolve around the classification of age and the difficulties it presents.

5.2.1 An Expanded Study to other Demographics

One opportunity for further exploration involves examining the classification of age brackets as a demographic component. Presently, this work considered gender and ethnicity. However, users are often much less willing to share age-related information online. Our data collection process, which uses the freely available Twitter API, involves randomly sampling user accounts with a capped limit on the number of server requests submitted. With the majority of Twitter users falling into the younger age brackets, the user profiles tended to be heavily skewed. With wider access to the Twitter API, the proposed modeling approaches could be adapted to extract age-related information similarly to the approaches presented in the gender inference section of this work.

5.2.2 An Analysis of Cancer-Related Discussions among Gender Groups

This work has established a method for accurately inferring a user's gender using a combination of user-derived features. However, we have yet to explore the usefulness of gender inference from a public health perspective. Therefore, in future studies we could leverage the gender inference approach and apply it to the analysis of the cancer-related tweets from a gender perspective.

5.3 Conclusion

This dissertation has introduced new approaches for detecting race/ethnicity and gender information from online user activity on Twitter. These approaches, building on existing work in the areas of machine learning and text mining, have increased the accuracy of gender inference and ethnicity detection among Twitter users. In addition, we introduced a new approach for automatically assigning hashtag labels to unlabeled tweets by combining a hashtag clustering approach with natural language processing techniques. We have also considered the application of the newly proposed ethnicity extraction technique for better understanding cancer-related discussion pattern disparities among African-American and Caucasian users. With this work, we hope to ease the burden of surveying large communities of users for understanding health patterns among various demographic groups by leveraging existing user-provided information on social media.

BIBLIOGRAPHY

[1] Man-Huei Chang, Ramal Moonesinghe, Heba M Athar, and Benedict I Truman. Trends in disparity by sex and race/ethnicity for the leading causes of death in the United States-1999-2010. *Journal of Public Health Management and Practice*, 2015.

[2] Robert N Anderson, Elizabeth Arias, et al. The effect of revised populations on mortality statistics for the United States, 2000. *National Vital Statistics Reports*, 51(9), 2003.

[3] Yuhua Ruan, Shu Liang, Junling Zhu, Xudong Li, Guangming Qin, Yujiang Jia, Qianping Liu, Benli Song, Qixing Wang, Hui Xing, et al. Gender and ethnic disparities of HIV and syphilis seroconversions in a 4-year cohort of injection drug users. *Southeast Asian Journal of Tropical Medicine and Public Health*, 44(5):842–53, 2013.

[4] Margaret E Beier and Phillip L Ackerman. Determinants of health knowledge: An investigation of age, gender, abilities, personality, and interests. *Journal of Personality and Social Psychology*, 84(2):439, 2003.

[5] Rachel A Freedman, Elena M Kouri, Dee W West, and Nancy L Keating. Racial/ethnic disparities in knowledge about one's breast cancer characteristics. *Cancer*, 121(5):724–732, 2015.

[6] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[7] Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, Department of Computer Science University of Massachusetts Amherst, USA, 2010.

[8] Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. Unfolding the event landscape on Twitter: Classification and exploration of user categories. In *Association for Computing Machinery 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 241–244, New York, NY, USA, 2012. ACM.

[9] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 15–24, New York, NY, USA, 2014. ACM.

[10] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How old do you think I am?" a study of language and age in Twitter. In *The International Conference on Weblogs and Social Media*, 2013.

[11] Dong Nguyen, Noah A Smith, and Carolyn P Rosé. Author age prediction from text using linear regression. In *5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics, 2011.

[12] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 149–156, Oct 2011.

[13] Toine Lagro-Janssen, Sylvie Lo Fo Wong, and Maria van den Muijsenbergh. The importance of gender in health problems. *European Journal of General Practice*, 14(s1):33–37, 2008.

[14] MJM Stoverincic, ALM Lagro-Janssert, and C Van Weel. Sex differences in health problems, diagnostic testing, and referral in primary care. *The Journal of Family Practice*, 43(6):567, 1996.

[15] Cecile MT Gijsbers van Wijk, Annemarie M Kolk, Wil JHM van Den Bosch, and Henk JM Van Den Hoogen. Male and female morbidity in general practice: The nature of sex differences. *Social science & Medicine*, 35(5):665–678, 1992.

[16] Elina Haavio-Mannila. Inequalities in health and gender. *Social Science and Medicine*, 22(2):141 – 149, 1986. Special Issue: Medical Sociology and the Who's Programme for Europe.

[17] Virginia M Miller. Why are sex and gender important to basic physiology and translational and individualized medicine? *American Journal of Physiology-Heart and Circulatory Physiology*, 306(6):H781–H788, 2014.

[18] Ronald M Epstein, Kevin Fiscella, Cara S Lesser, and Kurt C Stange. Why the nation needs a policy push on patient-centered health care. *Health Affairs*, 29(8):1489–1495, 2010.

[19] Janet B Henrich. Women's health education initiatives: Why have they stalled? *Academic Medicine*, 79(4):283–288, 2004.

[20] Barbara Zelek, Susan P Phillips, and Yvonne Lefebvre. Gender sensitivity in medical curricula. *Canadian Medical Association Journal*, 156(9):1297–1300, 1997.

[21] Emily M Fox, Todd W Miller, Justin M Balko, Maria G Kuba, Violeta Sánchez, R Adam Smith, Shuying Liu, Ana María González-Angulo, Gordon B Mills, Fei Ye, et al. A kinome-wide screen identifies the insulin/IGF-I receptor pathway as a mechanism of escape from hormone dependence in breast cancer. *Cancer Research*, 71(21):6773–6784, 2011.

[22] Betsy A Kohler, Recinda L Sherman, Nadia Howlader, Ahmedin Jemal, A Blythe Ryerson, Kevin A Henry, Francis P Boscoe, Kathleen A Cronin, Andrew Lake, Anne-Michelle Noone, et al. Annual report to the nation on the status of cancer, 1975-2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. *Journal of the National Cancer Institute*, 107(6):djv048, 2015.

[23] Salma Shariff-Marco, Juan Yang, Esther M John, Meera Sangaramoorthy, Andrew Hertz, Jocelyn Koo, David O Nelson, Clayton W Schupp, Sarah J Shema, Myles Cockburn, et al. Impact of neighborhood and individual socioeconomic status on survival after breast cancer varies by race/ethnicity: The neighborhood and breast cancer study. *Cancer Epidemiology Biomarkers & Prevention*, 2014.

[24] Robert Hines, Talar Markossian, Asal Johnson, Frank Dong, and Rana Bayakly. Geographic residency status and census tract socioeconomic status as determinants of colorectal cancer outcomes. *American Journal of Public Health*, 104(3):e63–e71, 2014.

[25] Mieke J Aarts, Carlijn BM Kamphuis, Marieke J Louwman, Jan Willem W Coebergh, Johan P Mackenbach, and Frank J Van Lenthe. Educational inequalities in cancer survival: A role for comorbidities and health behaviours? *Journal of Epidemiology and Community Health*, 2012.

[26] Ayal A Aizer, Tyler J Wilhite, Ming-Hui Chen, Powell L Graham, Toni K Choueiri, Karen E Hoffman, Neil E Martin, Quoc-Dien Trinh, Jim C Hu, and Paul L Nguyen. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. *Cancer*, 120(10):1532–1539, 2014.

[27] Carol A Parise and Vincent Caggiano. Disparities in race/ethnicity and socioeconomic status: Risk of mortality of breast cancer patients in the California Cancer Registry, 2000–2010. *BMC Cancer*, 13(1):1, 2013.

[28] Heather Orom, Marc T Kiviniemi, Willie Underwood, Levi Ross, and Vickie L Shavers. Perceived cancer risk: Why is it lower among nonwhites than whites? *Cancer Epidemiology Biomarkers & Prevention*, 19(3):746–754, 2010.

[29] Ezinne Grace Ndukwe, Karen Patricia Williams, and Vanessa Sheppard. Knowledge and perspectives of breast and cervical cancer screening among female african immigrants in the Washington DC metropolitan area. *Journal of Cancer Education*, 28(4):748–754, 2013.

[30] April Oh, Abdul Shaikh, Erika Waters, Audie Atienza, Richard P Moser, and Frank Perna. Health disparities in awareness of physical activity and cancer prevention: Findings from the National Cancer Institute's 2007 Health Information National Trends Survey (HINTS). *Journal of Health Communication*, 15(sup3):60–77, 2010.

[31] Cheryl G Healton, Kristen McCausland, M Lyndon Haviland, Donna Vallone, Ellen R Gritz, Kevin C Davis, and Ghada Homsi. Women's knowledge of the leading causes of cancer death. *Nicotine & Tobacco Research*, 9(7):761–768, 2007.

[32] Olúgbémiga T Ekúndayò and David B Tataw. Barriers to prostate cancer prevention and community recommended health education strategies in an urban african american community in Jackson, Mississippi. *Social Work in Public Health*, 28(5):520–538, 2013.

[33] Karen Kaiser, Kenzie A Cameron, Gina Curry, and Melinda Stolley. Black women's awareness of breast cancer disparity and perceptions of the causes of disparity. *Journal of Community Health*, 38(4):766–772, 2013.

[34] LJL Forbes, L Atkins, A Thurnham, J Layburn, F Haste, and AJ Ramirez. Breast cancer awareness and barriers to symptomatic presentation among women from different ethnic groups in East London. *British Journal of Cancer*, 105(10):1474–1479, 2011.

[35] Barbara D Powe, Dexter L Cooper, Lokie Harmond, Louie Ross, Flavia E Mercado, and Rachel Faulkenberry. Comparing knowledge of colorectal and prostate cancer among african american and hispanic men. *Cancer Nursing*, 32(5):412–417, 2009.

[36] Pauline M Green and Beatrice Adderley Kelly. Colorectal cancer knowledge, perceptions, and behaviors in african americans. *Cancer Nursing*, 27(3):206–215, 2004.

[37] Ann Scheck McAlearney, Katherine W Reeves, Stephanie L Dickinson, Kimberly M Kelly, Cathy Tatum, Mira L Katz, and Electra D Paskett. Racial differences in colorectal cancer screening practices and knowledge within a low-income population. *Cancer*, 112(2):391–398, 2008.

[38] Liliana Laranjo, Amaël Arguel, Ana L Neves, Aideen M Gallagher, Ruth Kaplan, Nathan Mortimer, Guilherme A Mendes, and Annie YS Lau. The influence of social networking sites on health behavior change: A systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 22(1):243–256, 2015.

[39] Melanie Hingle, Donella Yoon, Joseph Fowler, Stephen Kobourov, Michael Lee Schneider, Daniel Falk, and Randy Burd. Collection and visualization of dietary behavior and reasons for eating using Twitter. *Journal of Medical Internet Research*, 15(6):e125, 2013.

[40] Isabel De la Torre-Díez, Francisco Javier Díaz-Pernas, and Míriam Antón-Rodríguez. A content analysis of chronic diseases social groups on Facebook and Twitter. *Telemedicine and e-Health*, 18(6):404–408, 2012.

[41] Courtney R Lyles, Andrea López, Rena Pasick, and Urmimala Sarkar. "5 mins of uncomfyness is better than dealing with cancer 4 a lifetime": An exploratory qualitative analysis of cervical and breast cancer screening dialogue on Twitter. *Journal of Cancer Education*, 28(1):127–133, 2013.

[42] Yuya Sugawara, Hiroto Narimatsu, Atsushi Hozawa, Li Shao, Katsumi Otani, and Akira Fukao. Cancer patients on Twitter: A novel patient community on social media. *BMC Research Notes*, 5(1):699, 2012.

[43] Itai Himelboim and Jeong Yeob Han. Cancer talk on Twitter: Community structure and information sources in breast and prostate cancer social networks. *Journal of Health Communication*, 19(2):210–225, 2014.

[44] Rosemary Thackeray, Scott H Burton, Christophe Giraud-Carrier, Stephen Rollins, and Catherine R Draper. Using twitter for breast cancer prevention: An analysis of breast cancer awareness month. *BMC Cancer*, 13(1):508, 2013.

[45] Wen-Ying Sylvia Chou, Abby Prestin, and Stephen Kunath. Obesity in social media: A mixed methods analysis. *Translational Behavioral Medicine*, 4(3):314–323, 2014.

[46] Caroline A Bravo and Laurie Hoffman-Goetz. Tweeting about prostate and testicular cancers: Do twitter conversations and the 2013 movember Canada campaign objectives align? *Journal of Cancer Education*, pages 1–8, 2015.

[47] Atsushi Tsuya, Yuya Sugawara, Atsushi Tanaka, and Hiroto Narimatsu. Do cancer patients tweet? examining the Twitter use of cancer patients in Japan. *Journal of Medical Internet Research*, 16(5):e137, 2014.

[48] Vinay Prabhu, Ted Lee, Stacy Loeb, John H Holmes, Heather T Gold, Herbert Lepor, David F Penson, and Danil V Makarov. Twitter response to the United States preventive services task force recommendations against screening with prostate-specific antigen. *BJU International*, 116(1):65–71, 2015.

[49] Shoba Ramanadhan, Samuel R Mendez, Megan Rao, and Kasisomayajula Viswanath. Social media use by community-based organizations conducting health promotion: A content analysis. *BMC Public Health*, 13(1):1129, 2013.

[50] Deanna J Attai, Michael S Cowher, Mohammed Al-Hamadani, Jody M Schoger, Alicia C Staley, and Jeffrey Landercasper. Twitter social media is an effective tool for breast cancer patient education and support: Patient-reported outcomes by survey. *Journal of Medical Internet Research*, 17(7), 2015.

[51] Jennifer C Duke, Heather Hansen, Annice E Kim, Laurel Curry, and Jane Allen. The use of social media by state tobacco control programs to promote smoking cessation: A cross-sectional study. *Journal of Medical Internet Research*, 16(7):e169, 2014.

[52] Jenine K Harris. Local health department use of Twitter to disseminate diabetes information, United States. *Preventing Chronic Disease*, 10, 2013.

[53] Jenine K Harris, Sarah Moreland-Russell, Rachel G Tabak, Lindsay R Ruhr, and Ryan C Maier. Communication about childhood obesity on Twitter. *American Journal of Public Health*, 104(7):e62–e69, 2014.

[54] James Alexander, Harry T Kwon, Rachael Strecher, and Jill Bartholomew. Multicultural media outreach: Increasing cancer information coverage in minority communities. *Journal of Cancer Education*, 28(4):744–747, 2013.

[55] Mark Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012.

[56] Madhav Marathe and Naren Ramakrishnan. Recent advances in computational epidemiology. *IEEE Intelligent Systems*, 28(4):96, 2013.

[57] Valentina Beretta, Daniele Maccagnola, Timothy Cribbin, and Enza Messina. An interactive method for inferring demographic attributes in Twitter. In *26th ACM Conference on Hypertext & Social Media*, pages 113–122. ACM, 2015.

[58] Ann Colley and Zazie Todd. Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*, 21(4):380–392, 2002.

[59] John Russell Rickford. *African American Vernacular English: Features, evolution, educational implications*. Wiley-Blackwell, 1999.

[60] Tiffany A Pempek, Yevdokiya A Yermolayeva, and Sandra L Calvert. College students' social networking experiences on Facebook. *Journal of Applied Developmental Psychology*, 30(3):227–238, 2009.

[61] Dejin Zhao and Mary Beth Rosson. How and why people Twitter: The role that micro-blogging plays in informal communication at work, 2009.

[62] Oren Tsur and Ari Rappoport. What's in a hashtag?: Content-based prediction of the spread of ideas in microblogging communities. In *5th ACM International Conference on Web Search and Data Mining*, pages 643–652. ACM, 2012.

[63] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences (HICSS)*, pages 1–10. IEEE, 2010.

[64] Sara Rosenthal and Kathleen McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 763–772, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[65] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *3rd International Workshop on Search and Mining User-Generated Contents*, pages 37–44. ACM, 2011.

[66] Hasan Ali AL Akram and Amjad Mahmood. Predicting personality traits, gender and psychopath behavior of twitter users. *International Journal of Technology Diffusion*, 5(2):1–14, 2014.

[67] Bruce Thompson. *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. American Psychological Association, 2004.

[68] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, pages 457–500, 2007.

[69] Gaby Odekerken-Schröder, Kristof De Wulf, and Patrick Schumacher. Strengthening outcomes of retailer–consumer relationships: The dual impact of relationship marketing tactics and consumer personality. *Journal of Business Research*, 56(3):177–190, 2003.

[70] James W Pennebaker and Anna Graybeal. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93, 2001.

[71] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[72] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *ACL '02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

[73] Jalal S Alowibdi, Ugo Buy, Paul Yu, et al. Empirical evaluation of profile characteristics for gender classification on Twitter. In *2013 12th International Conference on Machine Learning and Applications (ICMLA)*, volume 1, pages 365–369. IEEE, 2013.

[74] Jalal S Alowibdi, Ugo A Buy, and Paul Yu. Language independent gender classification on Twitter. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 739–743. IEEE, 2013.

[75] Mohsen Sayyadiharikandeh, Giovanni Luca Ciampaglia, and Alessandro Flammini. Cross-domain gender detection in Twitter. *Computational Approaches to Social Monitoring*, 2016.

[76] Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. A comparative study of demographic attribute inference in Twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[77] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[78] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In *33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

[79] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in Twitter. In *1st Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[80] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[81] Xiaojun Ma, Yukihiro Tsuboshita, and Nei Kato. Gender estimation for SNS user profiling using automatic image annotation. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.

[82] Michele Merler, Liangliang Cao, and John R Smith. You are what you tweet... pic! gender prediction based on semantic analysis of social media images. In *2015 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2015.

[83] Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. Predicting the demographics of twitter users from website traffic data. In *Association for the Advancement of Artificial Intelligence*, pages 72–78, 2015.

[84] Junichi Ito, Takahide Hoshide, Hiroaki Toda, Tsuyoshi Uchiyama, and Keisuke Nishida. What is he she like?: Estimating Twitter user attributes from contents and social neighbors. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1448–1450. IEEE, 2013.

[85] Katja Filippova. User demographics and language in an implicit social network. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488. Association for Computational Linguistics, 2012.

[86] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our Twitter profiles, our selves: Predicting personality with Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 180–185, Oct 2011.

[87] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *1st Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[88] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to Twitter user classification. *International AAAI Conference on Web and Social Media*, 11:281–288, 2011.

[89] H. I. McCallum and R. M. Anderson. Systematic temporal changes in host susceptibility to infection: Demographic mechanisms. *Parasitology*, 89:195–208, 8 1984.

[90] Johanna M Seddon, Robyn Reynolds, Julian Maller, Jesen A Fagerness, Mark J Daly, and Bernard Rosner. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Investigative Ophthalmology & Visual Science*, 50(5):2044–2053, 2009.

[91] John McPartland. Demography and disease. *JAMA*, 281(20):1893–1893, 1999.

[92] Susan Scott and Christopher John Duncan. *Human Demography and Disease*. Cambridge University Press, 2005.

[93] Zhiheng Xu, Rong Lu, Liang Xiang, and Qing Yang. Discovering user interest on Twitter with a modified author-topic model. In *2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 422–429, Aug 2011.

[94] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In *20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1031–1040, New York, NY, USA, 2011. ACM.

[95] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 889–892, New York, NY, USA, 2013. ACM.

[96] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *International AAAI Conference on Web and Social Media*, 10:1–1, 2010.

[97] Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool, and Sungyoung Lee. Precise tweet classification and sentiment analysis. In *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*, pages 461–466. IEEE, 2013.

[98] Wang Meng, Lin Lanfen, Wang Jing, Yu Penghua, Liu Jiaolong, and Xie Fei. Improving short text classification using public search engines. In *Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 157–166. Springer, 2013.

[99] Stefano Mizzaro, Marco Pavan, Ivan Scagnetto, and Martino Valenti. Short text categorization exploiting contextual enrichment and external knowledge. In *1st International Workshop on Social Media Retrieval and Analysis*, pages 57–62. ACM, 2014.

[100] Jacques Savoy. Authorship attribution based on specific vocabulary. *ACM Transactions on Information Systems (TOIS)*, 30(2):12, 2012.

[101] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[102] Dell Zhang and Wee Sun Lee. Question classification using support vector machines. In *26th Annual International Conference on Research and Development in Informaion Retrieval*, pages 26–32. ACM, 2003.

[103] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49. ACM, 1999.

[104] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[105] Christiane Fellbaum. *Wordnet*. Wiley Online Library, 1998.

[106] George A Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[107] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential Twitterers. In *3rd Association for Computing Machinery International Conference on Web Search and Data Mining*, pages 261–270. ACM, 2010.

[108] Thorsten Joachims. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Springer, 1998.

[109] Kay H Brodersen, Cheng Soon Ong, Klaas E Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition (ICPR)*, pages 3121–3124. IEEE, 2010.

[110] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[111] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: An analysis of privacy leaks on Twitter. In *10th Annual ACM Workshop on Privacy in the Electronic Society*, pages 1–12. ACM, 2011.

[112] Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research*, 17(6), 2015.

[113] Erin E Kent, Abby Prestin, Anna Gaysynsky, Kasia Galica, Robin Rinker, Kaitlin Graff, and Wen-Ying Sylvia Chou. "obesity is the new major cause of cancer": Connections between obesity and cancer on Facebook and Twitter. *Journal of Cancer Education*, pages 1–7, 2015.

[114] National Institute of Health. Home Health Reference. http://ghr.nlm.nih.gov, January 2016.

[115] Bradford W Hesse, Galen E Cole, and Barbara D Powe. Partnering against cancer today: A blueprint for coordinating efforts through communication science. *Journal of the National Cancer Institute*, 2013(47):233–239, 2013.

[116] Kasisomayajula Viswanath and Karen M Emmons. Message effects and social determinants of health: Its application to cancer disparities. *Journal of Communication*, 56(s1):S238–S264, 2006.

[117] Maria Pérez, Julianne A Sefko, Deb Ksiazek, Balaji Golla, Chris Casey, Julie A Margenthaler, Graham Colditz, Matthew W Kreuter, and Donna B Jeffe. A novel intervention using interactive technology and personal narratives to reduce cancer disparities: African american breast cancer survivor stories. *Journal of Cancer Survivorship*, 8(1):21–30, 2014.

[118] Melany Cueva, Regina Kuhnley, Laura Revels, Nancy E Schoenberg, Anne Lanier, and Mark Dignan. Engaging elements of cancer-related digital stories in Alaska. *Journal of Cancer Education*, pages 1–6, 2015.

[119] Tracey L Thomas, Otis L Owens, Daniela B Friedman, Myriam E Torres, and James R Hébert. Written and spoken narratives about health and cancer decision making a novel application of photovoice. *Health Promotion Practice*, page 1524839912465749, 2012.

[120] Google. Google AdWords cities-DMA region, July 2016.

[121] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

[122] Max Kaufmann and Jugal Kalita. Syntactic normalization of twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India, 2010.

[123] Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432. Association for Computational Linguistics, 2012.

[124] Susannah R Stern. Expressions of identity online: Prominent features and gender differences in adolescents' world wide web home pages. *Journal of Broadcasting & Electronic Media*, 48(2):218–243, 2004.

[125] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE, 2011.

[126] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from Twitter. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 379–388. Association for Computational Linguistics, 2011.

[127] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[128] John S Justeson and Slava M Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(01):9–27, 1995.

[129] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[130] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[131] Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. Developing a successful semeval task in sentiment analysis of twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65, 2016.

[132] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.

[133] Malcolm Corney, Olivier De Vel, Alison Anderson, and George Mohay. Gender-preferential text mining of e-mail discourse. In *18th Annual Computer Security Applications Conference*, pages 282–289. IEEE, 2002.

[134] Harold Schiffman. Bibliography of gender and language, 2002.

[135] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *2010 conference on Empirical Methods in Natural Language Processing*, pages 207–217. Association for Computational Linguistics, 2010.

[136] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[137] Hugo Liu. Social network profiles as taste performances. *Journal of Computer-Mediated Communication*, 13(1):252–275, 2007.

[138] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.

[139] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 31–40, 2009.

[140] Maryam Dialameh and Mansoor Zolghadri Jahromi. A general feature-weighting function for classification problems. *Expert Systems with Applications*, 72:177–188, 2017.

[141] John R Smith and Shih-Fu Chang. Tools and techniques for color image retrieval. In *Electronic Imaging: Science & Technology*, pages 426–437. International Society for Optics and Photonics, 1996.

[142] Paul Spitalnic. Popular baby names - the United States Social Security Administration. http://www.socialsecurity.gov/OACT/babynames, May 2015.

[143] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *2nd International Workshop on Search and Mining User-Generated Contents*, pages 37–44. ACM, 2010.

[144] Liza Potts, Joyce Seitzinger, Dave Jones, and Angela Harrison. Tweeting disaster: hashtag constructions and collisions. In *29th ACM International Conference on Design of Communication*, pages 235–240. ACM, 2011.

[145] Marc Cheong. 'what are you tweeting about?': A survey of trending topics within twitter. *Clayton School of Information Technology, Monash University*, 2009.

[146] Anthony Mulac and Torborg Louisa Lundell. Linguistic contributors to the gender-linked language effect. *Journal of Language and Social Psychology*, 5(2):81–101, 1986.

[147] Social networking fact sheet. Technical report, Pew Research Center, 2015.