Detecting the Gender of a Tweet Sender

A Project Report
Submitted to the Department of Computer Science
In Partial Fulfilment of the Requirements
For the Degree of
Master of Science
In Computer Science
University of Regina

by

Thomas Oshiobughie Ugheoke
Regina, Saskatchewan

May 2014

ABSTRACT

Social media has been in existence for over two decades and, increasingly, people are using it to communicate, connect, share content, and socialize across the globe. Given the huge amount of data generated by the great popularity of social media sites, opportunities have emerged for researchers to study the demographic attributes of its social media users. Arguably, Twitter is one of the most popular of the social media sites. Acting as both a social media platform and as a micro-blogging site, Twitter has pioneered the use of the short-messaging system in social platforms. Often, this mode of conversation contains nonstandard language and, along with its requirement for brevity (i.e., 140 character limit), Twitter can be a challenging genre for natural-language processing. Twitter does not collect users' self-reported gender as do other social media sites (e.g., Facebook and Google+), but such information could be useful for targeting a specific audience for advertising, for personalizing content, and for legal investigation. These procedures, known as *authorship identification*, provide veritable information about the tweet author, for example, the gender of the author, their age, political affiliation, and occupation. And it is interesting to note that difference in writing patterns is known to exist between the male and female genders.

Utilizing these facts, this project employs a machine-learning approach to train a classifier to use manually-labeled data from Twitter to automatically detect the gender of a tweet sender. First, selection algorithms were used to evaluate the types of features that contain the distinguishing details of a particular gender and, second, the results of experiments performed using a number of features are presented.

ACKNOWLEDGEMENTS

The successful completion of this project is due solely to the concerted efforts, support, guidance, and assistance that I received from many.

First, I want to express profound gratitude to my supervisor, Dr. Robert Hilderman, for his support and guidance through all stages of this project. I am extremely grateful that I could work under his supervision and benefit from his advice and support.

Special thanks to Dr. Howard J. Hamilton who encouraged me throughout. It was his guidance that led me to do research in this area.

I would also like to thank Dr. Daryl Hepting and Dr. Lisa Fan for participating on my committee, the faculty and staff of the Department of Computer Science for their support, and the Faculty of Graduate Studies and Research for their financial support.

Finally, special thanks to my big brother and sponsor, Dr. Eghierhua A. Ugheoke, FACG for his encouragement and unflinching determination that I succeed. Also, thanks to other family members for their encouragement and support, as well as to all my friends.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

The internet is one of mankind's wonderful inventions that has, in no small measure, accelerated our ability to communicate immediately and globally. The World Wide Web (WWW3), which was built on the internet, has given individuals and groups the ability to communicate effectively. With the advent of Web 2.0 capabilities, new elements have been added to the existing platform. With the advent of these new dimensions, social networks, blogs, wikis, and so forth, came into existence. The features of Web 2.0 provided tools for large-scale communication among individuals, groups and organizations. With the extensive reach of the internet, virtually every part of the world is, in some way, connected through one or more of these platforms. The most popular of all these platforms is the social media network.

"Social media is defined as a group of internet-based applications built on the ideological and technological foundations of Web 2.0 that allow the creation and exchange of user-generated content" [1]. Social media has extremely affected the way we socialize and the way we make friends and it has pervaded our everyday lives. Prominent examples of social media include Facebook, MySpace, Instagram, LinkedIn, Google+, and Twitter. These media provide unhindered access for users to interact and share anything. Each site was built on different ideologies and each caters to a different niche of users. The increasing use of Twitter for communication and sharing creates large user-

generated content that represents users' thoughts, opinions, ideals, and emotions - a body of knowledge that cuts across every aspect of life. For example, as of March 2013 [2], Facebook had about 1.11 billion Monthly Active Users (MAUs); on the other hand, Twitter had over 500 million users [3]. In most cases, users voluntarily contribute to the large volume of user-generated content - content that cuts across text, photos, videos, and audio files.

The focus of this research project is on Twitter, one of the dominant social-media platforms. Twitter is a micro-blogging and social networking site that allows users to receive and send text-based messages that are usually 140-characters in length. With its large number of users, the potential of this service cannot be underestimated. Twitter has influenced the way in which we conduct business, our political orientation and, in fact, has been used in the organization and coordination of political protests around the world. For example, the 2011 Arab Spring that resulted in removal of the heads of government was largely coordinated through social networks, including Twitter [4].

Twitter's large repository of user-generated content contains data that can be mined. For example, previous research has shown that the linguistic choices associated with certain categories of people can be learned [15], [13] and that strong correlations between choice of language and such categories enable construction of "predictive models that are disarmingly accurate" [5]. But this leads to oversimplification that can present a misleading "picture of how language conveys personal identity" [5].

Many studies have shown that social media data, such as tweets, are rich sources of information about the real world. For example, studies have been conducted to understand the sentiment expressed in tweets [6] [33], in other work Tweets have been linked to polls to infer the effect of public opinion [7], and future events have been predicted by studying the behavior of users [8]; and future performance of the stock market has been predicted [9]

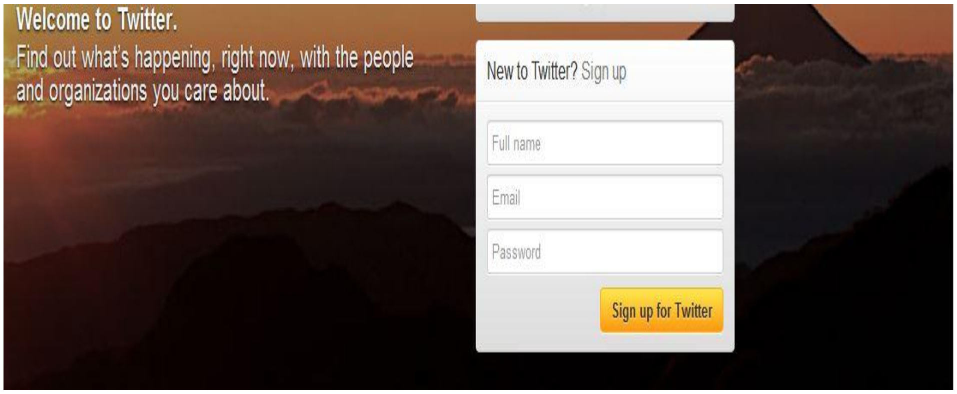
Several researches have been done to study the characteristics of Twitter users' demography, such as their political affiliation, their age, and their gender. And, there is much more that can be gleaned from these user-generated contents.

1.2 Statement of the Problem

Gender is an important life issue. Especially when dealing with life challenges, we tend to factor in specific solutions that best fit that gender. Whether discussion ranges from fashion, to health, to wealth, or to employment, gender distinction is an integral component. In each of these topics, discussion differs in respect to gender, in order to understand and to offer the best solution unique to each gender.

Social Media, such as Twitter, provides users with a means to communicate and to share with friends, families, organizations, and government. And, inherent in these voluntary communications are opinions, stances, styles, preferences, and ideas that may be unique to each gender. Understanding the user-generated content with respect to gender would be useful to individuals, companies, and even governments for personal recommendation, customization, targeted advertising, and policy formulation.

Typical profile information on most social networks and micro-blogging services contains detailed user metadata that includes name, location, gender, age, religious views, and so forth. Twitter, on the other hand, has a different approach to collecting the users' data, allowing some level of freedom in terms of metadata disclosure where personal information about age, gender, religious, and political orientation is conspicuously absent. Figure 1 shows Twitter's sign-up page where only a user's name and email address are requested. This page is in sharp contrast to Facebook's sign-up page that requires a user's birthday, gender, name and, occasionally, religious views. Twitter encourages the use of latent features.



Welcome to Twitter.

Find out what's happening, right now, with the people and organizations you care about.

Figure 1. Twitter sign up page.

The content generated on Twitter's user base and its coverage makes Twitter a rich source for research purposes. Yet Twitter's latent features do present challenges. These challenges could be addressed by using machine-learning approaches and by designing systems that can automatically and accurately predict Twitter's latent features to help build business intelligence and marketing, online customer reviews, purchase planning, public-opinion management, policy formation, sentiment search systems, and so forth.

1.3 Objectives of the Project

The objectives of this research project are to:

- identify the gender of the tweet sender

- demonstrate the importance of incorporating gender dimension in tweets that can help in research and innovation.
- analyze tweets with a view to assist in building automatic systems for business intelligence
- extract features in tweets and evaluate their effect on gender prediction
- understand the writing pattern peculiar to each gender on Twitter
- check the performance of SVM on corpus of tweets and compare results with baseline methods
- contribute to the body of knowledge in social-media analysis
- compare project results with well-known published work.

1. 4 Areas of Application

Some areas where the research findings could be applied include:

- online customer reviews
- personalization and customization
- public-opinion management
- online ads placement
- purchase planning
- detection of inflammatory text and cyber-bullying [10].

1.5 Organization of the Project Report

This report is organized as follows. In Chapter 2, some background information is provided, with a general overview of gender classification on Twitter. In Chapter 3, discussion includes work related to this field of research. In Chapter 4 the methodology is explained. Chapter 5 includes discussion of the experimental results of this project, which are compared with other well-known published works. The conclusions are presented in Chapter 6, with suggestions for future work.

CHAPTER TWO
BACKGROUND

Big data has increased exponentially in recent years, with social media contributing largely to volume of data generated. Hidden in these data is the potential for discovering useful knowledge that could improve our lives. Such knowledge contained in big data could make information transparent and usable, which could contribute to improvement in product development and policy formulation, thus contributing to growth in businesses [11]. A significant amount of these data is in text format and is largely generated by people. Specifically, social media creates platforms for social interaction and communication. Twitter, one of the dominant forms of social media, allows us to share and communicate, with little interference in the users' privacy, but neglects to collect the metadata that are so helpful when mining big data.

In Section 2.1 discussion provides an overview of gender and gender differences in language use; discussion in Section 2.2 dwells briefly on Twitter as a social media tool; Section 2.3 provides a general framework for gender classification on social media; Section 2.5 presents terminology that pertains to Twitter; and, lastly, Section 2.4 presents factors that affect tweet classification.

2.1 Overview of Gender and Differences in Use of Language

Is there variation in our written expression and in our use of language? To answer this question and to show that this not only true, but has also been proven [12], [13], [14], we look to research that has been conducted on this topic. It has been argued that variation exists in speech and that there is a pragmatic distinction between the genders in both language and speech 15], [16], and that the electronic media message is no different [17]. The male and female lexical use of language tends to be obvious in their choice and use of parts of speech (POS) Females tend to use more personal pronouns (e.g. he, me, I, him) and some specific noun modifiers, while males user use more noun specifiers (e.g. the, that, my, a) [12].

Some have argued that the difference in gender use of language stems from biological traits in male and female brain composition, which makes the female more verbally clever than the male [18], [19], [20]. Also, women tend to stick to the standard use of language, while men seem to consider reputation by use of nonstandard language that to a degree expresses toughness and asserts authority [5], [21]. Savicki et al [22] poses that there is a

> “tendency for women to be more polite, supportive, emotionally expressive, and less verbose than men in online public forums. Conversely, men are more likely to insult, challenge, and express sarcasm, use profanity, and send long messages. Discussion groups dominated by males have also been observed to use more impersonal, fact-oriented language”.

The group to which an author belongs influences his or her use of language, because the speaker adapts his/her use of language to the audience being addressed [23]. “Again, it follows that speakers can choose to show gender and age identity more or less explicitly in language use, depending on people’s perception of these variables, on their culture, the recipient of their utterance” [24]. Also, we can study the social identity of an author at different points in an interaction [24], [25] from a socio-linguistic aspect of language. These differences in speech between genders tend to concentrate more on interaction between the linguistic speaker and his or her linguistic context [12]. These distinctions have been widely and thoroughly studied in the literature dealing with the subject of gender and languages [26].

2.2 Twitter as a Social Media Tool

Twitter is a social media and micro-blogging site that was founded in 2006, which enables users to post short messages that are called *tweets* [27]. Usually micro-blogging is a short post delivered to network of associates and may also be referred to as micro-sharing or micro-updating [28]. Twitter users, known as *Tweeple* are allowed to post a message restricted to 140 characters in length. These posts can be about real-time happenings, or peer interactions, or expressions of sentiment and complain, and so forth. This is an easy way to communicate. When compared to other blogging sites, the enforced shorter post, or micro-blog, gives users a faster means of sharing and communicating and lessens the user's time requirement, as well as the thought needed to generate the content [29]. It is also worth noting that a user can post up to a thousand tweets per day, which a typical blogger may be unable to achieve. Paul [30] says that

> "Twitter is about discovering interesting people around the world. It can also be about building a following of people who are interested in you and your work/hobbies, and then providing those followers with some kind of knowledge value every day".

The status messages posted by a Twitter user on the social networking site and micro-blogging service appear on their timeline, as well as on those of their friends. Twitter users are usually identified by a user or screen name (i.e., it usually begins with '@' character), of the format, @username; where use of the user's real name is optional. Twitter allows all forms of characters to be part of the user profile name, with characters ranging from alpha-numeric to special characters. So it is not uncommon to see names such as "@*(*)ut%&#+3". A typical Twitter user can "follow" another user.

Consequently, the user who is followed can now receive the other user's status message (i.e., tweets) on his/her page. The user who is being followed has the option of following back. Twitter provides a direct-message feature that works as an email, enabling two users who follow each other to send private messages; however, if they are not following one another this is not possible.

Tweets can be grouped into topics by using hashtags (i.e., #), which are popular words, beginning with the "#" character. Hashtags allow users to efficiently categorize status messages into topics, which makes searching for a topic - based on tweets - possible. All users can *retweet* the status message of those they are following to their own followers.

As of June 2012, Twitter had over 500 million users [31]. Over 40 M users were from the US and over 200M were active monthly users [32]. This vast number of users, over 400 M tweets per day [TPD] [3], has opened up avenues for various areas of research [6], [33]. Both governments and organizations [34] have examined the research for knowledge that could be applied to decision making. In fact, to gauge the mood of customers towards their products and services, companies have become users themselves, interacting with both potential and existing customers.

Usually, tweets are 140 characters in length. This restriction on the number of characters per post results in using clever language and that the post is scan-friendly also assists users to track a number of interesting posts at a glance. In a world overwhelmed with content from various sources, this is ideal [30]. In most cases, posts of 140 characters in length result in a summary of an expression, which can make for an ambiguous message.

Facebook provides open and accessible metadata for the purpose of research. A quick glance at a Facebook user’s profile, immediately identifies the user’s birthday, their gender, name, and, occasionally, the option of their religious views. Twitter, on the other hand, provides scant knowledge of the user. Twitter exemplifies the latent use of users’ metadata. For proper use of the features inherent in these messages, we need enough user metadata that will assist in studying their patterns of writing. An accurate prediction of Twitter’s latent features in the content generated would be valuable in the field of business intelligence and marketing, online customer reviews, purchase planning, public opinion management, sentiment searches, and more. See Figure 2 for an example of a tweet.

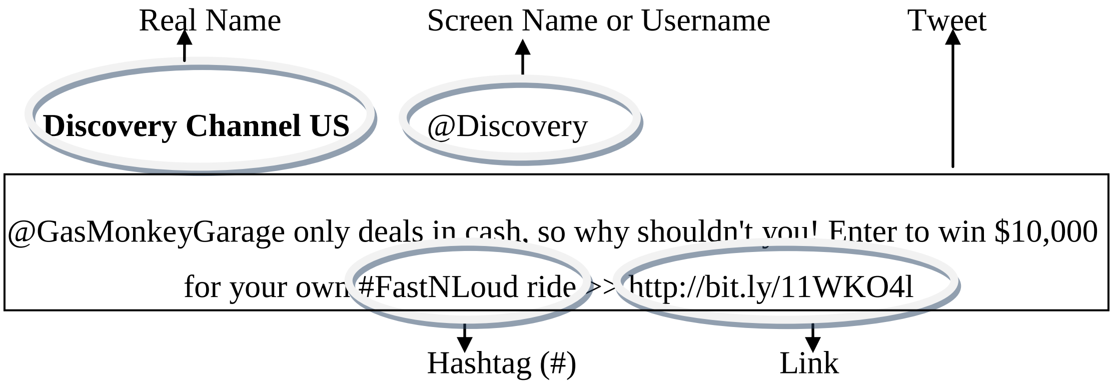

Real Name
Screen Name or Username
Tweet
Discovery Channel US
@Discovery
@GasMonkeyGarage only deals in cash, so why shouldn't you! Enter to win $10,000
for your own #FastNLoud ride >> http://bit.ly/11WKO4l
Hashtag (#)
Link

Figure 2. A tweet.

2.3 Twitter Terms

Twitter is a world on its own and, because of that, it has evolved specific terms that are peculiar to its micro-blogging site. Due to limiting the characters in a post, users have generated their own specific words to post. In on-line communities there is a high level of informality and the use of acronyms and slang is pervasive. Indeed, Twitter has evolved its own specific terms. Although these terms are not standard English, they do help users to quickly post updates and convey unique feelings. Some of these words have been globally adopted by users, and new words keep springing up. *Twictinary.pbworks.com and Twitonary.com* both have comprehensive listings of Twitter terms and their meanings. Some of the most popular terms are defined as follows.

Tweet. This is a short post/message/status/update/ from a user on Twitter's online message service, which is limited to 140 characters in length. This post could be about the user's activities, their communication with others, the sharing information, or retweeting other users' messages. The limit in the number of characters they can include influences a user's expression.

Tweeple. These are Twitter users, Twitter people, Twitter members.

Tweeps. These are users who follow each other from one social medial/network to another.

Twaffic. This means Twitter traffic.

Tweebay. This is used to describe selling (or promoting) an eBay item on Twitter.

Short URL. A user update or status may contain links to other blogs web-pages. Because a typical URL is, in most cases lengthy, Twitter condenses the URL to a short form, for example, http://bit.ly/1bifF5L.

Reply. The reply is a platform that Twitter provides to help users respond to an update (i.e., a tweet) that is directed to another user in response to their update. "An @reply is usually saved in the user's "Replies" tab. Replies are sent either by clicking the 'reply' icon next to an update or typing @username message (e.g., @user I saw that movie too)".[60]

***Hashtag*. "**The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet" [61] (e.g., #BarrackObama)."It was created organically by Twitter users as a way to categorize messages using minimal characters” [61]. Hashtag is a powerful function that helps users to categorize relevant keywords in a tweet. When the hashtag is clicked, all tweets containing that hashtag shows up. Certain hashtags sometime become very popular, which could turn them into trending topics. Hashtags are very useful for topic coordination and can provide insight into linguistic patterns in a particular domain. Hence tweets with a domain-specific hashtag can prove to be resourceful for keyword discovery.

Retweet. This is usually used to repost a tweet that was posted by another user to their followers. Sometimes the tweet may have been edited by a user who retweeted it; this is like forwarding an e-mail. When a user retweets a status, credit goes to the author of the tweet in this form RT @USER_NAME, for example, “@Nedunaija31m Really? RT @zynnnie: Joke of the century RT Nedu: Ngige is contesting again. He will win the election. His popularity in Anambra remains”

Mention. Mention acknowledges a user with the symbolic @ sign, but without using the Reply platform feature. For example, Wolf Blitzer tweeted ”@wolfblitzer I'm now in @CNNSitRoom which is about to begin”. Here CNN was mentioned.

Direct Message. This is a function of Twitter that enables users to send a private message to the people they follow and also to the people who follow them.

The above terms and acronyms were sourced from [35] and the online resources were sourced from [36] and [37].

2.4 General Features for Detecting Gender on Twitter

Machine-learning tools have proven to be very useful in mining the data by helping to discover hidden knowledge, critical to our decision-making process as our lives have become more and more complex, with overwhelming, readily available data, machine-learning tools work to simplify these complexities. While Twitter has championed online, short-messaging services and brought together different interest groups and social interactions, there is a need to better understand these users. But, in trying to understand the gender composition of Twitter users, we need to look further to identify indicators that best assist in determining a user's gender. Thus, discussion includes such as the user profile, user tweeting behavior, the linguistic style content of the user message, and the user's social network [5], [38] .

2.4.1 User Profile

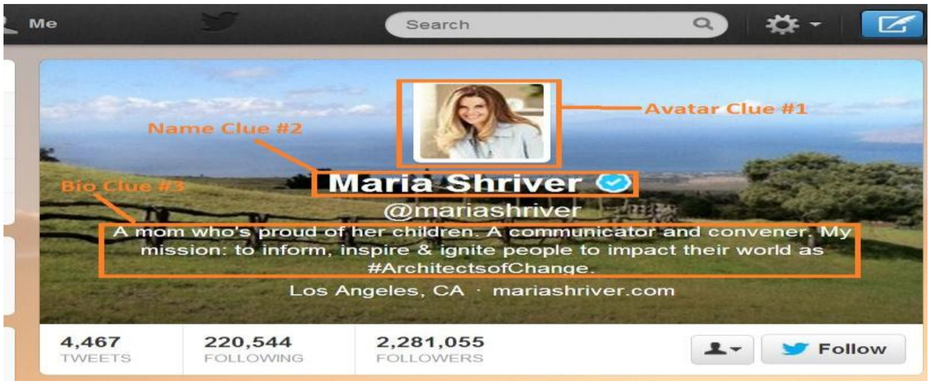
Twitter collects less user information upon sign up when compared to other social-media sites that make such user information publicly available. Typical user-profile information includes user name, location of the user, a short bio, a profile avatar, and a website. When properly mined, the information of these users can reveal much about them. Profile avatar can not only reveal your ethnicity and gender, but also more could be revealed if the user posted a picture that clearly associated him or her to political group or sport. Marco & Ana-Maria [38] found that the ethnicity of about 50% of 15,000 random users was correctly identified, while 57% correlated with a specific gender. However, because users may not use their correct avatar, this could be misleading. Marco & Ana-Maria [38] also found that 20% of the users did not post their picture at all, but sometimes the photo of a celebrity or another person.

User name has the potential for detecting both the ethnicity and gender of a user. Government departments, such as the US Social Security Administration, provide public access to their database that stores the baby names of both males and females, and this includes the most popular names for both genders. Thus, in cases where a user gives a real name, it can be appropriately classified.

User bio provides another good cue for demographical prediction. Twitter provides this for users to describe themselves and this short description can reveal a lot about their demographic makeup. For example, such gender-revealing user clips as “a mom who is proud of her children,” “I’m his biological dad” (see Figures 3 and 4). A typical, well-defined bio can be useful. For some users, their URL is another cue that can link to another aspect of online activities. A number of Twitter users link their profile to blogs and other social media such as Facebook. These media blogs and social network sites usually have well-defined bio where more can be learned about the author. Burger et al [35] automatically followed about 184,000 users and were able to determine their gender. A user’s bio can provide further verification and accuracy.

2.4.2 User Tweeting Behavior

There is research to support the claim that a user`s interactions on the micro-blogging service can be measured. Such measurable interactions include determining the statistics of the user tweeting rate; the average number of messages per day; the number of retweets per day; the number of replies, et cetera. And, there is research that both support and refute this.



Maria Shriver

@mariashriver

A mom who's proud of her children. A communicator and convener. My mission: to inform, inspire & ignite people to impact their world as #ArchitectsofChange.

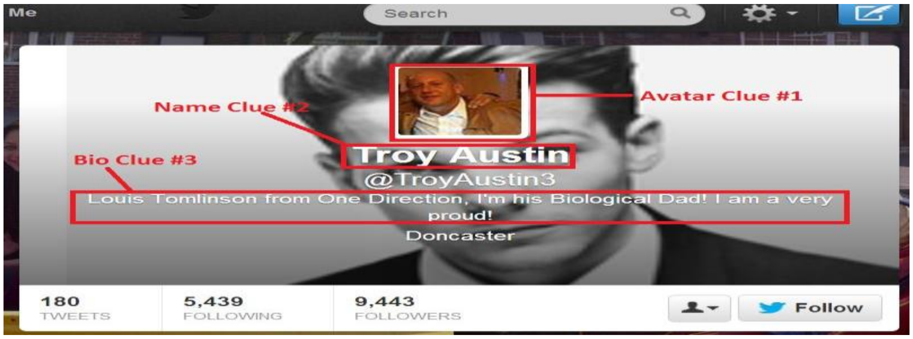
Los Angeles, CA · mariashriver.com

4,467 TWEETS

220,544 FOLLOWING

2,281,055 FOLLOWERS

Figure 3. Female Twitter profile.



Troy Austin

@TroyAustin3

Louis Tomlinson from One Direction, I'm his Biological Dad! I am a very proud!

Doncaster

180 TWEETS

5,439 FOLLOWING

9,443 FOLLOWERS

Figure 4. Male Twitter profile.

Akshay et al [40] argue that some users are information seekers on the micro-blogging site who rarely post status updates (i.e., tweets), while users who frequently post URL on their status updates are information providers. Another argument against measuring interactions on the micro-blogging service is [40], who suggest that linguistic features provide more cues for user classification, than does a user`s tweeting behavior. Marco & Ana-Maria [38] probed the claims of [39] and [40] by further experimentation that included more tweeting behaviors. Such behaviors included the number of tweets posted per user; the average number of hashtags and URLs per tweet; the number and fraction of tweets that are retweets and replied, and so forth. In measuring labels between males and females, they claimed an accuracy rate of 88%.

2.4.3 Linguistic Style

Users' lexical usage and linguistic style can provide helpful hints when classifying users by examining various media such as the forms from formal text, blogs, spoken conversations, search sessions, and, more recently, social media [38]. Computational linguists have used a variety of linguistics features to study the content of tweets. Style is still prevalent in our expressions; about 55% of the words we use are style words, even though only 0.05% of the English vocabulary is composed of style words which also includes articles and prepositions [41]. One may argue whether this applies to Twitter but, again, studies have proven this to be true [5], [42]. Danescu-Niculescu-Mizil et al [43] argued that "Twitter exchanges reflect the psycholinguistic concept of communication accommodation, where participants in conversations tend to converge to one another's communicative behavior; they coordinate using a variety of dimensions including choice of words, syntax, utterance length, pitch and gestures". Linguistic style is unconsciously expressed and is advocated to be a biological trait that is identical to each gender [18], [19], [20], [45].

2.4.4 Social Network

The saying that "birds of a father flock together" somewhat applies to social networking, in general, where one of users ultimate and most likely goal is to make friends. In the case of Twitter users, this entails being followed and following. Intuitively, sport fans are more likely to follow, tweet, and retweet notable footballers and sports-related topics. This is also applicable in the political world where a member of

Conservative Party of Canada will most likely follow Conservative leaders than to follow leaders of the New Democratic Party [5], [38].

2.5 Factors Affecting Gender Detection on Twitter

Tweets are predominantly text and in the text mining field are relegated to the category of Natural Language Processing (NLP). Social media is a recent form of communication that continues to evolve in new ways of interacting and sharing, and this includes new vocabularies. Twitter pioneered the famous 140-character short message online communication and enforced restriction on the number of characters per message. These characteristics have positively influenced the platform, as indicated by its continuous usage and its increased user base. Nevertheless, restrictions on the characters per post poses a different challenge because users have devised new ways to summarize their expressions which, in most cases, come at the cost of standard use of language. The following accentuates some of the present challenges.

2.5.1 Incomplete and short

Tweets are inherently short, which is Twitter ideology, so that users must attempt to adapt to this new form of expression. Sometimes these attempts result in an incomplete post and sometimes the brevity of the tweets may not capture the essence of the meaning because relevant information has been omitted. And, this omission could be a potential loss of information that is relevant to aspects of classification.

2.5.2 Spam

Spam is a general challenge in online communities and social media is no exception. Twitter, also, faces this increasing problem. Social media drives traffic to traditional websites makes it easy for spammers to target users. A study shows that more than 3% of messages on Twitter are spamming [46]. In most cases, spammers post malicious links to lure users to their site.

2.5.3 Deviation from Traditional Sociolinguistic Cues

Twitter is a community on its own that exerts influence on users' communication patterns. Here, the traditional socio-linguistics cues that hold in other genre are not so clear. Twitter failed to reveal certain behavioral characteristics that could distinguish between genders [40].

2.5.4 New Vocabularies

In a number of aspects, the limits that Twitter imposes have a positive influence, but these limitations have seriously affected the standard use of language. For instance, standard English is not necessarily use of formal expression, but on Twitter the users want to communicate and share and thus numerous vocabularies, alien to the standard dictionary, have developed. Even within the platform, Twitter, each niche may create their own words to assist them to share and communicate within Twitter unique 140-character limit. Even if one wants to accept these new words, the problem of general acceptability arises. It is very difficult to standardize these words. Most text-classification methods perform better with the standard language, such as that shown in the following tweets.

> ***Mark Dunk*** *@unklar1h: The* ***twaffic*** *is* ***atwocious****. (@ 288 S)*
> *http://4sq.com/1bsuUJz*
>
> ***Danielle Smith*** *@DanielleISmith25 Jul:* ***Twettiquette****,* ***Tweeple*** *or* ***Tweeps****,*
> ***Twaffic****, and* ***Tweetup****—the language of #Twitter. #SRSC*
>
> ***Amy Brown*** *@RhodyMam****a:Yo Tweeps: Twaffic Exchange!: To play along and increase your Twitter twaffic: we need to meet some new people.b...***
> *http://bit.ly/awIaqB*

Of course terms such as *Twettiquette, Tweeple,* or *Tweeps* and other such words are for exclusive use on Twitter**.**

2.5.5 Lack of Prosodic Cues

Unlike emails, blogs, and conversational speech which are characterized by prosodic cues helpful when learning attributes of author, Twitter is characterized by short and informal text [40].

2.5.6 Abbreviation/Acronyms

Abbreviations and acronyms are other areas of concern. Sometimes Twitter users may compose an entire tweet devoid of normal words, but of different words and acronyms (e.g., LOL, PRT, TMB).

> ***Obey ShadeY*** *@uhShadeY :Ummmmmmm who do you like? Hehehehhechuckle chuckle* ***lmfao omg omg lol lol lmao lmfao omg omg omg hehehe*** *cu... — wat http://ask.fm/a/52h4g7lm*

2.5.7 Informal Texts

Sometimes users deliberately misspell standard words in their tweets and, in most cases, every user misspells words uniquely and that greatly affects classification efforts. For instance, see the following tweets as examples of deliberate misspelling.

> *@autumndiabo:* ***Yesss****&omg you and cris are totally twinning it, did you guys come from... —* ***Yeeaah*** *omg I'm 4 minutes older than him http://ask.fm/a/3ck260o5*

> *@itsYSG9h:S☀P☀E☀C☀I☀A☀L☀F☀R☀I☀D☀A☀Y☀S☀H☀O☀U☀T ☀O☀U☀T☀T☀O*--------------------*>>>>>>>>>>>>>>> @bigmoNaija*

CHAPTER THREE
RELATED WORK

In general, social media has attracted a barrage of research since it first evolved. Twitter`s famous 140-character message ideology has bemused everyone from ordinary users - who sees the platform as a means to communicate and share with friends and others - to the researcher who views this platform as more than just a tool for sharing. Due to Twitter`s wide acceptance that cuts across all spectrums and walks of life – from politics, to social life, to the fields of entertainment, as well as industry - the researcher must accept the responsibility to learn the new platforms in order to access the vast store of user-generated data that is being turned out every second.

Continuous advancement in text-classification methods and machine-learning techniques has paved the way for research in this domain of social media. Machine-learning algorithms, such as Support Vector Machines (SVM), Naïve Bayes, Modified Balance Winnow, Gradient Boosted Decision Trees (GBDTs) have been used to classify gender with varying accuracy. The following examines some of the methods and techniques that are most relevant to this project.

3.1 Recent Work

Classifying the latent user attributes in Twitter was introduced by [40], by manually annotating 500 English users whose gender was labeled. Their work is the first-of-its kind application of latent-author-attribute discovery on Twitter data. Their work involved classifying the latent user attributes to determine their gender, political orientation, age, and regional origin. Their goal was to (a) study the effectiveness of language-content processing to identify latent user attributes from status messages and to (b) automatically discover some of the latent features through communication behavior, social-network structure, and status updates (i.e., tweets). They investigated stacked-SVM-based classification algorithms to study these attributes. Their study also included a detailed analysis to learn which features and methods worked well and those that did not. Their work was a supervised learning problem that was applied to tweet content in two ways: (a) sociolinguistic influenced features and (b) lexical-feature-based approaches. Three classification models were utilized: (a) the Ngram-feature, (b) the sociolinguistic-feature, and (c) the stacked model, which combined the results of the two previous models. They found that Twitter provided different socio-linguistic cues when compared to other online media. For example, the presence of consecutive, combined exclamation marks (!) and question marks (?) was indicative of a female user. Also they found that certain kinds of emoticons, as <3, are strong indicators of female authors. Men tend to laugh by using *LMFAO*, while women use *LOL*. Women were found to use more excitement-inclined words such as *OMG* [40]. Delip et al [40] also observed that possessives features and words following ‘my’ are a high indicator of predictive features of gender. For example, “my_wife”, “my_gf” are high indicators of a male user, while “my_bf”, “my_dress”, “my_husband” are indicators of a female user. They derived unigrams and bigrams of text from tweets, while preserving emoticons and other punctuation sequences. Insight into distinctions in language use across gender, age, regional and political orientation, current trends, and informal communication are also shown in [40].

Burger et al [34] argue that profile elements are not very useful for predicting gender when using supervised learning approaches to classify tweets with the purpose of identifying gender. Instead, they combed through the users who included URL in their profile field, which enabled them to link to blogs and other online media that clearly identify gender. They generated approximately 180,000 Twitter users whose gender was identified and labeled. Also, they manually conducted a small-scale quality assurance to ascertain the correctness of the automatically generated dataset that were labeled with the gender of 1000 users. This was done by checking the description field on user profiles for conspicuous indicators. Their work showed that female users are more likely to show clear gender cues; also, female users perform a status update more often than do the male users. Their work focused on multilingual tweets where they investigated the performance of three different machine-learning algorithms, namely Naïve Bayes, Balanced Winnow2, and Support Vector Machines (SVM). Using the n-grams feature of tweets - the full name, description, and screen name - they constructed a simple Boolean indicator that showed the presence or absence of a specific word or character n-gram in the string of text connecting a particular user profile field. They used combination of various fields and divided their dataset into development, training, and testing. Their work showed that predicting gender from a full name field provided a higher level of accuracy than using screen names and description. Although the full name field from user profile fields proved to be very informative, the status message (i.e., tweets) performed better which, they argue, is more informative in predicting gender. They reported that the

Balanced Winnow algorithms performed better in accuracy, speed, and robustness, even when the not-so-informative features were included. To verify the efficacy of their classifiers, they used Amazon Mechanical Turk (AMT) to annotate the dataset labeled with gender, and their classifiers outperformed AMT.

Employing neural networking and data-mining techniques, William et al [47] were able to identify the gender of Twitter users. They demonstrated that stream-based neural networks could automatically discriminate a user's gender on manually labeled tweets. Their work utilized several feature-selection algorithms from WEKA, and for the task of classification they adopted the three methodologies of (a) Mistake Driven Online Learner, (b) Balanced Winnow Neural Networks, and (c) Modified Balanced Winnow Neural Network. They extracted features from the unigrams and bigrams of tweets to create separate, relevant features; however, the modified balanced winnow algorithm performed best having the most relevant features, as compared to a combination of features. Some of their most relevant features were consistent with the research of [34]. Although their work utilized a small dataset as compared to the dataset used in previous research, they did achieve a higher level of accuracy than baseline performance.

There is much in the literature about using the content of the user status message (i.e., tweets) to determine gender; however, Wendy and Derek [48] argued that using only the user's first name can predict the gender of Twitter users. They particularly argued that the work described in [34] work was not representative of the user base of Twitter's social network. Thus, they gathered their own data and, using 1990 US census names, they evaluated the inference accuracy of three different gender-predicting methods: baseline, integrated, threshold methods. Their approach utilized the Amazon

Mechanical Turk (AMT) to manually construct gender-labeled datasets, which relied on the profile pictures of selected Twitter users. Then to find a match, they compared the names gathered from the 1990 US census to the data they had collected. In addition, they investigated [34] approach by incorporating other content of profile information and received a higher level of accuracy than the baseline. They argue that when their method is combined with full profile information, it yields higher level of accuracy.

Approaching the task of gender classification from a more responsive and efficient point of view, Wendy et al [50] applied [49] approach to a specific domain. Their work focused on identifying the gender composition of the population in Toronto, Canada, that used transportation, such as cars, public transportation, and bikes, to commute. They gathered their data from different accounts dedicated to advertising their particular mode of transportation. In their work, they manually hard coded male and female users, using their profiles, profile pictures, usernames, and tweets. Then they applied the demographic inference algorithm, SVM. Their findings for Twitter users' three modes of transportation - cars, public transportation, and bikes – aligned with the data provided by Statistics Canada 2007. This was evidence that Twitter's platform could be capable of measuring the components of a particular community.

Clay et al [51] presented a region-specific example to identify the gender of Twitter users from the content of their tweets. They applied supervised learning to content of Nigerian Tweet users. They employed unigram features from the tweets only, omitting other, predictive features that had proven useful in determining the gender composition of Twitter users. Though they still relied on names to manually build training sets by searching popular Facebook pages to obtain unique names, they did not use these names in their experiment. They found that females used more emotion-laden words, such as *aww, sad, happy,* and emoticons such as (e.g.☺, ☹), while male users referenced soccer as *game, arsenal.* Their work lent support to the finding of previous research by maintaining that tweets contain information that is highly predictive of gender on Twitter.

Bamman et al [5] argued against using the information to identify the gender of Twitter users as a binary variable. Their work focused on the style, stance and social networks to identify the gender of Twitter users, employing a more subtle approach. They suggested a relationship between gender, linguistic style, and social networks. By clustering Twitter feeds, they constructed a corpus of tweets from 14,000 users, to discover that there is a relationship between styles and interests that reflects a multifaceted interaction between gender and language. Some users' style of language use was in agreement with language-gender statistics; although others' language use contradicted the language-gender statistics. They also did some investigative work on Twitter users whose language use was synonymous with a difference in gender. They found that the composition of a user's network of friends does influence their language use and, in general, the social network homophile associated with the use of same-gender language markers.

CHAPTER 4

METHODOLOGY

This research project utilized a supervised, machining-learning approach to detect the gender of a Tweet sender. In this chapter, discussion includes the setup of the research and the methodology used. Section 4.1 provides a statement of assumptions; Section 4.2 outlines the research approach; Section 4.3 presents the selection techniques used in the project; Section 4.4 discusses the classifier that was used, and lastly, Section 4.5 discusses the steps involved in the classification system.

4.1 Assumptions

This project is predicated on the following assumptions.

1. Within the online community, as on other social networks, users are at liberty to use any name of their choice. For example, females may choose to use a male name and males may choose to use a female name. Beyond their self-declared user names, no effort is made to verify names.
2. The dataset represents users from the United States of America and may be ineffective if a user has changed his or her location.
3. The corpus contains tweets in American English; British English is not considered.
4. Twitter users are at liberty to use screen names which can include letters of the alphabet, as well as numeric and special characters. For instance, @#79hg, @_-yte$, and @)(*jhdwe are possible screen names, but such names are not included in this research.

5. Because of Twitters' restriction to the number of characters that can be included in a tweet, users sometime include a URL to other websites. As well, tweets may include links that Twitter's system generates to shorten more than 140 characters. Because of Twitter's enforced limitation of characters in a tweet, all URLs and links to other websites are disregarded for purposes of this research project.
6. Hashtags (i.e., #) are popular on Twitter to coordinate topics. These hashtags may consist of normal words, or acronyms, or abbreviations, or any other combination of characters. Normal words in hashtags have been used in this project, but if hashtags do not form a normal word, they have not been used. For example, #Moore and #Obama, would be included in the research, but not #USGShutdown.
7. The convention applied to user names in this research project is that the user name begins with first name and is followed by the last name. Therefore, for example, Elizabeth Mathew would be acceptable to this research; however, Mathew Elizabeth would not be acceptable.

4.2 Description of Our Approach

In this project, we used a general machine-learning framework for social network user classification that included the user profile, the user tweeting behavior, the linguistic contents of the user post, and the social network [5], [49], [50], [52], [34]. However, our specific focus was on two of these frameworks - the content of user tweets or messages and user names using support vector machines (SVM). SVM have been widely used in gender identification [34].

4.2.1 User Name

Users on Twitter's social network are free to use any name they choose. As such, they occasionally choose to use weird names such as *Sharksnake* or *puff-adder*, which do not provide a clue about the gender. Others may use a combination of apha-numeric and special characters. Although, these characters are of common use in a user's screen name, they are sometimes used in real names. And many users do choose to use their real names. One can identify the gender of people by their name, because it seems the norm to assign a name that is unique to each gender. For example, Jacob, and Michael, and Matthew are exclusively male names, whereas Abigail, and Elizabeth, and Emily are exclusive to female names. However, there are exceptions to these norms, where the given name could apply to either a female or a male, such as the names Morgan, and Ashley, and Ashton.

The reason for using this model of a machine-learning framework is that not only are most names unique to particular gender, but also a prime factor was the wide availability of a public name resource such as the US census data (http://www.census.gov) and the US Social Security Administration (http://www.ssa.gov). Mislove et al [52] were the first to propose this technique. Based on the assumption that users' self-reported names are in the order of first name then last name, we manually annotated the training data for each gender by matching each name with the names from above-mentioned websites. Where there was a match, we extracted the user from the corpus, alongside their content from Tweet. Of particular interest were names that occurred at a higher percentage in the census and social-security data.

4.2.2 Nonconventional Names

Most names are highly associated with either the male or the female gender, while others are not [49]. Because is not mandatory for users to reveal their identity on Twitter, the use of unknown names is also prevalent. Some unknown user names are discussed in the following.

4.2.3 Nicknames and Abbreviations

Non-names for example, “??????????????”, “Circle the Moon” “The Man on the Moon”, and “J.J.C”, are sometimes self-claimed names or a combination of words from their names that cannot be matched with any specific gender name. These word forms do not follow any kind of naming convention and have no meaning in any dictionary of names. These word forms are usually common to those who want to associate themselves with some phenomenon or celebrity.

4.2.4 Amalgamated Names

Amalgamated names such as “JennaMoran” “Jenny_Meszaros” “Jan-Eric Sundgren” appear to be real names except they are joined together most times, with some special character that makes it difficult for proper clarification.

4.3 Feature Selection

Data generated from social media sites are, in most cases, very noisy, vast, unstructured, distributed, and dynamic in nature [53]. These inherent characteristics pose challenges to efficiently classify text. For purposes of classification, it is often advantageous to identify the most relevant and most informative features. By removing inaccurate and irrelevant features, *feature selection* reduces the original features and, in many cases, leads to an increased rate of accuracy [54]. In addition, the running time of the classification algorithm is decreased when irrelevant features are reduced from feature sets. Most feature selection algorithms calculate a score for each singular feature, based on some predefined measure and, based on the measure, return the top-ranked features.

4.3.1 Characteristics of Salient features

The following present the salient features that were used in the research.

1. **Features must be frequent enough.** More features could make the learning process difficult, because some occur frequently and some, infrequently. However, frequent features help the machine-learning process. With the use of ranking algorithms, the features can be ranked, based on number of occurrences in a corpus. Highly ranked features make the learning process easier and give a better accuracy rate.
2. **Features must be distinctive enough.** Machine-learning systems need to be able to discriminate between features that are peculiar to a certain class. Features are of interest if they are unique to a particular class, but absence in the other. For example, women and men use possessive words distinctively. *My_nigga, my_woman, my_wife* are a common expression of males, while *my_husband, my_nails, my_eyebrows* are common to females [40].
3. **Features should be comprehensive and representative of the entire corpus.** Typically, most corpuses contain a large volume of data which, in turn, contains subsets that comprise the entire corpus. Some features may be very accurate in representing a particular subset, while other features may represent another subset.

4.3.2 Feature-Selection Methods

Features can be very large; some may be very useful, while others may not. Using feature- ranking filters can produce feature sets with less noise that result in those features that are most predictive. To rank the features, this project utilized ranker-filter algorithms from the WEKA toolkit. The following algorithms were used in this research project - Chi-Square, Information Gain Ratio, Information Gain, Relief, and Symmetrical Uncertainty. All these feature-selection algorithms rank features, based on the score computed for each feature.

4.3.3 Features

Research in this area proposed a rich set of features derived from the content of various Twitter users [34], [49], [38], [55], [50]. This project adopted some of these feature sets. The manually labeled sets were divided into the two groups of male and female, and features were then extracted, based on the following.

- ***k-top words.*** These represent the *k* most discriminating words used by each labeled set. The individual features were included into the training set as feature sets.
- ***k-top stems.*** Many English words are the derivative of another word. Action words in verb form and plural words can cause certain words to be treated separately, and this may weaken the word signal. For example, “fishing”, “fisher”, “fished” will be treated as separate words if not stemmed or reduced to their base word “fish”. To achieve this, all words were filtered through the popular stemmer, Lovins Stemmer, that is built into WEKA [56]. This process returned words to their base word and were included in the training sets.
- ***k-top digrams and trigrams:*** As described above, digram and trigram *k* most discriminating words in each labeled group were included as feature sets.

```
@relation 'training set'
@attribute birthdai numeric
@attribute bitch numeric
@attribute dont_know numeric
@attribute black numeric
@attribute cute numeric
@attribute leav numeric
@attribute live numeric
@attribute lmao numeric
@attribute nigga numeric
@data
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,f
0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,f
0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,m
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,m
```

Figure 5. Sample of converted data

4.4 Classifier

This research project used support vector machines which have been widely used in prior works with quite a high rate of performance [34], [38], [48], [49], [50]. The avalanche of prior research that have utilized this machine-learning classifier made it a choice classifier for this research. Another reason for choosing this classifier is that we were able to compare our results with the results of the research of [34] and of [49], [48].

According to Jakkula [57]

> “support vector machines can be defined as systems which use hypothesis space of a linear function in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory”.

SVM is a binary, linear non-probabilistic classifier that inputs data and predicts an output for every given input. The output shows which of the two possible classes to which it could belong. SVM machine-learning algorithm builds models based on the training sets. The model that is built is used to predict new sets, based on the categories in the model.

4.5 Classification

For the purposes of classification, this research project used WEKA data-mining software. The classification task began with data preprocessing, where the dataset are cleaned to remove unwanted features that include stop words, for example, “in,” “the,” “for,” “come”, et cetera. The data are then tokenized and converted to Attribute-Relation File Format (ARFF). WEKA presents powerful feature-selection algorithms that help to rank features, with the most useful ranking above the others. This helps to speed up the process and thereby less memory is consumed. Different feature-selection algorithms were used to generate high-ranking differentiating features to train the classifier, SVM. The features threshold started from 4 to 8 were used to first select the features, before they were ranked. A baseline feature of 4 (word frequency) was used to represent all features before increasing. The two classification approaches used were Baseline method and Integrated method. The baseline method has no user name association, whereas the integrated method has a user name association. The classifier was then trained on each feature set, using ten-fold cross validation. Figures 5 and 6 show the detailed classification process.

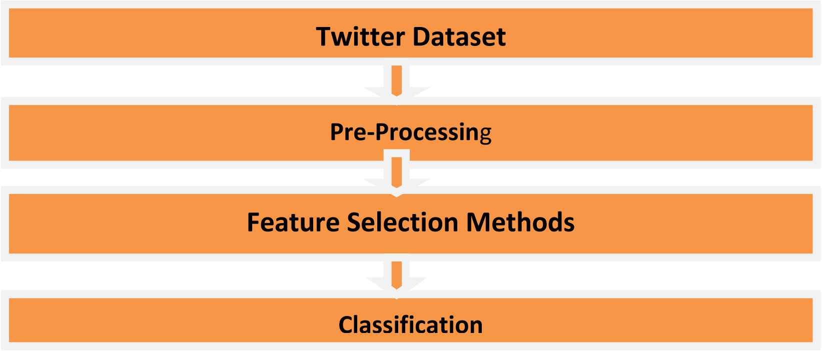


Figure 6. Methodology of the classification process.

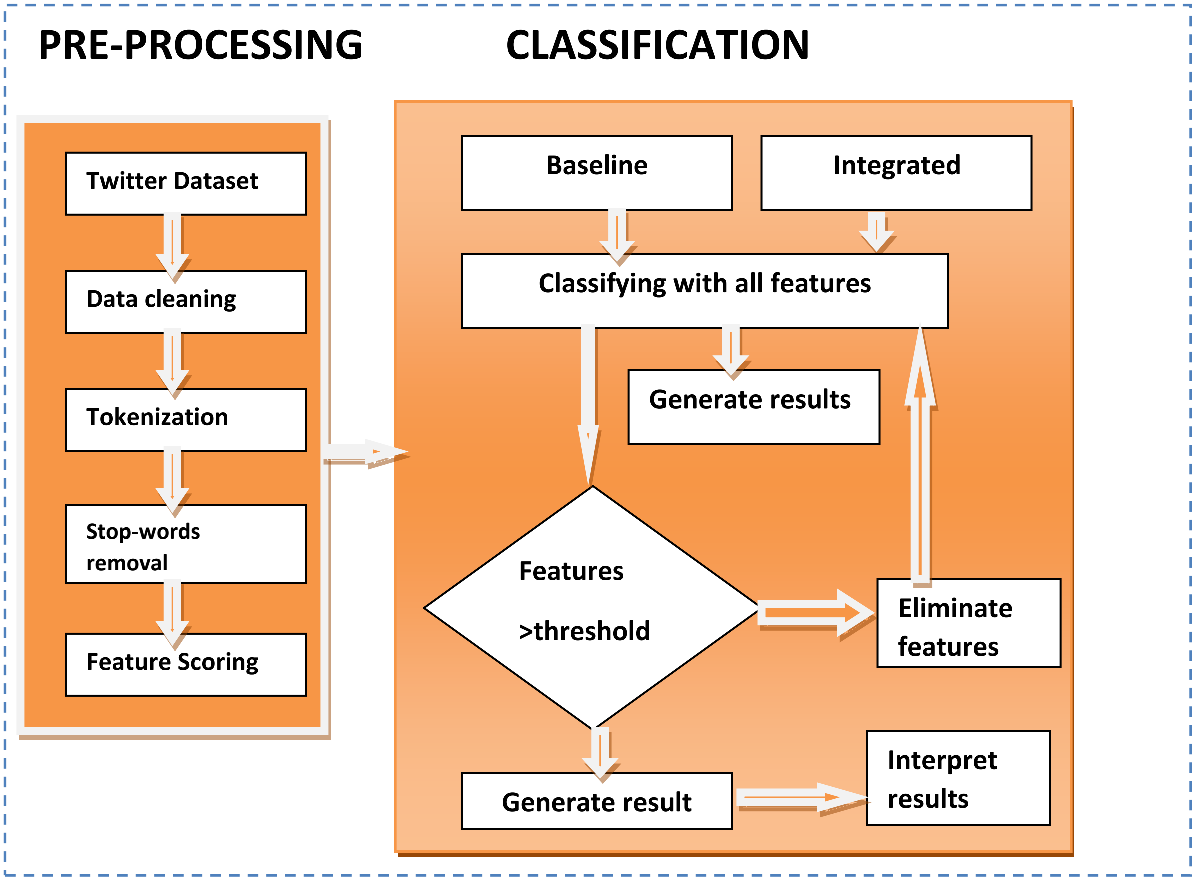


Figure 7. Classification execution.

CHAPTER 5
EXPERIMENTAL RESULTS

In this chapter we discuss various experiments executed and present results. In section 5.1 we discuss the details of hardware used. In section 5.2, we will give details of the data mining software we used to conduct our experiments. In section 5.3 we will discuss the details of Twitter dataset. In section 5.4, we will detail yardstick for measuring the experiments. In section 5.5, we will show our results. In section 5.6 we will discuss our results.

5.1 Hardware and Software.

This work utilized two machines to conduct the experiments. The first machine was a Samsung Notebook 530U4B-S01 with installed Windows 7 operating system. Processor Intel(R) Core(TM) i5-2467M CPU @ 1.60GHz, 1601 MHz, 2 Core(s), 4 Logical Processor(s), 1Terabyte of Hard drive, 6GB of RAM. The other machine was Dell desktop PC with Windows XP Professional with 2GB RAM, Intel Pentium (R) 4 CPU 3.4GHz processor and 250G hard disk space. We also used Microsoft Excel 2010 Java SE for data preprocessing

5.2 WEKA

WEKA is an acronym for Waikato Environment for Knowledge Analysis. WEKA is a collection of state-of-the-art machine learning algorithms for data mining tasks. It also contains data preprocessing, classification, regression, clustering, data visualization tools, and association rule. WEKA is written in Java under GNU General Public License open source software with cross platforms compatibility. The machine learning algorithms can either be utilized directly to a dataset or called from personal Java code. It is also appropriate for developing new machine learning schemes [59]. ARFF (Attribute-Relation File Format), CSV, C4.5 and binary format are the types of files that all WEKA algorithms take as their input, it can also be in the form of a single relational or generated from an SQL database (using JDBC) or read from a URL [58].

WEKA has gained popularity in performing data mining task with strong classification capabilities. It comes with four interfaces; the Explorer, the KnowledgeFlow, the Experimenter and the command line interface. The explorer contains tools that can be used for data pre-processing operations such as attribute selection, normalization, and transformation in classification, regression, clustering, association rule, and data visualization. The Experimenter provides means for performance comparison of different learning algorithms for classification and regression problems with capability to write results in file or database. The KnowledgeFlow helps to provide alternative to the Explorer as a graphical frontend to WEKA's core machine learning algorithms. However, the KnowledgeFlow can support limited function when compared with the Explorer. The KnowledgeFlow presents an imaginative data-flow interface to WEKA. It provides the user the option to select WEKA's components from a toolbar, position them on a layout that is canvas and join them together in order to produce the knowledge flow.

The command line interface which can be accessed by entering textual commands provides low level access to basic functionality. It also accepts Java syntax as command.

5.3 Datasets

As of now, Twitter allows public access through its API for the general public to get tweets. Twitter does prevent sharing of tweets to other parts other than the intended user; as a result there are no publicly available tweets datasets. For this work, US users of Twitter were our focus for two basic reasons; (1) our focus was on English speakers and (2) availability of gender tag name resources from US Social Security Administration. For each Twitter user, Twitter provides geo-location information, this helps when gathering tweets based on location of the user. We were able to gather 30,000 tweets with user names, screen names and content of tweet as shown in table 5.1. Although each of the profile details is completely optional when a user is registering on the social network, hence the use of some profile details that tell little about the users. The user profile features are not very useful when trying to apply supervised learning methods to learn gender attributes. Other research have used manual labor-intensive method to label users based on demographical attributes in [34] in user web link fields available on their profiles were crawled to link to other websites where they clearly defined their gender. In [40] a manual methodology to annotate users based on their gender was used. In this work, we manually annotated user tweets based on gender.

Screen Names	User Names	Tweets
Andrey Fadeev	Fadandr	I just unlocked the Epic Swarm badge on @foursquare for checking in with 999 other people! http://t.co/BmAz0wMg
Emily	theonlyemilyb	I tell myself you don't have the power to ruin my mood, yet every time it's still always you
Mike Fortman	courtneyykaay	i think it's kinda funny how much effort people put in trying to hurt you.. #jokesonyou
Courtney	Franklin_MEWX	Overcast and 48 F at Bar Harbor Automatic Weather Observing / Reporting ME Winds are South at 9.2 MPH (8 KT). The pres http://t.co/vEd6rVUr
Frederick Gloria	Fredzillasaurus	I want to meet Shakira so that I can ask her hips the truth about love life meaning existence and reality because they dont lie!
Just George	G_A_Fegley	Its my last compulsory school day ever and Ive eaten so many doughnuts I could open a bakery in my gut hahaha)

Table 5.1 Tweets

The details of tweets above are not constant; they can be changed at any given time or deleted. To annotate our training set, we manually compare a user provided name with gender names publicly available on US Social Security Administration. From this process, 500 users for each gender (male & female) were annotated. We partitioned our dataset into training and testing sets. The annotated users served as our training set.

5.4 Interest Measures

To measure most classifier performance, recall, precision, f-measure and accuracy has been highly used. Recall measure the ability of a classifier to classify all relevant terms into a class. The higher the recall, the fewer the false negative classification (positive tweets which have been wrongly classified as negative), while lower recall means higher false negatives. Accuracy measures the overall correctness of a classifier; it is the sum of correct classification divided by the total sum of classifications. Precision is the measures of exactness of a classifier. The higher the precision, the fewer false positives (negative tweets which have been wrongly classified as positive), while lower precision means higher false positives. The f-measure “is the average of precision and recall but gives higher scores when precision and recall results are closer together” [63]. It is the harmonic mean of recall and precision

$$Recall = \frac{tp}{tp+fp}$$

$$Precision = \frac{tp}{tp+fp}$$

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall+Precision}$$

$$Accuracy = \frac{tp+tn}{tp+fp+tn+fn}$$

where *tp* is true positive, *fp* represent false positive, *fn* represent false negative and *tn* represents true negative.

Here we focus on accuracy. Results for the other measures are similar and can be found in Appendix

5.5 Results

Two gender inference methods were used during experimentation; baseline and integrated employed by [48]. SVM was the classifier used throughout this research because it gives good results when conducting text classification. The rest of the chapter will be devoted to discussing our results and comparing with other works.

5.5.1 Results from Baseline Inference Method

The baseline inference method deals with features of each class alone with no user names association. This was done to enable comparison with the other case where gender names are associated with each class. Table 5.2 shows the summary of ten-cross validation accuracies with different number of features threshold.

Feature threshold	Number of instances	Accuracy
12	**1400**	**86.8**
10	1400	85.6
8	1400	82.5
6	1400	79.8

Table 5.2 Accuracy for Baseline Inference Method

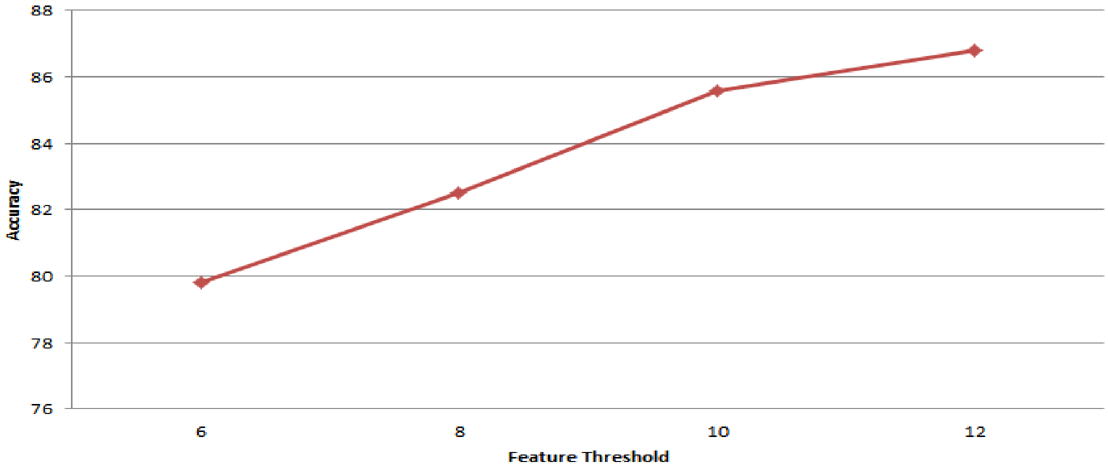


Figure 8. Accuracy for Baseline Inference Method

5.5.2 Results from Integrated Inference Method

The integrated inference method deals features that are associated with user names. This was done to ascertain the contribution of username associability. However, we found out that associating usernames with features improved accuracy as shown in the Table 5.3 below.

Feature threshold	Number of Instances	Accuracy
12	1400	**95.3**
10	1400	94.7
8	1400	94.5
6	1400	91.3

Table 5.3 Accuracy for Integrated Inference Method

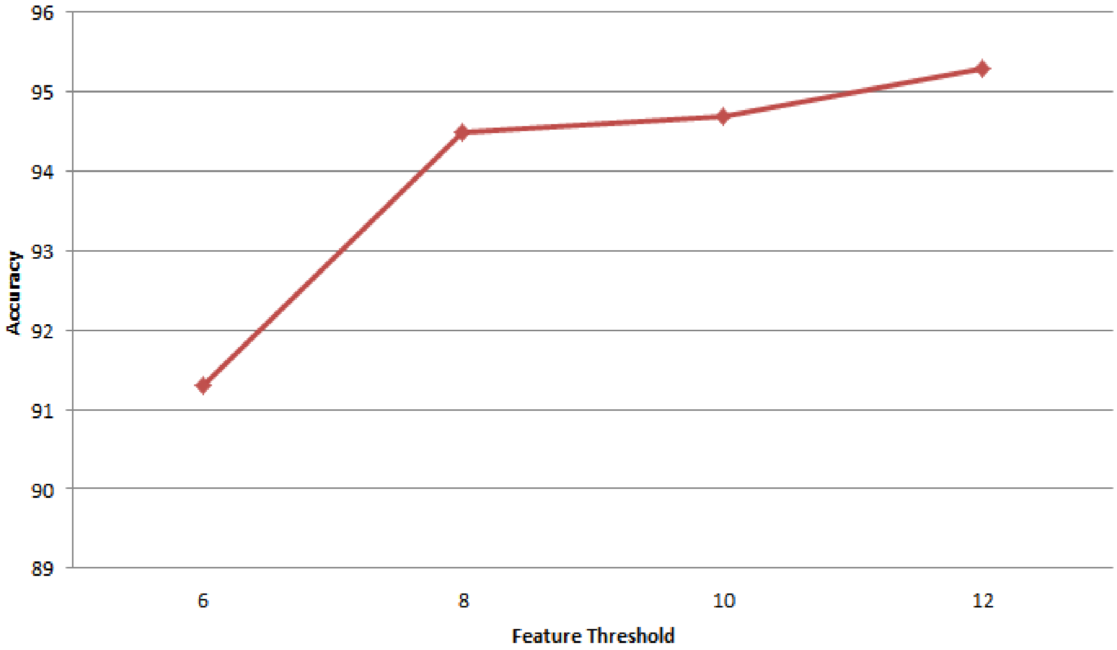


Figure 9 Accuracy for Integrated Inference Method

5.6 Discussion

In this section, we discuss the results from our experiments presented in the previous section and make comparison with other work.

5.6.1 Performance Analysis

Table 5.2 and Table 5.3 shows the summary of results for the baseline and integrated methods. We noticed an increment in feature threshold, results in higher accuracy. This is as a result of increase in uniqueness of features; the higher the threshold the more distinct the feature. The integrated inference method gave the highest accuracy when compared to the baseline inference method. The increase in accuracy for integrated method is as a result of inclusion of gender names as features. On the other hand, the baseline inference method gave a lower accuracy of 79.8% when the feature threshold was set to 6. In the baseline method in Table 5.2, the SVM classifier solely rely on features generated from tweets and does not have the privilege to benefit from names association with tweets hence the reason for the lower accuracies. We got highest accuracy of 95.3% when feature threshold was set to 12. Both baseline and integrated inference methods does have commonality which can be observed in the way feature thresholds affects the accuracies as shown in figure 5.1 and 5.2. It is intriguing to note that, in both cases, accuracy grows proportionately to number of threshold.

5.6.2 Comparison with Other Results

In this research, we have leveraged on some approaches used in previous work. First we compare our results with those reported in [48] were the highest accuracy achieved 85.2%. They, however, used profile pictures as the means to identify gender. This method, however, has one major problem. Many users who tend to like a particular person may choose to use that person’s picture as their profile picture. This is usually very common of fans; they sometimes use celebrities and other political bigwigs’ pictures as their profile picture.

Method	**Accuracy**
Baseline and Integrated using gender names with higher percent of occurrence as part of features and different feature thresholds (this work)	**95.3%**
Baseline and Integrated using Profile Picture to manually hard code label [48].	85.2%
Social linguistic, Stacked and n-gram by crawling male and female Twitter accounts products to identify gender [40].	72.33%

Table 5.4 Accuracy comparison

Another comparison is the work done by [40] which used social linguistic, n-gram and stacked as features selection methods. They crawled accounts of gender specific products on Twitter to label to gender. This not very useful as many of those who follow these gender specific accounts may not all belong to one gender as the authors presumed. This may not be unconnected to the low accuracy 72.33% they reported.

In each case, this research outperformed both with an accuracy of 95.3%. This can be attributed to the method used in manually coding gender for training the SVM classifier. We relied on user self-declared names by comparing those names with US Social Security publicly available gender labeled names. We took note of those names that were strictly unique to each gender. The US Social Security is an authentic source of gender labeled baby names.

5.7 Drawbacks

Social media data are generally unstructured in nature and a great deal of work is often required to clean the data. In so many cases, the meaning and structures of words changes significantly from conventional English language. This causes distortion in some words thereby affecting the accuracy achieved. Hashtag (#) is very popular on Twitter which is mostly used for topic discovery and coordination. Sometimes, hashtags contains unique feature needed to perform classification, but in cases where the hashtag were used to coordinate gender specific topics (hashtag used maybe a meaningful word unique to a specific gender), we are not able to properly distinguish gender because such hashtags were used for gender precise topics coordination. The inclusion of these hashtags influenced the performance of the classifier and also affected the results. We relied on word frequency to extract features from each gender.

Another major drawback is the reliability of users' self-declared names, since we did not make effort beyond what the users claimed to be their names. This is particularly a problem because for some reasons best known to them, users may choose not to use their real gender name; for instance a male user may claim to be female by using female username or verse versa.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Both organizations and individuals are increasingly dependent on social media as a means of sharing, communication and social interaction,. As a result researchers have been provided with new data sources and opportunities for mining this data to improve knowledge and service delivery. Twitter, a very popular social media site, has received increased research attention in recent times as a means of monitoring, understanding, even sometimes predicting real-life phenomenon. Twitter gender authorship detection of users can, however, be very complicated because tweets are inherently short and in very many instances they contain informal text, slangs words, even deliberate typos. This makes it very difficult to find informative features to help in the task of gender identification.

In this research, we present results of experiment conducted using machine learning approach to predict a tweet sender's gender. Since Twitter does not collect gender information on users, we devised an approach by manually annotating a training set using users' self-declared names. We relied on US Social Security Administration (SSA) publicly available gender associated names data to annotate the gender.

The training set enabled us to train an SVM classifier using a set of features. The trained classifier was tested on a test set with and without names associated features.

Our results showed a higher performance with both inference methods giving accuracies of 86.8% and 95.4% when feature thresholds were set to 12. However, we observed an increase in accuracy of 8.5% when gender names were associated with tweet content (integrated inference method) over the baseline inference method. We can conclude that gender names integration contributes to higher accuracy. We also observed that increasing the feature thresholds contribute to better accuracies since these increases helps to bring out more refined features.

Topic based coordination of tweets is a popular way of bringing together interest group using hashtags (#), we hope to conduct further study on topic discovery unique to each gender using hashtag . This research can also be extended to identify influence of gender on retweets; messages of a follower or a followed user re-post on their timeline,

REFERENCES

[1] Andreas M. Kaplan, Michael Haenlein (2010). Users of the World, Unite! The challenges and opportunities of Social Media. Kelley School of Business, Indiana University.

[2] Monthly Active Users MAU"Facebook Reports First Quarter 2013 Results". Facebook. URL: http://investor.fb.com/releasedetail.cfm?ReleaseID=761090 Retrieved 17th May 2013.

[3] Hayley Tsukayama, Twitter turns 7: Users send over 400 million tweets per day. URL: http://articles.washingtonpost.com/2013-0321/business/37889387_1_tweets-jack-dorsey-Twitter. Retrieved 7th June, 2013.

[4] Philip N. Howard, Aiden Duffy, Deen Freelon, Muzammil Hussain,Will Mari, Marwa Mazaid.. Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? URL: http://pitpi.org/wp-content/uploads/2013/02/2011_Howard-Duffy-Freelon-Hussain-Mari-Mazaid_pITPI.pdf. Retrieved 5th August, 2013

[5] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. (2012). Gender in Twitter: Styles, stances, andsocial networks. Technical Report 1210.4567, arXiv, October.

[6] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011): Sentiment analysis of Twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38

[7] Brendan O’Connor (2010). From Tweet To Polls: Linking Text Sentiment to Public Opinion Time Series, Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC. Pp. 122-129,

[8] Sitaram Asur, Bernardo A. Huberman, (2010). Predicting the Future with Social Media. Proceeding WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01pp. 492-499

[9] Johan Bollena, Huina Maoa, Xiaojun Zengb, (2010). Twitter mood predicts the stock. In Journal of Computational Science, Volume 2, Issue 1, pp.1–8

[10] Duric, A., Song,F. (2011). Feature selection for sentiment analysis Based on Content and Syntax Models, Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT, Portland, Oregon, USA pp 96–103,

[11] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. (2011). Big data: The next frontier for innovation, competition, and productivity. Report McKinsey Global Institute. URL:ttp://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation Retrieved 22nd June, 2012. Retrieved 5th, July, 2013

[12] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni (2003). Gender, genre, and writing style in formal written texts. TEXT pp. 321--346...

[13] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression, First Monday, 2007- firstmonday.org

[14] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni.. (2002). Automatically Categorizing Written Text by Author Gender. Literary and Linguistic Computing.

[15] Eckert, P. (1997). Gender and sociolinguistic variation, in J. Coates ed., Readings in Language and Gender (Blackwell, Oxford), pp. 64-75.

[16] Holmes, J. (1990). Hedges and boosters in women's and men's speech, Language & Communication.

[17] Herring, S. (1996). Two Linguistic Correlates of Gender and Sex. English World Variants of an Electronic Message Schema, in S. Herring ed., Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives (John Benjamins, Amsterdam), pp. 81-106.

[18] Chambers, J. K. (1992). Linguistic correlates of gender and sex. English World-Wide 13(2):173–218.

[19] Chambers, J. K. (1995). Sociolinguistic theory: Linguistic variation and its social significance. Oxford: Blackwell.

[20] Labov, William. (1966). The social stratification of English in New York City. Washington, D.C.: Center for Applied Linguistics.

[21] Trudgill, Peter. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. Language in Society 1(2):179–195.

[22] Savicki, V., Lingenfelter, D. and Kelley, M. (1996), Gender Language Style and Group Composition in Internet Discussion Groups. Journal of Computer-Mediated Communication.

[23] Bell, A. (1984). Language style as audience design. Language in society 13(2):145–204. 22.

[24] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, Theo Meder (2013). “How Old Do You Think I Am?”: A Study of Language and Age in Twitter. A Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media

[25] Holmes, J., and Meyerhoff, M. (2003). The handbook of language and gender. Oxford: Blackwell.

[26] Schiffman, H. (2002). Bibliography of Gender and Language. URL: http://ccat.sas.upenn.edu/~haroldfs/ popcult/bibliogs/gender/genbib.htm. Retrieved 20th June, 2013

[27] Dorsey, Jack (March 21, 2006). "just setting up my twttr". Twitter. Com. URL: https://Twitter.com/jack/status/20. Retrieved 23rd February, 2013.

[28] Bernard J. Jansen, Mimi Zhang, Kate Sobel (2009). Twitter Power: Tweets as Electronic Word of Mouth. Journal of the American Society for Information Science and Technology pp. 2169–2188

[29] Finin, Tim Tseng, Belle, (2007). Why We Twitter : Understanding Microblogging. Joint 9th WEBKDD and 1st SNA-KDD Workshop on Web mining and social network analysis. San Jose, California , USA . pp. 56-65

[30] Paul Gil, About.com Guide, (2012). What Exactly Is 'Twitter'? What Is 'Tweeting'? URL: http://netforbeginners.about.com/od/internet101/f/What-Exactly-Is-Twitter.htm. Retrieved 21st July 2013,

[31] Ingrid Lunden ,(Monday, July 30th, 2012). Analyst: Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta ‘Biggest Tweeting’ City. Techcrunch.com. URL: http://techcrunch.com/2012/07/30/analyst-Twitter-

passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/. Retrieved 16th April, 2013

[32] Twitter [Twitter] (18th December, 2012). There are now more than 200M monthly active @Twitter users. You are the pulse of the planet. We're grateful for your ongoing support! URL:https://Twitter.com/Twitter/status/281051652235087872. Retrieved 5th July, 2013

[33] Dey, L., Haque, S. M., Khurdiya, A., and Shroff, G. (2011). Acquiring Competitive Intelligence from Social Media. In Proceedings of the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data

[34] Burger, J.; Henderson, J.; and Zarrella, G. (2011). Discriminating Gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301-1309

[35] Hemant Purohit, Andrew Hamptonb, Valerie L. Shalinb, Amit P. Shetha, John Flach, Shreyansh Bhatt (2013).What Kind of #Conversation is Twitter? Mining #Psycholinguistic Cues for Emergency Coordination. Journal of Computer In Human Behavior. pp. 2438-2447

[36] Vangie Beal, (2010). Twitter Dictionary: A Guide to Understanding Twitter Lingo, http://www.webopedia.com/quick_ref/Twitter_Dictionary_Guide.asp, Retrieved 25th June, 2013.

[37] Pete Cashmore, (2008). Twitterspeak: 66 Twitter Terms. URL:http://mashable.com/2008/11/15/Twitterspeak/. Retrieved 22nd June, 2013

[38] Marco Pennacchiotti, Ana-Maria Popescu (2011). A machine learning approach to Twitter user classification. In Proceedings of AAAI Conference on Weblogs and Social Media.

[39] Akshay Java , Xiaodan Song , Tim Finin , Belle Tseng (2007). Why we Twitter: understanding microblogging usage and communities, Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web mining and social network analysis San Jose, California. pp. 56-65

[40] Delip Rao , David Yarowsky , Abhishek Shreevats , Manaswi Gupta, (2010). Classifying Latent User Attributes in Twitter, Proceedings of the 2nd international workshop on Search and mining user-generated contents. Toronto, ON, Canada. pp. 37-44

[41] Y. R. Tausczik and J. W. Pennebaker, (2010). The Psychological Meaning of Words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology. pp. 24-54

[42] Emre Kiciman. (2010). Language differences and metadata features on Twitter. In Proc. SIGIR 2010 Web N-gram Workshop, pages 47--51..

[43] Cristian Danescu-Niculescu-Mizil Michael Gamon, Susan Dumai (2011). Mark my Words!: Linguistic style Accommodation in Social Media. Proceeding WWW '11 Proceedings of the 20th international conference on World Wide Web. pp. 745-754

[44] Hemant Purohit, Andrew Hamptonb, Valerie L. Shalinb, Amit P. Shetha, John Flach, Shreyansh Bhatt (2013).What Kind of #Conversation is Twitter? Mining #Psycholinguistic Cues for Emergency Coordination, Preprint submitted to Computer In Human Behavior.

[45] W. Levelt and S. Kelter (1982). Surface Form and Memory in Question Answering. Cognitive Psychology, 14(1):78-106.

[46] Analytics, P., “Twitter study- August 2009”, URL: http://www.peranalytics.com/blog/wpcontent/uploads/2010/05/Twitter-Study-August-2009.pdf, Retrieved 15th July, 2013

[47] William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, Wei Hu (2012). Gender Identification on Twitter Using the Modified Balanced Winnow. A Scientific Research publication in Communications and Network. pp. 189-195

[48] Wendy Liu. and Derek Ruths, (2013). What ’ s in a Name ? Using First Names as Features for Gender Inference in Twitter. InSymposium on Analyzing Microtext.

[49] Zamal, Faiyaz Liu, Wendy and Ruths, Derek. (2012). Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In Proceedings of the International Conference on Weblogs and Social Media.

[50] Wendy Liu and Faiyaz Al Zamal and Derek Ruths (2012). Using Social Media to Infer Gender Composition of Commuter Populations, Association for the Advancement of Artificial Intelligence (www.aaai.org). AAAI Technical Report WS-12-04 When the City Meets the Citizen

[51] Clay Fink,a Jonathon Kopecky,a Maksym Morawskib (2012). Inferring Gender from the Content of Tweets: A Region Specific Example. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media

[52] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, J. Niels Rosenquist. (2011). Understanding the Demographics of Twitter Users. In Proceedings of the International Conference on Weblogs and Social Media

[53] Pritam Gundecha, Huan Liu (2012). Mining Social Media: A Brief Introduction. A tutorial in operation research. Inform 2012

[54] Ian H. Witten and Eibe Frank, (2005). Data mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco

[55] Mayur Rustagi, Rajendra Prasath, Sumit Goswami, Sudeshna Sarkar (2009). Learning Age and Gender of Blogger from Stylistic Variation. In Proceedings of the International Conference on Pattern Recognition and Machine Learning.

[56] Lovins, Julie Beth (1968). "Development of a Stemming Algorithm". Mechanical Translation and Computational Linguistics. pp. 22–31.

[57] V. Jakkula (2006). Tutorial on Support Vector Machine (SVM). School of EECS, Washington State

[58] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations.

[59] Mark Hall & Peter Reutemann (2008); WEKA KnowledgeFlow Tutorial for Version 3-5-8, The University of Waikato, New Zealand. July 14, 2008.

[60] http://www.webopedia.com/TERM/_/_reply.html, Retrieved 7th July, 2013

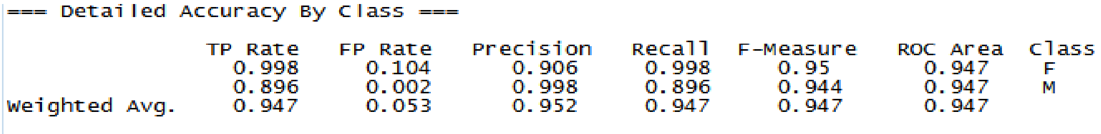
[61]. How to Hashtag - Don't Abuse the Hashtag! (n.d.). URL: http://www.howtohashtag.com/ Retrieved 6th June, 2013

[62]. Corinna Cortes, Vladimir Vapnik. (1995). "Support-vector networks". Machine Learning 20 (3): 273

[63]. Jennifer Zaino, (April 14, 2010). “Everything I Need To Know About Sentiment Analysis” URL: http://semanticweb.com/everything-i-need-to-know-about-sentiment-analysis_b578. Retrieved 7th July, 2013

APPENDIX

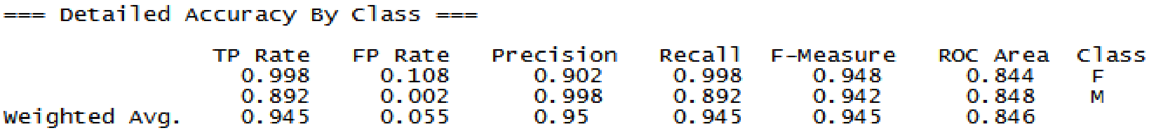
Data/text Feature threshold:12 Number of Instances: 1400 Method: Integrated



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.104      0.906     0.998     0.95       0.947    F
                 0.896     0.002      0.998     0.896     0.944      0.947    M
Weighted Avg.    0.947     0.053      0.952     0.947     0.947      0.947
```

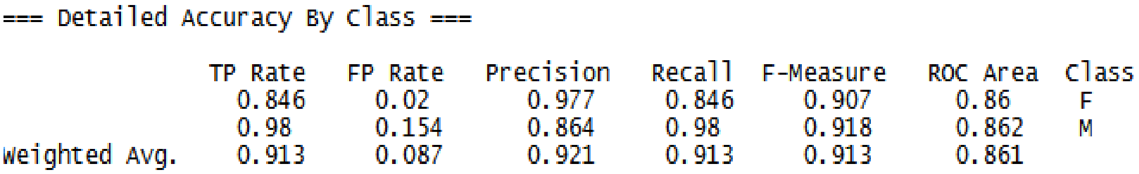
Data/text Feature threshold:10 Number of Instances: 1400 Method: Integrated



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.108      0.902     0.998     0.948      0.844    F
                 0.892     0.002      0.998     0.892     0.942      0.848    M
Weighted Avg.    0.945     0.055      0.95      0.945     0.945      0.846
```

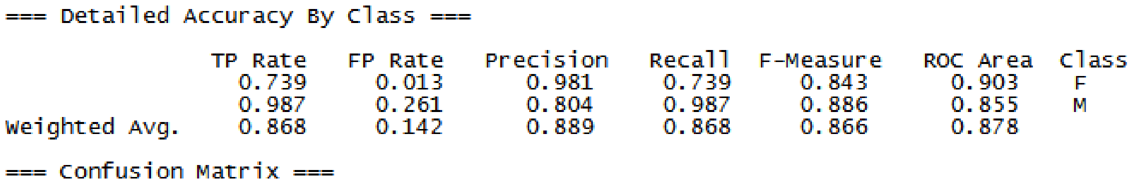
Data/text Feature threshold:8 Number of Instances: 1400 Method: Integrated



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.846     0.02       0.977     0.846     0.907      0.86     F
                 0.98      0.154      0.864     0.98      0.918      0.862    M
Weighted Avg.    0.913     0.087      0.921     0.913     0.913      0.861
```

Data/text Feature threshold:12 Number of Instances: 1400 Method: Baseline

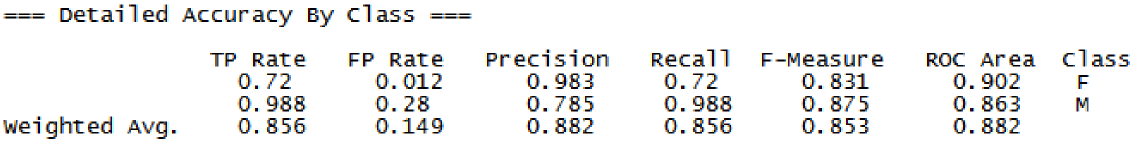


```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.739     0.013      0.981     0.739     0.843      0.903    F
                 0.987     0.261      0.804     0.987     0.886      0.855    M
Weighted Avg.    0.868     0.142      0.889     0.868     0.866      0.878

=== Confusion Matrix ===
```

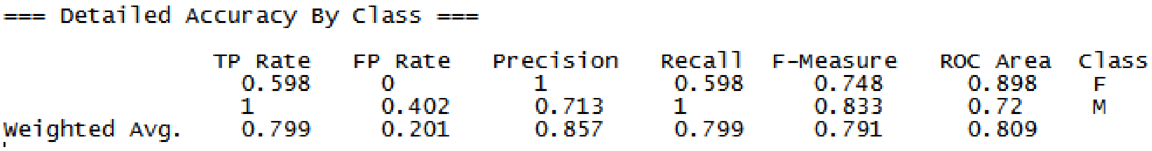
Data/text Feature threshold:10 Number of Instances: 1400 Method: Baseline



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.72      0.012      0.983     0.72      0.831      0.902    F
                 0.988     0.28       0.785     0.988     0.875      0.863    M
Weighted Avg.    0.856     0.149      0.882     0.856     0.853      0.882
```

Data/text Feature threshold:8 Number of Instances: 1400 Method: Baseline



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.598     0          1         0.598     0.748      0.898    F
                 1         0.402      0.713     1         0.833      0.72     M
Weighted Avg.    0.799     0.201      0.857     0.799     0.791      0.809
```