Detecting Sarcasm on Twitter: A Behavior Modeling Approach

by

Ashwin Rajadesingan

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved September 2014 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Subbarao Kambhampati
Heather Pon-Barry

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

Sarcasm is a nuanced form of language where usually, the speaker explicitly states the opposite of what is implied. Imbued with intentional ambiguity and subtlety, detecting sarcasm is a difficult task, even for humans. Current works approach this challenging problem primarily from a linguistic perspective, focussing on the lexical and syntactic aspects of sarcasm. In this thesis, I explore the possibility of using behavior traits intrinsic to users of sarcasm to detect sarcastic tweets. First, I theorize the core forms of sarcasm using findings from the psychological and behavioral sciences, and some observations on Twitter users. Then, I develop computational features to model the manifestations of these forms of sarcasm using the user's profile information and tweets. Finally, I combine these features to train a supervised learning model to detect sarcastic tweets. I perform experiments to extensively evaluate the proposed behavior modeling approach and compare with the state-of-the-art.

DEDICATION

To Amma, Appa, Deepi and Siddy

ACKNOWLEDGEMENTS

I would like to thank my advisor, mentor and life coach, Dr. Huan Liu, for his immense support throughout the course of my Masters degree. His pragmatism and astute advice helped me through my journey here, technical and otherwise. I am sure the lessons learnt will significantly enable my future endeavors.

I would also like to thank committee members Dr. Heather Pon-Barry and Dr. Subbarao Kambhampati for consenting without hesitation to be on my thesis committee. Suggestions and constructive criticism from them helped improve this thesis immensely, making the work more interesting and engaging than what I had initially envisioned.

Special thanks to Reza Zafarani, my friend and collaborator, who patiently answered all my queries and doubts throughout this thesis work. I also thank Fred Morstatter, Suhas Ranganath and other DMMLers for making my time spent here so much more meaningful. I learnt a lot and made great memories. I am also grateful to the Office of Naval Research whose financial support helped me throughout my Masters studies. This work was supported, in part, by the Office of Naval Research grant N000141410095 and Minerva grant N000141310835.

Life in Tempe would have been a lot less fun if I didn't have an amazing group of friends. Thanks Mukund, Mani, Arpit, Rashmi, Mouna, Nikhil, Megha, Niranjan, Sandeep, Malvika and Shibani - couldn't have done it without you guys.

Of course, thank you mom, dad, Deepi and Siddy for encouraging me and standing by me throughout my life - words don't do justice to everything you've done for me. Also, thank you Oviya, for the countless nights (and days) of encouragement and reassurance - you are amazing!

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In recent years, social networking sites such as Twitter have gained immensely in popularity and importance. According to a Pew Research Center study[1], as of September 2013, 74% of all online adults use social networking sites, up from less than 30% in 2008. These sites have not only gained users but also multiple functionalities - they have become an ad-hoc source of entertainment, news, information et cetera (Whiting and Williams, 2013; Kwak *et al.*, 2010). These sites have evolved from simple platforms where users connect to each other and keep in touch, to large ecosystems where users, among other things, express their ideas and opinions uninhibitedly. Nowadays, with social media forming a part of our everyday lives, users candidly share a wide breadth of information, from the relatively mundane to the highly personal. From a sales and marketing perspective, companies have unbridled access to this unique ecosystem to gain critical insights into the mindset and thought process of their customers and to better serve their needs. They can tap into public opinion on their products or services and even provide real-time customer assistance through social media. Not surprisingly, most large companies have a social media presence and a dedicated social media team working on marketing, after-sales service, and consumer assistance.

Given the high velocity and volume of social media data, companies rely on automated social media management tools such as HootSuite[2], to analyze data and to

[1]PewResearch Internet Project, http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

[2]Hootsuite, https://hootsuite.com/

provide customer service. These tools perform tasks such as content management, sentiment analysis and extraction/filtering of relevant messages for the company's customer service representatives to take action. While these tools perform well for basic tasks, they lack the necessary sophistication to decipher more nuanced forms of language such as sarcasm, in which the meaning of a message is not always obvious and explicit. This is quite a handicap, especially in the context of social media where the relative ambiguity and the ability to hide behind computer screens often encourages snarky, rude and sarcastic posts. The lack of a viable sarcasm detection mechanism imposes an extra burden on the company's social media team, who are already inundated with customer messages, to identify these sarcastic messages and respond appropriately. Table 1.1 provides two examples where the customer service representatives fail to detect sarcasm. Such public gaffés not only upset the already disgruntled customers but also ruin the public images of companies.

Interestingly in June 2014, the United States Secret Service also issued a work order seeking social media software capable of detecting sarcasm [3], explicitly stating that social media tools currently in the market do not have the capability of detecting nuanced forms of language such as sarcasm.

Our goal in this study is to tackle the challenging problem of sarcasm detection on Twitter. While sarcasm detection is inherently complex and difficult, the style and nature of content on Twitter further complicate the process. Compared to other, more conventional sources such as news articles and novels, Twitter [i] is more informal in nature with an evolving vocabulary of slang words and abbreviations and [ii] has a limit of 140 characters per tweet which provides fewer word-level cues thus adding more ambiguity. However, Twitter provides other information such as social graphs, past tweets and profile bio details, which when used effectively, may help overcome

[3]Solicitation Number: HSSS01-14-Q-0182, https://www.fbo.gov/?id=8aaf9a50dd4558899b0df22abc31d30e

the aforementioned challenges.

Current research on sarcasm detection on Twitter (Tsur *et al.*, 2010; González-Ibáñez *et al.*, 2011; Liebrecht *et al.*, 2013; Riloff *et al.*, 2013) primarily analyze information obtained only from the text of tweets. These techniques treat sarcasm as a linguistic phenomenon, with limited emphasis on the psychological aspects of sarcasm. However, sarcasm has been extensively studied in the psychological and behavioral sciences and theories explaining when, why, and how sarcasm is expressed have been established. These theories can be extended and employed to automatically detect sarcasm on Twitter. For example, Rockwell (Rockwell, 2007) identified a positive correlation between cognitive complexity and the ability to produce sarcasm. A high cognitive complexity of an individual may be manifested in the language complexity of her tweets on Twitter.

We follow a systematic approach to sarcasm detection; we first theorize the core forms of sarcasm using existing psychological and behavioral studies. Next, we develop computational features to capture these forms of sarcasm using user's current and past tweets. Finally, we combine these features to train a learning algorithm to detect sarcasm. The major contributions of this thesis are:

1. We identify different forms of sarcasm and demonstrate how these forms may be manifested on Twitter.

2. We introduce behavioral modeling as a new, effective approach for detecting sarcasm on Twitter; we propose and evaluate the **SCUBA** framework — **S**arcasm **C**lassification **U**sing a **B**ehavioral modeling **A**pproach.

3. We investigate and demonstrate the importance of historical information discerned from past tweets for sarcasm detection.

In the next chapter, we review related sarcasm detection research. In Chapter 3, we formally define sarcasm detection in Twitter. Then, we discuss different forms of sarcasm and outline SCUBA, our behavior modeling framework for detecting sarcasm. In Chapter 4, we demonstrate how different forms of sarcasm can be identified within Twitter and construct features that model these forms. In Chapter 5, we describe in detail the data collection process, experiment set up and discuss baseline approaches used for comparison. Then, we perform extensive experiments and evaluate our framework. Chapter 6 concludes this thesis with discussions and directions for future work. In the appendix section, we provide a brief introduction of Twitter and twitter parlance crucial to the understanding of our framework.

Table 1.1: Examples of Misinterpreted Sarcastic Tweets.

Examples	Users	Tweets
1	User 1	you are doing great! Who could predict heavy travel between #Thanksgiving and #NewYearsEve. And bad cold weather in Dec! Crazy!
	Major U.S Airline	We #love the kind words! Thanks so much.
	User 1	wow, just wow, I guess I should have #sarcasm
2	User 2	Ahhh..**** reps. Just had a stellar experience w them at Westchester, NY last week. #CustomerSvcFail
	Major U.S Airline	Thanks for the shout-out Bonnie. We're happy to hear you had a #stellar experience flying with us. Have a great day.
	User 2	You misinterpreted my dripping sarcasm. My experience at Westchester was 1 of the worst I've had with ****. And there are many.

Chapter 2

RELATED WORK

Sarcasm has been widely studied by psychologists, behavioral scientists and linguists for many years. Theories explaining the cognitive processes behind sarcasm usage such as the echoic reminder theory (Kreuz and Glucksberg, 1989), allusional pretense theory (Kumon-Nakamura *et al.*, 1995) and implicit display theory (Utsumi, 2000) have been well researched and detailed.

The echoic reminder theory (Kreuz and Glucksberg, 1989) states that recognizing sarcasm depends on the listener's allusion of some previous state of affairs. Positive statements such as "you're an amazing friend" may be viewed as sarcastic without the need for explicit allusion (which may instead be implicit), comparing to the often unsaid but established, conventional norms and traditions. However, negative statements "you're a terrible friend" may require explicit antecedents to be understood. Positive sarcastic statements may not require explicit antecedents as most customs and conventions are generally positive whereas, negative sarcastic statements may need an explicit antecedent for better understanding. Importantly, the echoic reminder theory accounts for this asymmetry between positive and negative sarcastic statements. This work also identifies and discusses the motivations behind why sarcasm is used when the same situation may be expressed without sarcasm. One of the interesting motivations discussed is that using sarcasm to describe a certain situation not only gives an objective evaluation of the situation but also reflects the users attitude and perception towards the situation. For example, the positive sarcastic statement "the weather is so lovely", not only indicates that the weather is bad but also indicates the user's disdain for the weather, whereas, its equivalent statement

"the weather is bad" does not provide us an insight into the attitude of the speaker.

The allusional pretense theory (Kumon-Nakamura *et al.*, 1995) describes ironic situations as when the speaker strives to allude to the listener to a particular failed expectation. This is done by displaying faux sincerity, drawing attention to the failed expectation and to the speaker's viewpoint on the same. Two primary factors which are necessary to convey irony is discussed: [1] Allusion to differences between what is expected and what the reality actually is. [2] Pragmatic insincerity which may be conveyed by showing a semantic contrast, being overly polite, counterfactual, uninterested etc. However, the authors also concede that these two factors may not be sufficient conditions for irony and briefly touch upon other possible preconditions such as mutual knowledge. Mutual knowledge between participants in the conversation may be established by community membership, physical co-presence and linguistic co-presence. They also present a new motivation for using irony by viewing it as a tool to convey negative attitudes with humor and wit without directly confronting the subject (Jorgensen, 1996).

The implicit display theory (Utsumi, 2000) claims that an ironic utterance implicitly displays an *ironic environment* which the authors describe as having three important properties - expectation, incongruity, emotional attitude. Essentially, the theory states that the speaker has a set expectation which fails. This incongruity between what is expected and what is observed in reality results in the speaker having a negative emotion, leading to an ironic/sarcastic utterance. Utsumi proposed a simple computational framework to detect the degree of irony based on the degrees of [1] allusion, [2] pragmatic insincerity, [3] indirect expression of the negative attitude expressed through the utterance, [4] context-independent polarity and [5] manifestness of the expectations that create the ironic environment. However, the framework is quite theoretical and no experiments were performed to evaluate the computational

effectiveness of the framework.

Automatic detection of sarcasm is a relatively new, less researched topic and is deemed a difficult problem (Pang and Lee, 2008). While works on automatic detection of sarcasm in speech (Tepperman *et al.*, 2006) utilizes prosodic, spectral and contextual features, sarcasm detection in text has relied on identifying text patterns (Davidov *et al.*, 2010) and lexical features (González-Ibáñez *et al.*, 2011; Kreuz and Caucci, 2007).

Davidov et al. (Davidov *et al.*, 2010) devised a semi-supervised technique to detect sarcasm in Amazon product reviews and tweets. They used an interesting pattern-based (using high frequency words and content words) and punctuation-based features to build a classification model using a weighted k-nearest neighbor classifier to perform sarcasm detection. González-Ibánez et al. (González-Ibáñez *et al.*, 2011) devised a detection technique using numerous lexical features (derived from LWIC (Pennebaker *et al.*, 2001), Wordnet Affect (Strapparava and Valitutti, 2004)) and pragmatic features such as emoticons and replies. Reyes et al. (Reyes *et al.*, 2012) focussed on developing classifiers to detect verbal irony based on ambiguity, polarity, unexpectedness and emotional cues derived from text. Liebrecht et al. (Liebrecht *et al.*, 2013) used unigrams, bigrams and trigrams as features to detect sarcastic dutch tweets using a Balanced Winnow classifier. More recently, Riloff et al. (Riloff *et al.*, 2013), used a well constructed lexicon-based approach to detect sarcasm based on an assumption that sarcastic tweets are a contrast between a positive sentiment and a negative situation. Table 2.1 gives a brief overview of the aforementioned current research related to automatic sarcasm detection.

As described above, current works on sarcasm detection have heavily focussed on sarcasm's linguistic aspects and utilized primarily, the content of the tweet. In contrast, we believe that our framework provides a systematic approach towards better

Table 2.1: Overview of Related Work

Authors & Year	**Overview of methodology**
Riloff *et al.* (2013)	Lexicon-based approach contrasting positive sentiment and negative situation
Liebrecht *et al.* (2013)	Unigram, bigram and trigram features used to train a Balanced Winnow classifier
Reyes *et al.* (2012)	Ambiguity, polarity, emotional cues etc., to train decision trees
González-Ibáñez *et al.* (2011)	lexical and pragmatic features to train SMO classifier
Davidov *et al.* (2010)	Patterns and punctuations based features used to train weighted k-nearest neighbor classifier

sarcasm detection by not only analyzing the content of tweets but by also exploiting the behavioral traits of users derived from their past activities. Furthermore, the user's past activities also aid in incorporating contextual awareness to our behavior modeling framework to improve the classification process. Contextual awareness has been acknowledged within psychology research as being a necessary condition for identifying sarcasm (Capelli *et al.*, 1990; Woodland and Voyer, 2011).We map research on (1) what makes people use sarcasm, (2) when they use it and (3) how they use it, to observable user behavior on Twitter and build a comprehensive supervised framework to detect sarcasm. A somewhat similar behavior modeling approach has been used by Zafarani et al. (Zafarani and Liu, 2013; Zafarani *et al.*, 2014) to connect users accross social networks using minimum information.

Chapter 3

BEHAVIOR MODELING FRAMEWORK

Before describing our approach to detect sarcasm and detailing our behavior modeling framework, we formally state the problem at hand.

3.1 Problem Statement

Sarcasm, while quite similar to irony, differs in that it is usually viewed as being negative, caustic and derisive. Some researchers even consider it to be aggressive humor (Basavanna, 2000) and a form of verbal aggression (Toplak and Katz, 2000). While researchers in linguistics and psychology debate on what exactly constitutes sarcasm, for the sake of clarity, we use the Oxford dictionary's[1] definition of sarcasm as *a way of using words that are the opposite of what you mean in order to be unpleasant to somebody or to make fun of them* and formally define the sarcasm detection problem on Twitter as follows:

Definition of sarcasm detection on Twitter: *Given an unlabeled tweet* $\mathbf{t}$ *from user* $\mathbf{U}$ *along with a set of* $\mathbf{U}$*'s past tweets* $\mathbf{T}$*, a solution to sarcasm detection aims to automatically detect if* $\mathbf{t}$ *is sarcastic or not.*

In addition to following a behavior modeling approach, our problem is different from past research on sarcasm detection which use only text information from $\mathbf{t}$ and do not consider the user's past tweets $\boldsymbol{T}$ which are available in Twitter. This is a very important distinction as the usage of past tweets in our classification process helps put the tweets that we are examining into context. We made this conscious

[1] http://oald8.oxfordlearnersdictionaries.com/dictionary/sarcasm

decision of using past tweets based on the aforementioned psychological theories on sarcasm which unilaterally stress on past customs and expectations being factors behind generating and recognizing sarcasm.

3.2 Behavior Modeling Approach

In Twitter, tweets are not always created in isolation. When posting a sarcastic tweet, the user makes a conscious choice to express her thoughts through sarcasm. The user may decide to use sarcasm as a response to a certain situation, observation or emotion. This behavior is informed by the user's individual characteristics, moods etc., which may be observed and analyzed through her activities on Twitter.

Further, it is observed that some people have more difficulty in generating and recognizing sarcasm than others due to cultural differences, language barriers etc. Therefore, some individuals have a higher propensity to use sarcasm than others. Hence, we factor in the user's likelihood of being a *sarcastic person* or otherwise, by analyzing historical data in the form of the user's past tweets.

Using existing research on sarcasm and our observations on Twitter, we find that sarcasm generation can be characterized as one (or a combination) of the following:

Sarcasm as a contrast of sentiments

A popular perception of sarcasm among researchers is that sarcasm is a contrast of sentiments. A classical view of sarcasm, based on the traditional pragmatic model (Grice, 1975), argues that sarcastic utterances are first processed in the literal sense and if the literal sense is found incompatible with the present context, only then is the sentence processed in its opposite (ironic) form. This perceived contrast may be expressed through multiple facets such as mood, affect or sentiment.

Sarcasm as a complex form of expression

Rockwell (Rockwell, 2000) showed that there is a small but significant correlation between cognitive complexity and the ability to produce sarcasm. A high cognitive complexity involves understanding and taking into account, multiple perspectives to make cogent decisions. Furthermore, expressing sarcasm requires determining if the environment is suitable for sarcasm, creating an appropriate sarcastic phrase and assessing if the receiver would be capable of recognizing sarcasm. Therefore, sarcasm is a complex form of expression needing more effort than usual from the user (McDonald, 1999).

Sarcasm as a means of conveying emotion

Sarcasm is primarily a form of conveying one's emotions. While sarcasm is sometime interpreted as aggressive humor (Basavanna, 2000) and as form of verbal aggression (Toplak and Katz, 2000), it also functions as a tool of self expression. Past studies (Grice, 1978), recognize that sarcasm is usually expressed in situations with negative emotions and attitudes.

Sarcasm as a function of familiarity

Friends and relatives are found to be better at recognizing sarcasm than strangers (Rockwell, 2003). Further, it has been demonstrated that the familiarity of language (Cheang and Pell, 2011) and cultural factors (Rockwell and Theriot, 2001; Katz *et al.*, 2004) also play an important role in the recognition and usage of sarcasm.

Sarcasm as form of written expression

In psychology, sarcasm has been studied primarily as a spoken form of expression. However, sarcasm is quite prevalent in the written context as well,

especially with the advent of online social networking sites. Through time, users have become more adept at conveying sarcasm in writing by including subtle markers that indicate to the unassuming reader, that the phrase is sarcastic. For example, while "you're so smart" does not hint at sarcasm, "Woowwww you are SOOOO cool"[2] elicits some doubts on the statement's sincerity.

We believe that when expressing sarcasm, the user would invariably exhibit one or more of the aforementioned forms of sarcasm. Therefore, we build a behavior modeling framework for sarcasm detection that utilizes features which model these different forms. These extracted features are used to train a supervised classification model to determine if the tweet is sarcastic or not. As the novelty of approach lies in the behavior modeling and not the actual classifier itself, we explain more in detail on how sarcasm is modeled and incorporated into the framework. If the reader is unfamiliar with Twitter, a brief introduction of Twitter is included in the Appendix section. Readers who are well acquainted with Twitter are encouraged to proceed to the next chapter which describes the feature construction in detail.

[2]An original tweet collected.

Chapter 4

REPRESENTING FORMS OF SARCASM

Users' efforts in generating sarcasm are manifested in many ways on Twitter. In this section, we describe how different forms of sarcasm are realized in Twitter and how one can construct relevant features to capture these forms in the context of Twitter.

4.1 Sarcasm as a Contrast of Sentiments

4.1.1 Contrasting Connotations

A common means of expressing sarcasm is to employ words with contrasting connotations within the same tweet. For example, in *I love getting spam emails!*, *spam* has an obvious negative connotation while *love* is overwhelmingly positive. To model such occurrences, we construct features based on (1) affect and (2) sentiment scores. We obtain affect score of words from a dataset compiled by Warriner et al. (Warriner *et al.*, 2013). This dataset contains affect (valence) scores for 13,915 English lemmas which are on a 9-point scale, with 1 being the least pleasant.

The sentiment score is calculated using SentiStrength (Thelwall *et al.*, 2010). SentiStrength is a lexicon-based tool optimized for tweet sentiment detection based on sentiments of individual words in the tweet. Apart from providing a ternary sentiment result {positive, negative, neutral} for the whole tweet, SentiStrength outputs two scores for each tweet. A negative sentiment score from -1 to -5 (not-negative to extremely-negative) and a positive sentiment score from 1 to 5 (not-positive to extremely-positive). Here, we use SentiStrength's lexicon to obtain word-level sentiment scores. From these sentiment and affect scores, we calculate different scores as

follows:

$$A = \{\, \text{affect}(w) \mid w \,\epsilon\, t\} \tag{4.1}$$

$$S = \{\, \text{sentiment}w) \mid w \,\epsilon\, t\} \tag{4.2}$$

$$\Delta_{\text{affect}} = \max(A) - \min(A) \tag{4.3}$$

$$\Delta_{\text{sentiment}} = \max(S) - \min(S) \tag{4.4}$$

where t is the tweet and w is a word in t. The affect(w) outputs the affect score of w. The sentiment(w) outputs the sentiment score of w. Δ_{affect} and $\Delta_{\text{sentiment}}$ indicate the level of contrast in terms of sentiment and affect infused into the tweet by the user. We use Δ_{affect} and $\Delta_{\text{sentiment}}$ as features (2 features).

SentiStrength and the approach by Warriner et al. (Warriner *et al.*, 2013) provide sentiment and affect scores only for unigrams. However, there are many words which when viewed individually may not have sentiment value but when analyzed together may convey a positive or negative connotation. For example, "working on sundays" is conventionally viewed with disdain while the individual words themselves do not allude any emotion. Hence, we construct a lexicon of positive and negative sentiment bigrams and trigrams used on Twitter following an approach similar to Kouloumpis et al. (Kouloumpis *et al.*, 2011) as follows:

1. We collect about 400,000 tweets with positive sentiment hashtags such as #love, #happy, #amazing and 400,000 tweets with negative sentiment hashtags such as #sad, #depressed, #hate, among others.

2. From these tweets, we extracted bigrams and trigrams along with their respective frequencies. We filter out bigrams and trigrams with frequencies less than 10.

3. For each bigram or trigram b, we find its associated sentiment score S_b,

$$S_b = \frac{POS(b) - NEG(b)}{POS(b) + NEG(b)} \tag{4.5}$$

where $POS(b)$ is the number of occurrences of b in the positive tweets dataset and $NEG(b)$ is the number of occurrences of b in the negative tweets dataset. We filter out bigrams or trigrams with marginal sentiment scores $\in (-0.1, 0.1)$. This sentiment measure is similar to association scores produced by Liu et al. (Liu and Ruths, 2013)

Using the generated lexicon, we include as features, the number of bigrams and trigrams with positive sentiment scores, negative sentiment scores and their respective sum of scores (4 features).

4.1.2 Contrasting Present with the Past

While users often use contrasting words in the same tweet to express sarcasm, often times, a user may set up a contrasting context in her previous tweet and then, choose to use a sarcastic remark in her current tweet. This behavior may be more prevalent on Twitter as a result of the 140 character limit.

To model such behavior, we obtain the sentiment expressed by the user (i.e., positive, negative, neutral) in the previous tweet and the current tweet using SentiStrength. Then, we include the type of sentiment transition taking place from the past tweet to the current tweet (for example, *positive* $\rightarrow$ *negative*, *negative* $\rightarrow$ *positive*) as a feature (1 feature). In total, there are nine such transitions involving the combinations of positive, negative and neutral sentiments.

To provide a historical perspective on the user's likelihood for such sentiment transitions, we compute the probability for all nine transitions using the user's past

tweets. The transition probabilities along with the probability score of the current transition are included as features (10 features).

4.2 Sarcasm as a Complex Form of Expression

4.2.1 Readability

As sarcasm is widely acknowledged to be hard to read and understand, we adapt standardized readability tests to measure the degree of complexity and understandability of tweets. We use as features: number of words, number of syllables and number of syllables per word in the tweet derived from the Flesch-Kincaid Grade Level Formula (Flesch, 1948). We also include number of polysyllables[1] and the number of polysyllables per word in the tweet derived from SMOG grade for readability (McLaughlin, 1969) as features (5 features).

Inspired by the average word length feature used in the Automated Readability Index (Kincaid *et al.*, 1975), we formulate a more comprehensive set of features involving the word length distribution $L = \{l_i\}_{i=1}^{19}$ constructed from tweet t as follows:

1. For each word w in t, we compute its character length $|w|$. For convenience, we ignore words of length 20 or more. We construct a word length distribution $L = \{l_i\}_{i=1}^{19}$ for t, where l_i denotes the number of words in the tweet with character length i.

2. L may be represented succinctly using the following 6-tuple presentation:

$$< \mathbb{E}[l_w], med[l_w], mode[l_w], \sigma[l_w], \min_{w \in t} l_w, \max_{w \in t} l_w > \tag{4.6}$$

   where $\mathbb{E}$ is the mean, *med* is the median, *mode* is the mode and σ is the standard deviation of word length distribution L.

[1]Polysyllables are words containing three or more syllables.

We include the 6-tuple representation as features (6 features).

Further, given the availability of the user's past tweets, we examine if there is a noticeable difference in the word length distribution between the user's current tweet and her past tweets. It must be noted that while sarcastic tweets may also be present in the user's past tweets, because of their relative rarity, the past tweets when taken in entirety, would *average out* any influence possibly introduced by a few past sarcastic tweets. Therefore, any difference from the norm in the word length distribution of the current tweet can be captured. To capture differences in word length distribution, we perform the following steps:

1. From the user's current tweet, we construct a probability distribution D_1 over length of words in the tweet.

2. From the user's past tweets, we construct a probability distribution D_2 over length of words in all the past tweets.

3. To calculate the difference between the world length distribution of the current tweet and the past tweets, we calculate the Jenson-Shannon (JS) divergence between D_1 and D_2:

$$JS(D_1||D_2) = \frac{1}{2}KL(D_1||M) + \frac{1}{2}KL(D_2||M) \quad (4.7)$$

where $M = \frac{D_1+D_2}{2}$ and KL is the KL-divergence:

$$KL(T_1||T_2) = \sum_i \ln(\frac{T_1(i)}{T_2(i)})T_1(i)$$

We include the JS-divergence value also as a feature (1 feature).

4.3 Sarcasm as a Means of Conveying Emotion

4.3.1 Mood

Mood represents the user's state of emotion. Intuitively, the mood of the user may be indicative of her propensity to use sarcasm; if the user is in a bad (negative) mood, she may choose to express it in the form of a sarcastic tweet. Therefore, we gauge the user's mood using sentiment expressed in her past tweets. However, we cannot assume that the user's mood is encapsulated in her last n tweets. Therefore, we capture the mood using her past tweets as follows:

1. For each past tweet t, we compute its positive sentiment score, $\text{pos}(t)$ and its absolute negative sentiment score, $\text{neg}(t)$ using SentiStrength.

2. We divide the user's past tweets into overlapping buckets based on the number of tweets posted prior to the current tweet.

3. Each bucket b_n consists of the previous n tweets posted by the user. We select $n \in \{1, 2, 5, 10, 20, 40, 80\}$.

4. In each b_n, we capture the user's perceived mood using two tuples. The first tuple is:

$$< \overset{+}{\sum}, \overset{-}{\sum}, P, \max(\overset{+}{\sum}, \overset{-}{\sum}) >, \tag{4.8}$$

   where $\sum^+$ and $\sum^-$ are the total positive and negative sentiment scores of tweets in b_n:

$$\overset{+}{\sum} = \sum_{t \in b_n} \text{pos}(t), \tag{4.9}$$

$$\overset{-}{\sum} = \sum_{t \in b_n} \text{neg}(t), \tag{4.10}$$

$$P = \begin{cases} +, & \text{if } \sum^{+} \geq \sum^{-} \\ -, & \text{otherwise} \end{cases} \quad (4.11)$$

The second tuple is:

$$< n_+, n_-, n_0, n, Q, \max(n_+, n_-, n_0) > \quad (4.12)$$

where n_+ is the number of positive tweets, n_- is the number of negative tweets, n_0 is the number of neutral tweets present in b_n(found using SentiStrength). n is the total tweets present in b_n and Q indicates the majority sentiment of tweets, i.e., $Q \in \{+, -, 0\}$.

$$Q = \begin{cases} +, & \text{if } n_+ = max(n_+, n_-, n_0) \\ -, & \text{if } n_- = max(n_+, n_-, n_0) \\ 0, & \text{if } n_0 = max(n_+, n_-, n_0) \end{cases} \quad (4.13)$$

We include both tuples for each b_n as features ($7 \times 10 = 70$ features).

As one's mood remains constant for a limited amount of time, we also gauge the user's mood within a specific time window. However, again, we cannot assume that the user's mood is encapsulated within any t minutes. Therefore, we divide the user's past tweets into buckets b_t, which consists of all the tweets posted by the user within t minutes from the current tweet. Here, $t \in \{1, 2, 5, 10, 20, 60, 720, 1440\}$ minutes (1440 minutes = 1 day). For each bucket b_t, we include the tuples in (6.8) and (6.12) also as features ($8 \times 10 = 80$ features).

4.3.2 Affect and Sentiment

As sarcasm is a combination of affect and sentiment expression, we explore the possibility of observing differences with respect to how affect and sentiment is ex-

pressed in a sarcastic tweet. To this end, we construct a sentiment score distribution SS in which each count is the number of words in the tweet with sentiment score i where $i \in [-5, 5]$. We also construct an affect score distribution AS in which each count is the number of words in the tweet of affect score j where $j \in [1, 9]$. We normalize counts in SS and AS. We include as features both these distributions (20 features). Similar to (4.6), we represent these distributions as 6-tuples and include them as features (12 features). We also included the number of affect words, number of sentiment words and the sentiment expressed (positive, negative and neutral) which are obtained from SentiStrength as features (3 features).

To capture the difference in sentiment expression, we compare the sentiment score distribution of the user's past tweets to that of her current tweet. Following a procedure similar to (4.7), we calculate the JS-divergence between the past and current sentiment score distributions and include it as a feature (1 feature).

In order to gain insights into the range of sentiments expressed by the user to gauge how she uses Twitter as a tool to express emotion, we construct a normalized distribution over the sentiment score $[-5, 5]$ of each word of her past tweets and include the distribution as a feature (11 features). This distribution given a perception of how expressive the user is, on Twitter. This is crucial as different users use Twitter for different reasons. Some Twitter users tweet objective facts, news articles and are generally information while other users are quite informal and tweet personal issues, emotions, opinions etc.

4.3.3 Frustration

When individuals observe or experience an unjust situation, they sometimes turn to social media which act as effective outlets for their complaints and frustrations (Bi and Konstan, 2012). This frustration is often expressed in the form of sarcasm (Gibbs,

2000) (example, tweets in Table 1.1). Usually, sarcasm is not premeditated; it is a spontaneous reaction to certain unpleasant/disturbing events or scenarios. Therefore, quantifying the spontaneity of a tweet can provide insights into whether the tweet is sarcastic or not.

To model spontaneity, using the user's past tweets, we construct an expected tweet posting time probability distribution which describes the regular tweeting norms of the user. From each of the user's past tweets, we extract the tweet creation time, using which, we build a normalized 24 bin distribution TT (one for each hour). TT approximates the probability of the user tweeting at each hour. For each examined tweet, using the respective user's TT, we find the likelihood of a user posting the tweet at that hour. The lower the likelihood, the more divergent the tweet is from the user's usual tweeting patterns. Low likelihood scores indicate that the user is not expected to tweet at that particular time and that the user has gone out of her way to tweet at that time, therefore, in some sense, the tweet is spontaneous in nature. We include as a feature, actual likelihood score of the user tweeting at that particular hour (1 feature).

We also observe that users tend to post successive tweets in short quick bursts when they vent out their frustrations, therefore, we include as a feature, the time difference between the examined tweet and the previous tweet posted by the user (1 feature). Another common way to express frustration is through the usage of swear words. Using Wang et al's (Wang *et al.*, 2014) compilation of most common swear words, we check for the presence of such words in the tweet and include their presence as a boolean feature (1 feature).

4.4 Sarcasm as a Function of Familiarity

4.4.1 Familiarity of Language

Intuitively, one would expect a user who uses a form of language as complex as sarcasm to have good command over the language. Therefore, we obtain a profile of the user's language skills by measuring features inspired from standardized language proficiency cloze tests. As part of the cloze test (Oller, 1972), proficiency is evaluated based on vocabulary, grammar, dictation and reading levels. As dictation and reading levels pertain to the oratory and reading skills of the user which cannot be measured from written text, we concentrate our efforts on constructing features that best represent vocabulary and grammar skills.

Using past tweets from the user, we determine the size of her vocabulary. We include as features, the total words, total distinct words used and the ratio of distinct words to total words used, to measure the user's redundancy in word usage (3 features).

Grammar skills are measured in terms of the usage of different parts-of-speech(POS). The POS tags for words in the tweet are generated using TweetNLP's (Owoputi *et al.*, 2013) POS tagger.The tags produced may be *interjections*, *emoticons*, etc. The complete list of 25 POS tags is provided in Owoputi et al. (Owoputi *et al.*, 2013). We obtain the POS tag for every word in the tweet and build a corresponding normalized POS distribution and include it as features (25 features).

Oftentimes, location is a major confounding factor in how language is spoken. Further, it has been shown that people in different regions perceive and use sarcasm differently. For example, comparing northerners and southerners in the U.S, Dress et al. (Dress *et al.*, 2008) showed that the northerners formulate more sarcastic sentences compared to the southerners. Therefore, we try to infer the approximate location of

the user. However, as the location field in Twitter is a *free-text* field in which any text may be inputted, it is often noisy. Therefore, we approximate the user's location with her time zone and include it as a feature (1 feature).

We also include as a feature, the number of past occurrences of the #sarcasm and #not hashtags (1 feature). This feature indicates if the user is familiar with sarcasm as a form of expression.

4.4.2 Familiarity of Environment

Generally, users express sarcasm better when they are well acquainted with the environment. Just as people are less likely to use sarcasm at a new, unfamiliar setting, we believe that users would take some time to get themselves acclimatized with Twitter before they post sarcastic tweets. Therefore, we measure familiarity in terms of the number of tweets posted, number of days since the user created her Twitter profile (twitter age), number of tweets divided by the user's twitter age and use them as features (3 features). These features give an indication of the duration for which the user has been using Twitter.

We also measure familiarity in terms of the user's frequency of Twitter usage with respect to time. From the user's past tweets, we calculate the time intervals between each pair of successive tweets. We represent these times as a 6-tuple, similar to (4.6) and include them as features (6 features). To capture how active the user is on Twitter and her familiarity with *twitter parlance*, we include as features, the number of retweets, mentions and hashtags used in her past tweets (3 features). We also quantify the user's familiarity with Twitter by identifying how embedded she is in Twitter's social graph by including as features, the number of friends and followers (2 features). To adjust for longevity, we divide the number of friends and followers by the user's Twitter age and include the same also as features (2 features).

Most regular and experienced twitter users often use shortened words (by removing vowels, using numbers etc.) to circumvent the 140 character limit. Therefore, we include as features, the presence of alphanumeric words (boolean), presence of words without vowels (boolean) as well as the percentage of dictionary words present in the tweet (3 features).

4.5 Sarcasm as a Form of Written Expression

While low pitch, high intensity and a slow tempo (Rockwell, 2000) are vocal indicators of sarcasm, users attempting to express sarcasm in writing are devoid of such devices. Therefore, users may be forced to innovate and use certain styles of writing to compensate for the lack of visual and verbal cues. We categorize variations stemming from such behavior as either (i) prosodic or (ii) structural.

4.5.1 Prosodic Variations

Prosody has been studied and identified as one of the major cues of sarcasm (Wang *et al.*, 2006; Nakassis and Snedeker, 2002; Woodland and Voyer, 2011; Capelli *et al.*, 1990). Prosodic variations refer to changes made to writing styles in order to express intonation and stress. Language in social media is continuously evolving as users find simple, yet effective ways to better express themselves within the constraints imposed by the social networking site.

Users often repeat letters in words to stress and over-emphasize certain parts of the tweet (for example, *sooooo*, *awesomeeee*) to indicate that they mean the opposite of what is written. We capture such usage by including as boolean features, the presence of repeated characters (3 or more) and the presence of repeated characters (3 or more) in sentiment-loaded words (such as, loveeee) (2 features). We also include as features, the number of characters used, and the ratio of the number of distinct

characters to the total characters used in the tweet (2 features).

We also observe that users often capitalize certain words to emphasize changes in tone (if the tweet were to be read out loud). We account for such changes by including as features, number of capitalized words in the tweet (1 feature). It is also commonly observed that some users capitalize certain parts-of-speech(POS) to exaggerate or to vent their frustration. Using TweetNLP, we obtain the POS tag for each capitalized word in the tweet. Then, we compute the probability of observing such tags and include the same as features (25 features).

Furthermore, users also use certain punctuations to express non-verbal cues that are crucial for sarcasm deliverance in speech. For example, users use "*" to indicate emphasis, "..." to indicate pause, "!!!" for exclamations (sometimes overdone to indicate sarcasm). Therefore, we include as features, the normalized distribution of common punctuation marks(.,!?'*") (7 features). To compare the user's current usage of punctuations to her past usage, similar to (4.6), we calculate the JS-divergence measure between the current and past punctuation distribution, and include the same as a feature (1 feature). This comparison puts the punctuation usage into perspective, taking into account users who may have a tendency to use a disproportionate number of punctuations in their everyday tweets.

4.5.2 Structural Variations

Structural variations are inadvertent variations in the POS composition of tweets to express sarcasm. We observe that sarcastic tweets sometimes have a certain structure wherein the user's views are expressed the first few words of the tweet, while in the later parts, a description of a particular scenario is put forth (for example, I love it when my friends ignore me). To capture possible syntactic idiosyncrasies arising from such tweet construction, we use as features, the POS tags of the first three words

and the last three words in the tweet (6 features). We also include the position of the first sentiment-loaded word (0 if not present) and the first affect-loaded word (0 if not present) as a feature (2 features).

Given the structure followed in constructing sarcastic tweets, we also check for positional variations in the hashtags present in the tweet. We trisect the tweet based on the number of words present and include as features, the number of hashtags present in the each of the three parts of the tweet (3 features).

To capture differences in syntactic structures, we examine the parts of speech sequence present in the tweet. Similar to (4.6), we construct a probability distribution over the POS-tagged current tweet as well as POS-tagged past tweets and include as a feature, its Jenson-Shannon divergence measure (1 feature).

Existing works on quantifying linguistic style (Hu *et al.*, 2013) use lexical density, intensifiers and personal pronouns as important measures to gauge the writing style of the user. Lexical density is the fraction of information carrying words present in the tweet (nouns, verbs, adjectives and adverbs). Intensifiers are words that maximize the effect of adverbs or adjectives (for example, so, very). Personal pronouns are pronouns denoting a person or group (for example, me, our, her). We include as features the lexical density, the number of intensifiers used and the number of first-person singular, first-person plural, second-person and third-person pronouns present in the text (6 features).

In total we construct 327 features based on the behavioral aspects of sarcasm. Figure 4.1, gives an overview of the features constructed.

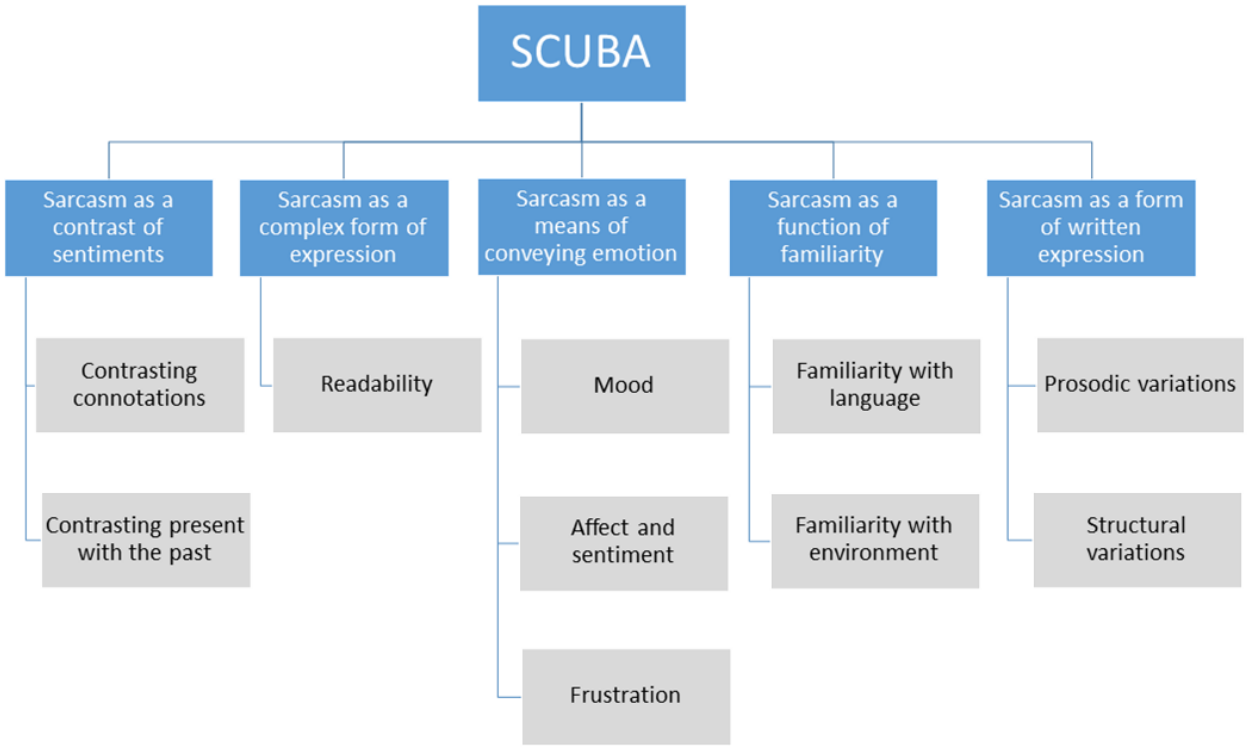

Figure 4.1: Overview of Features Constructed

Chapter 5

EXPERIMENTS AND EVALUATION

5.1 Data Collection

We validate our framework using a dataset[1] of tweets from Twitter. To obtain a set of sarcastic tweets, we query the Streaming API using keywords #sarcasm and #not filtering out non-english tweets and retweets. We also remove tweets containing mentions and URLs as obtaining information from media and URLs is computationally expensive. We limit our analysis to tweets which contain more than three words as we found that tweets with fewer words were very noisy or clichéd (e.g., *yeah, right! #sarcasm*). Davidov et al. (Davidov *et al.*, 2010) noted that some tweets containing the #sarcasm hashtag were *about* sarcasm and that the tweets themselves were not sarcastic. To limit such occurrences, we include only tweets that have either of the two hashtags as its last word; this reduces the chance of obtaining tweets that are about sarcasm but are themselves not sarcastic. After preprocessing, we obtained about 9104 sarcastic tweets which were self described by the user as being sarcastic using the appropriate hashtags. We remove the #sarcasm and #not hashtags from the tweets before proceeding with the evaluation.

In order to collect a set of general tweets (not sarcastic), we used Twitter's Sample API which provides a random sampling of tweets. We remove tweets that contain #sarcasm or #not from this random sample. It is true that this random sample may yet contain tweets that are sarcastic (but without the sarcasm hashtags) and fully acknowledge that the random dataset collected may not be pure. However, we believe

[1]The dataset can be obtained by contacting the author

that the possible proportion of sarcastic tweets in the random sample is extremely low and that when these tweets are taken in entirety, its effect would be miniscule. These tweets were subjected to the same aforementioned preprocessing technique.

Finally, for each tweet in the collected dataset, we extract the user who posted the tweet and then, we obtained that user's past tweets (we obtain the past 80 tweets for each user).

Some examples of tweets in the dataset are:

1. *This paper is coming along... #not*

2. *Finding out your friends' lives through tweets is really the greatest feeling. #sarcasm*

The above examples illustrate the difficulty of the task at hand. The first tweet may or may not be sarcastic purely depending on the context (which is not available in the tweet). Even if some background is available to us, as in the case of the second tweet, clearly, it is still a complicated task to map that information to sarcasm.

It must also be noted that, to avoid confusion and ambiguity when expressing sarcasm in writing, the users choose to explicitly mark the sarcastic tweets with appropriate hashtags. The expectation is that these tweets, if devoid of these hashtags, might be difficult to comprehend as sarcasm, even for humans. Therefore, our dataset might be biased towards the hardest forms of sarcasm. Using this dataset, we evaluate our framework and compare it with existing baselines.

5.2 Experiment Setup

As seen in the previous section, we have labeled data in the form of tweets with and without the sarcasm hashtags. Using this labeled data, we model our sarcasm detection problem as a supervised classification problem. A schematic diagram of the

experiment set up is given in figure 5.1.

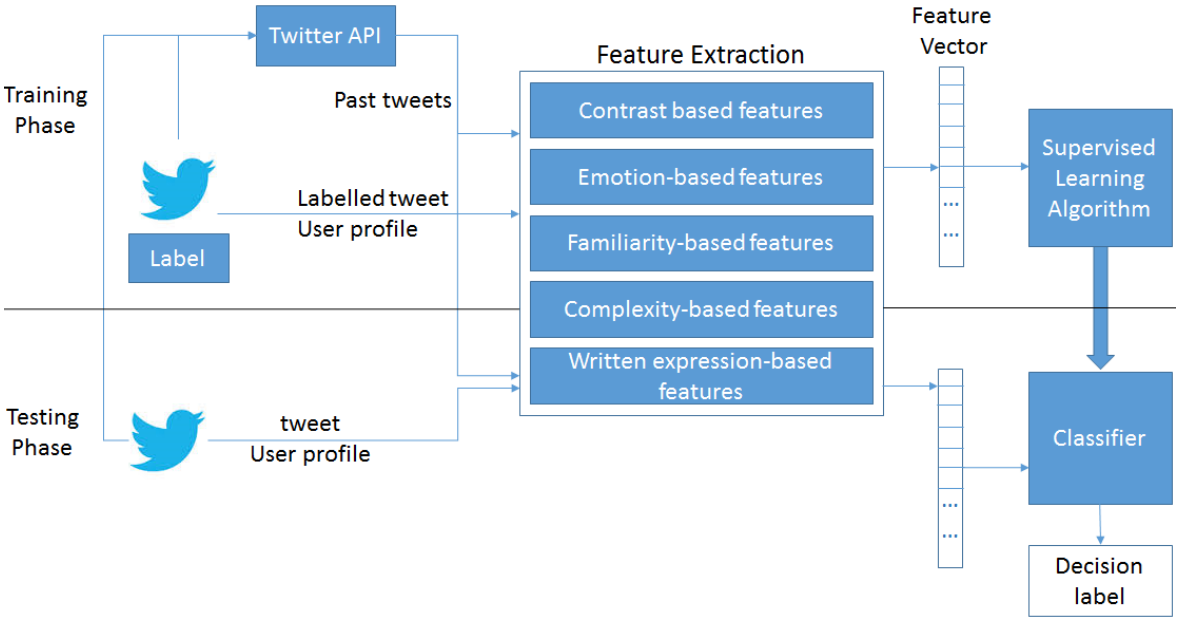

Figure 5.1: SCUBA's Ssupervised Learning Framework

Training Phase

In the training phase, as described earlier, we have access to the labelled tweets. Each tweet is obtained in JSON format and contains numerous tweet-based fields such as text, hashtags, tweet creation time etc., and user profile-based fields such as account creation time, number of past statuses, friends, followers etc. A sample tweet JSON object showcasing the list of raw fields available to us is shown in figure 5.2. The user object which is embedded in the tweet object is shown in figure 5.3 for want of space.

For each tweet in the dataset, we identify the user who posted the tweet and then, use Twitter's API to obtain past tweets from that user. Each of the past tweet is also in the JSON format shown in figure 5.1.

```
{
  "contributors": null,
  "truncated": false,
  "text": "Today's been a lovely day! #sarcasm",
  "in_reply_to_status_id": null,
  "id": 42578925834023,
  "favorite_count": 0,
  "source": "web",
  "retweeted": false,
  "coordinates": null,
  "entities": {
    "symbols": [],
    "user_mentions": [],
    "hashtags": [
      {"indices": [27,35],
        "text": "sarcasm"
      }
    ],
    "urls": []
  },
  "in_reply_to_screen_name": null,
  "id_str": "42578925540357734",
  "retweet_count": 0,
  "in_reply_to_user_id": null,
  "favorited": false,
  "user": see user object figure
  "geo": null,
  "in_reply_to_user_id_str": null,
  "lang": "en",
  "created_at": "Wed Jan 22 00:37:23 +0000 2014",
  "filter_level": "medium",
  "in_reply_to_status_id_str": null,
  "place": null
}
```


Figure 5.2: Tweet Object

Using the tweets and past tweets, we construct relevant features as described in the previous chapter. This feature extraction process produces feature vectors used to train a supervised learning algorithm which produces a classifier model classifying tweets as sarcastic or otherwise.

```
"user": {
  "follow_request_sent": null,
  "profile_use_background_image": true,
  "default_profile_image": false,
  "id":  82394239,
  "verified": false,
  "profile_image_url_https": "https://pbs.twimg.com/9a8d95573635f.png",
  "profile_sidebar_fill_color": "DDFFCC",
  "profile_text_color": "333333",
  "followers_count": 388,
  "profile_sidebar_border_color": "FFFFFF",
  "id_str": " 82394239",
  "profile_background_color": "9AE4E8",
  "listed_count": 3,
  "profile_background_image_url_https": "https://pbs.twimg.com/9a8d95573635f.png",
  "utc_offset": -21600,
  "statuses_count": 4322,
  "description": "I have a really bad YouTube channel. Please watch my sh*t.",
  "friends_count": 183,
  "location": "Canada",
  "profile_link_color": "0084B4",
  "profile_image_url": "https://pbs.twimg.com/9a8d95573635f.png",
  "following": null,
  "geo_enabled": false,
  "profile_banner_url": "https://pbs.twimg.com/9a8d95573635f.png",
  "profile_background_image_url": "https://pbs.twimg.com/9a8d95573635f.png",
  "name": "Robert",
  "lang": "en",
  "profile_background_tile": true,
  "favourites_count": 44,
  "screen_name": "932so3er",
  "notifications": null,
  "url": "http://www.youtube.com/932so3er",
  "created_at": "Sat Nov 08 15:52:00 +0000 2008",
  "contributors_enabled": false,
  "time_zone": "Central  Time (US & Canada)",
  "protected": false,
  "default_profile": false,
  "is_translator": false
},
```


Figure 5.3: User Object

Testing Phase

In the testing phase, again, we query the Twitter API to obtain past tweets and perform the feature extraction process. The feature vector outputted by the extraction

process is then taken as input for the classifier. The classifier returns whether the inputted tweet is sarcastic or not.

5.3 Selecting Suitable Learning Algorithm

Before evaluation, we must choose a suitable supervised classifier for SCUBA. Given the large number of features constructed, some of which may not be good predictors, we need to ensure that the chosen classifier identifies a good subset of features from all the features constructed and learns the classification task without overfitting. It appears that ℓ_1 regularization is a good candidate for this scenario as ℓ_1 encourages sparse representation and performs implicit feature selection by driving some feature weights to zero. This type of regularization not only ensures that the feature weights are low, hence preventing over-fitting but also, encourages sparse representation which allows for easy storage and fast computation.

To put our theory to test, we evaluate SCUBA using multiple learning algorithms[2] including J.48 decision tree, ℓ_1-regularized logistic regression and ℓ_1-regularized ℓ_2-loss SVM to obtain an accuracy of 78.06%, 83.46% and 83.05% respectively on the collected dataset (with class distribution 1:1). Clearly, ℓ_1 regularization appears beneficial with superior performances. We choose the ℓ_1-regularized logistic regression version of our framework for comparison with the baselines.

5.4 Baselines

We compare our framework against a state-of-the-art lexicon-based technique by Riloff et al. (Riloff *et al.*, 2013). The basic premise of their method is that sarcasm can be viewed as a contrast between a positive sentiment and a negative situation. They constructed three phrase lists (positive verb phrases, positive predicative ex-

[2]We use Weka, LIBLINEAR and Scikit-learn library

pressions and negative situations) from 175,000 tweets using a parts-of-speech aware bootstrapping technique extracting relevant phrases. Different combinations of these phrase lists were used to decide if a tweet is sarcastic or not. Using these phrase lists, we re-implement two of their most successful approaches:

1. **Contrast Approach**, which marks a tweet as sarcastic if it contains a positive verb phrase or positive predicative expression along with a negative situation phrase.

2. **Hybrid Approach**, which marks a tweet as sarcastic if the tweet was marked sarcastic either by bootstrapped-lexicon approach or by a bag-of-words classifier trained on unigrams, bigrams and trigrams.

In order to provide a comparable framework to the Hybrid Approach, we embed as a feature, results from the aforementioned bag of words classifier into SCUBA as well. We call our n-gram augmented framework, SCUBA++.

To quell doubts that SCUBA merely labels all tweets from users who have previously used #sarcasm or #not as sarcastic, we completely remove that particular feature and perform the same classification task. We also include as baselines, a random classifier which classifies the tweets randomly into sarcastic and non-sarcastic, a majority classifier which classifies all tweet into the majority class (obtained with knowledge of the class distribution) and the aforementioned n-grams model.

5.5 Evaluation Metrics

Naturally, the class distribution over tweets is skewed towards the non-sarcastic tweets. Similar to previous works (Liebrecht *et al.*, 2013), we evaluate the SCUBA framework using different class distributions (1:1, 10:90, 20:80, where 1:1 means for every sarcastic tweet in the dataset, we introduce 1 tweet that is not sarcastic.). We

Table 5.1: Performance Evaluation using 10-Fold Cross-validation

Techniques	Dataset distributions					
	1:1		20:80		10:90	
	Acc.	AUC	Acc.	AUC	Acc.	AUC
SCUBA	83.46	0.83	88.10	0.76	92.24	0.60
Contrast Approach	56.50	0.56	78.98	0.57	86.59	0.57
SCUBA++	86.08	0.86	89.81	0.80	92.94	0.70
Hybrid Approach	77.26	0.77	78.40	0.75	83.87	0.67
SCUBA - #sarcasm	83.41	0.83	87.53	0.74	91.87	0.63
N-grams	78.56	0.78	81.63	0.76	87.89	0.65
Majority classifier	50.00	0.50	80.00	0.50	90.00	0.50
Random classifier	49.17	0.50	50.41	0.50	49.78	0.50

include AUC (Area under the ROC Curve) apart from accuracy as a performance measure as AUC is robust to class imbalances (Fawcett, 2006). This analysis gives an insight into how well SCUBA performs under varied distributions. We use the standard 10-fold cross validation technique to evaluate the performance, the results of which are given in Table 5.1.

From the results, we observe that SCUBA++ clearly outperforms all other techniques for every class distribution on both performance measures. It is interesting to note that only SCUBA and SCUBA++ perform better than the majority classifier for highly skewed distributions (90:10). We also observe that while the Hybrid Approach performs much better than the Contrast Approach, it is still not very effective for skewed distributions. Also, we notice that when the past sarcasm feature is removed from SCUBA, we obtain similar performance measures showing the minimal effect

of using this feature on the framework performance. Both random classifier and the majority classifier obtain an AUC score of 0.50, which is the minimum possible AUC score attainable.

5.6 Contrasting SCUBA with Contrast and Hybrid Approaches

A possible reason for the Hybrid and Contrast approach underperforming is that these approaches operate with the assumption that a sarcastic tweet contains a positive and negative sentiment phrase. However, this appears to be a very simplistic assumption and may not always hold true in real world settings. For example, the tweet, "Linkedin: 'Sean, you're getting noticed.' Ooh, do tell #sarcasm" is sarcastic, yet there are no positive or negative sentiments. However, SCUBA makes no such assumptions about sentiments.

Furthermore, SCUBA takes into account the user's past activities on Twitter which help provide contextual awareness aiding in better decision making. It is important to note that SCUBA's context awareness is with respect to user's emotion, mood, characteristics etc., deciphered from her past activities on Twitter. However, the baseline approaches do not consider the possibility of using historical information to better their classification. Also, SCUBA takes a behavior modeling approach, focusing on the psychological and behavioral aspects to sarcasm while the Contrast and Hybrid approaches view sarcasm from purely a linguistic perspective. This highlights the core differences in the approaches taken and may shed light on why SCUBA performs better than the baseline approaches.

5.7 Feature Set Analysis

In the previous section, we noted that SCUBA and SCUBA++ performs well even in skewed distributions. However, to gain insights into which specific types of features

have good predictive power, we perform the following feature set analysis. We divide the list of features used into sets depending on the different forms of sarcasm from which they were derived - features based on complexity, based on contrast, based on expression of emotion, based on familiarity and based on expression in written form. This analysis allows us to make an informed decision about which feature sets to consider if we are computationally constrained. It also provides an insight into the type of sarcasm that is prevalent in Twitter. A low predictive power for a feature set may indicate fewer instances of sarcastic tweets originating from that form of sarcasm. However, this analysis must be taken with a grain of salt as the predictive power is also a function of how well the features model the observations.

Table 5.2 shows the performance of SCUBA using each of the feature sets individually. While all feature sets may contribute to SCUBA's performance, they do so unequally. Clearly, all feature sets perform much better than random (50%). This further shows the need to view sarcasm through its varied facets and not trivialize it to a particular form of expression (such as contrast seeking).

It is interesting to note that complexity and familiarity based features, which are quite unique to the behavior modeling approach adopted, perform very well compared to some of the other, more intuitive feature sets. Written expression-based features perform the best among all feature sets which is not surprising as injecting prosodic and structural variations to account for the lack of verbal cues have become more common, especially with users becoming more experienced using social media. However, it is important to note that the feature sets themselves are not completely independent, for example, some features constructed from the contrast aspect of sarcasm may be incorporated into the emotion expression aspect and vice-versa.

Table 5.2: Feature Set Analysis

Features	accuracy
All feature sets	83.46
complexity-based features	73.00
contrast-based features	57.34
emotion expression-based features	71.52
familiarity-based features	73.67
written expression-based features	76.72

5.8 Feature Importance Analysis

As observed, different feature sets have different effects on the performance. To gain deeper insights into which specific features are most important for detecting sarcasm, we perform feature ranking analysis. While we may use many features to detect sarcasm, clearly, some features may be more important than others. Therefore, we perform a thorough analysis of features to determine the set of features that contribute most to detecting sarcasm. We use the odds-ratio (coefficients from ℓ_1-regularized logistic regression) for the importance analysis.

As described earlier, ℓ_1 regularization performs implicit feature selection and hence some of the feature weight values are zero. Given below are the top 10 features in decreasing order of importance:

1. Percentage of emoticons in the tweet. (-)
2. Percentage of adjectives in the tweet. (+)
3. Percentage of past words with sentiment score 3. (+)

4. Number of polysyllables per word in the tweet (-)
5. Lexical density of the tweet. (-)
6. Percentage of past words with sentiment score 2. (+)
7. Percentage of past words with sentiment score -3. (+)
8. Number of past sarcastic tweets posted. (+)
9. Percentage of positive to negative sentiment transitions made by the user. (+)
10. Percentage of capitalized hashtags in the tweet. (-)

We observe that features derived from all forms of sarcasm: text expression-based features (1, 2, 5, 10), emotion-based features (3, 6, 7), familiarity based features (8), contrast-based features (9) and complexity-based features (4) rank high in discriminative power.

Interestingly, the most important feature is the percentage of emoticons present in the tweet. It appears to be negatively correlated with sarcasm. This may be because sarcasm is inherently ambiguous and needs to be that way to produce the intended effect. However, when users use emoticons, the tweets, in some sense, lose ambiguity as emoticons are a very obvious form of emotion expression. Therefore, users may refrain from using emoticons in their sarcastic tweets.

Another interesting observation is that five of the top ten features (3, 5 ,6, 7, 8) are context-based features which have been derived from the past activities of the user. This further highlights the importance of using past information in the decision making process.

5.9 Evaluating Effectiveness of Historical Information

In SCUBA, we have included the user's historical information on Twitter in the form of past tweets to detect sarcasm. However, it might be computationally expensive to process and use all the past tweets for classification. Furthermore, it would be imprudent of us to assume that Twitter would continue to provide access to so many past tweets for each user. Therefore, it is imperative that we identify the optimum number of past tweets to be used in detecting sarcasm. To do this, we measure the gain in performance by executing the sarcasm classification multiple times while varying the number of past tweets available to us.

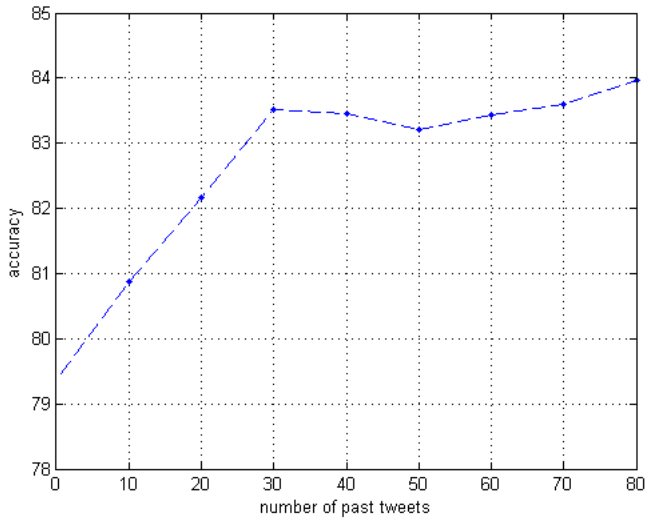

Figure 5.4: Effect of Historical Information on Performance

Figure 5.1 shows the performance obtained by varying past tweets (smoothened using a moving-average model). We observe that without using any historical information, we obtain an accuracy of 79.38%, which is still better than all the baselines. Interestingly, using only the user's past 30 tweets, we obtain a considerable gain (+4.14%) in performance. However, adding even more tweets does not significantly

improve the performance. Therefore, if computationally constrained, we can use only the past 30 tweets and expect a comparable performance. This result is of high significance as it makes SCUBA feasible to be used in real world, real-time environments by establishing a reasonable bound on the amount of computational power required.

Chapter 6

DISCUSSIONS AND FUTURE WORK

In this work, we propose SCUBA, a behavior modeling framework for sarcasm detection. With SCUBA, we take a systematic approach outlined as follows:

1. We identify and discuss the different forms of sarcasm: (1) as a contrast of sentiments, (2) as a complex form of expression, (3) as a means of conveying emotion, (4) as a function of familiarity and (5) as a form of written expression.
2. We construct relevant features to identify these forms on Twitter.
3. We train a supervised learning algorithm using the constructed features to detect sarcastic tweets.
4. Through multiple experiments, we demonstrate that SCUBA is effective in detecting sarcastic tweets.

Unlike current approaches to sarcasm detection, SCUBA takes a holistic view, taking into account not just the actual tweet content but also, the user's overall propensity to use sarcasm. SCUBA distinguishes itself from current approaches in the following ways.

1. SCUBA is the first behavior modeling framework proposed for sarcasm detection. It models sarcasm taking into account not only the linguistic aspects but also the psychological and behavioral aspects of sarcasm.
2. SCUBA is the only sarcasm detection approach which makes uses of historical information which help provide context to the tweet.

3. Importantly, we have demonstrated that even using limited amount of historical information may greatly aid in improving the efficiency of the classification process. The resilience of SCUBA's performance to limited information makes it a good fit for real world, real-time applications which may have higher computational constraints.

It is important to note that while we perform our evaluation and experiments on a Twitter dataset, SCUBA can be generalized to other social networking sites. It can be easily expanded by including other features specific to the target social networking site. This further widens the scope of applicability of SCUBA to different social networking sites.

With nearly all major service oriented companies having a social media presence to provide consumer assistance, SCUBA can co-exist with existing sentiment analysis technologies to better serve the needs of the company's social media team. With consumer assistance teams aiming for a zero-waiting time response to customer queries through social media, undetected sarcasm can amount to embarrassing gaffes and potential PR disasters. Using SCUBA, social media teams can better detect sarcasm and deliver appropriate responses to sarcastic tweets.

As automatic sarcasm detection research is still in its infancy, there are numerous extensions to our approach that may be evaluated. One of the more interesting avenues for future research is to identify how a user's social network and her past interactions affect sarcasm generation. This comports well with existing research (Rockwell, 2003) which suggests that users are more likely to use sarcasm with friends than with strangers. Furthermore, the strength of ties in social networks may also be quantified and leveraged to identify sarcasm directed at specific individuals. Currently, SCUBA does not consider sarcasm directed at specific individuals. This research gap may be explored to identify social factors that influence sarcasm usage.

Further, having observed the advantages of using a behavior modeling approach to detect sarcasm, we wish to apply the same to detect other non-literal forms of language such as humor. The behavior modeling aspects may complement the existing linguistic research to provide improved performance on such difficult tasks.

This thesis focusses on Twitter as a platform for sarcasm detection experiments. However, this approach may be applied to other social networking sites such as Facebook as the core structure of these sites are very similar. SCUBA primarily relies of the user's tweet, profile and past tweets in Twitter for classification. This information is not unique to Twitter and most social networking sites have similar information available which may be leveraged for sarcasm detection.

REFERENCES

Basavanna, M., *Dictionary of psychology* (Allied Publishers, 2000).

Bi, F. and J. A. Konstan, "Customer service 2.0: Where social computing meets customer relations.", IEEE Computer **45**, 11, 93–95 (2012).

Capelli, C. A., N. Nakagawa and C. M. Madden, "How children understand sarcasm: The role of context and intonation", Child Development **61**, 6, 1824–1841 (1990).

Cheang, H. S. and M. D. Pell, "Recognizing sarcasm without language: A cross-linguistic study of english and cantonese.", Pragmatics & Cognition **19**, 2 (2011).

Davidov, D., O. Tsur and A. Rappoport, "Semi-supervised recognition of sarcastic sentences in twitter and amazon", in "Proceedings of the Fourteenth Conference on Computational Natural Language Learning", pp. 107–116 (Association for Computational Linguistics, 2010).

Dress, M. L., R. J. Kreuz, K. E. Link and G. M. Caucci, "Regional variation in the use of sarcasm", Journal of Language and Social Psychology **27**, 1, 71–85 (2008).

Fawcett, T., "An introduction to {ROC} analysis", Pattern Recognition Letters **27**, 8, 861 – 874, URL `http://www.sciencedirect.com/science/article/pii/S016786550500303X`, {ROC} Analysis in Pattern Recognition (2006).

Flesch, R., "A new readability yardstick.", Journal of applied psychology **32**, 3, 221 (1948).

Gibbs, R. W., "Irony in talk among friends", Metaphor and symbol **15**, 1-2, 5–27 (2000).

González-Ibáñez, R., S. Muresan and N. Wacholder, "Identifying sarcasm in twitter: A closer look.", in "ACL (Short Papers)", pp. 581–586 (Citeseer, 2011).

Grice, H. P., "Logic and conversation", in "Syntax and semantics", edited by P. Cole and J. L. Morgan, vol. 3 (New York: Academic Press, 1975).

Grice, H. P., "Some further notes on logic and conversation", in "Syntax and Semantics 9: Pragmatics", edited by P. Cole, pp. 113–127 (1978).

Hu, Y., K. Talamadupula, S. Kambhampati *et al.*, "Dude, srsly?: The surprisingly formal nature of twitter's language.", in "ICWSM", (2013).

Jorgensen, J., "The functions of sarcastic irony in speech", Journal of Pragmatics **26**, 5, 613–634 (1996).

Katz, A. N., D. G. Blasko and V. A. Kazmerski, "Saying what you don't mean social influences on sarcastic language processing", Current Directions in Psychological Science **13**, 5, 186–189 (2004).

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel", Tech. rep., DTIC Document (1975).

Kouloumpis, E., T. Wilson and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!", ICWSM **11**, 538–541 (2011).

Kreuz, R. J. and G. M. Caucci, "Lexical influences on the perception of sarcasm", in "Proceedings of the Workshop on computational approaches to Figurative Language", pp. 1–4 (Association for Computational Linguistics, 2007).

Kreuz, R. J. and S. Glucksberg, "How to be sarcastic: The echoic reminder theory of verbal irony.", Journal of Experimental Psychology: General **118**, 4, 374 (1989).

Kumar, S., F. Morstatter and H. Liu, *Twitter Data Analytics* (Springer, 2014).

Kumon-Nakamura, S., S. Glucksberg and M. Brown, "How about another piece of pie: The allusional pretense theory of discourse irony.", Journal of Experimental Psychology: General **124**, 1, 3 (1995).

Kwak, H., C. Lee, H. Park and S. Moon, "What is twitter, a social network or a news media?", in "Proceedings of the 19th international conference on World wide web", pp. 591–600 (ACM, 2010).

Liebrecht, C., F. Kunneman and A. van den Bosch, "The perfect solution for detecting sarcasm in tweets #not", WASSA 2013 p. 29 (2013).

Liu, W. and D. Ruths, "What's in a name? using first names as features for gender inference in twitter", in "Analyzing Microtext: 2013 AAAI Spring Symposium", (2013).

McDonald, S., "Exploring the process of inference generation in sarcasm: A review of normal and clinical studies", Brain and Language **68**, 3, 486 – 506, URL `http://www.sciencedirect.com/science/article/pii/S0093934X99921247` (1999).

McLaughlin, G. H., "Smog grading: A new readability formula", Journal of reading **12**, 8, 639–646 (1969).

Nakassis, C. and J. Snedeker, "Beyond sarcasm: Intonation and context as relational cues in childrens recognition of irony", in "Proceedings of the Twenty-sixth Boston University Conference on Language Development [en línea]. Disponible en: http://www. wjh. harvard. edu/˜ lds/pdfs/Nakassis& Snedeker_2002. pdf[Links]", (2002).

Oller, J., John W., "Scoring methods and difficulty levels for cloze tests of proficiency in english as a second language", The Modern Language Journal **56**, 3, pp. 151–158, URL `http://www.jstor.org/stable/324037` (1972).

Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters", in "Proceedings of NAACL-HLT", pp. 380–390 (2013).

Pang, B. and L. Lee, "Opinion mining and sentiment analysis", Foundations and trends in information retrieval **2**, 1-2, 1–135 (2008).

Pennebaker, J. W., M. E. Francis and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001", Mahway: Lawrence Erlbaum Associates p. 71 (2001).

Reyes, A., P. Rosso and D. Buscaldi, "From humor recognition to irony detection: The figurative language of social media", Data & Knowledge Engineering **74**, 1–12 (2012).

Riloff, E., A. Qadir, P. Surve and Silva, "Sarcasm as contrast between a positive sentiment and negative situation.", in "EMNLP", pp. 704–714 (ACL, 2013), URL `http://dblp.uni-trier.de/db/conf/emnlp/emnlp2013.html#RiloffQSSGH13`.

Rockwell, P., "Lower, slower, louder: Vocal cues of sarcasm", Journal of Psycholinguistic Research **29**, 5, 483–495 (2000).

Rockwell, P., "Empathy and the expression and recognition of sarcasm by close relations or strangers", Perceptual and motor skills **97**, 1, 251–256 (2003).

Rockwell, P., "The effects of cognitive complexity and communication apprehension on the expression and recognition of sarcasm", Hauppauge, NY: Nova Science Publishers (2007).

Rockwell, P. and E. M. Theriot, "Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis.", Communication Research Reports **18**, 1, 44 – 52, URL `http://search.ebscohost.com.ezproxy1.lib.asu.edu/login.aspx?direct=true&db=ufh&AN=9471018&site=ehost-live` (2001).

Strapparava, C. and A. Valitutti, "Wordnet affect: an affective extension of wordnet.", in "LREC", vol. 4, pp. 1083–1086 (2004).

Tepperman, J., D. R. Traum and S. Narayanan, "" yeah right": sarcasm recognition for spoken dialogue systems.", in "INTERSPEECH", (2006).

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text", Journal of the American Society for Information Science and Technology **61**, 12, 2544–2558 (2010).

Toplak, M. and A. N. Katz, "On the uses of sarcastic irony", Journal of Pragmatics **32**, 10, 1467–1488 (2000).

Tsur, O., D. Davidov and A. Rappoport, "Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews.", in "ICWSM", (2010).

Utsumi, A., "Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony", Journal of Pragmatics **32**, 12, 1777–1806 (2000).

Wang, A. T., S. S. Lee, M. Sigman and M. Dapretto, "Neural basis of irony comprehension in children with autism: the role of prosody and context", Brain **129**, 4, 932–943 (2006).

Wang, W., L. Chen, K. Thirunarayan and A. P. Sheth, "Cursing in english on twitter", in "Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing", pp. 415–425 (ACM, 2014).

Warriner, A. B., V. Kuperman and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas", Behavior research methods pp. 1–17 (2013).

Whiting, A. and D. Williams, "Why people use social media: a uses and gratifications approach", Qualitative Market Research: An International Journal **16**, 4, 362–369, URL `http://www.emeraldinsight.com/doi/abs/10.1108/QMR-06-2013-0041` (2013).

Woodland, J. and D. Voyer, "Context and intonation in the perception of sarcasm", Metaphor and Symbol **26**, 3, 227–239 (2011).

Zafarani, R., M. A. Abbasi and H. Liu, *Social Media Mining: An Introduction* (Cambridge University Press, 2014).

Zafarani, R. and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach", in "Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 41–49 (ACM, 2013).

APPENDIX A

TWITTER FUNDAMENTALS

Twitter is an online social networking site where users may connect with each other and post messages called "tweets". Each tweet may have a maximum of 140 characters. In Twitter, the social network is directed, that is, a user may connect to other users by simply following them. Any user may follow any other user (unless the account is protected) without explicit consent. The users following a user are called followers, while users that are being followed by a user are called friends in Twitter parlance.

Each user has a "timeline" which contains tweets from accounts that the user follows in reverse chronological order. It is important to note that Twitter, unlike Facebook, does not filter or algorithmically curate timelines. A user may share tweets from other users with their followers by retweeting. This feature is similar to sharing posts on Facebook. Similar to the "Like" feature in Facebook, users may "favorite" a tweet. A user may also reply to tweets and mention other users using the "@" symbol.

An interesting feature which was quite unique to Twitter but is now prevalent throughout the online social networking sphere is the usage of hashtags. Hashtags, though quite common in Internet Relay Channels, were not initially part of Twitter's design. They were conceived by Twitter users as a way to group tweets and users on topics of interest. Nowadays, hashtags are ubiquitous and function as a simple mechanism to converse with specific groups, search for specific topics, make tweets more visible etc. More recently, hashtagged tweets also function as cheap, readily available labelled data for supervised learning algorithms.

A more detailed introduction to Twitter is available here[1]. The book "Twitter Data Analytics" (Kumar *et al.*, 2014) is an excellent resource detailing how to obtain data from Twitter using their APIs. Shown in figure 5.1, a sample tweet for illustration purposes.

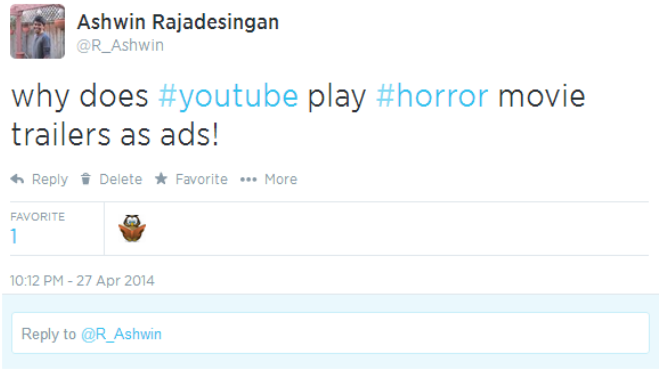

Figure A.1: A Typical Tweet

[1]https://support.twitter.com/articles/215585-getting-started-with-twitter