Summer 2019

A Multimodal Approach to Sarcasm Detection on Social Media

Dipto Das
Missouri State University, Dipto175@live.missouristate.edu

A MULTIMODAL APPROACH TO SARCASM DETECTION ON SOCIAL MEDIA

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Computer Science

By

Dipto Das

August 2019

A MULTIMODAL APPROACH TO SARCASM DETECTION ON SOCIAL MEDIA

Computer Science

Missouri State University, August 2019

Master of Science

Dipto Das

ABSTRACT

In recent times, a major share of human communication takes place online. The main reason being the ease of communication on social networking sites (SNSs). Due to the variety and large number of users, SNSs have drawn the attention of the computer science (CS) community, particularly the affective computing (also known as emotional AI), information retrieval, natural language processing, and data mining groups. Researchers are trying to make computers understand the nuances of human communication including sentiment and sarcasm. Emotion or sentiment detection requires more insights about the communication than it does for factual information retrieval. Sarcasm detection is particularly more difficult than categorizing sentiment. Because, in sarcasm, the intended meaning of the expression by the user is opposite to the literal meaning. Because of its complex nature, it is often difficult even for human to detect sarcasm without proper context. However, people on social media succeed in detecting sarcasm despite interacting with strangers across the world. That motivates us to investigate the human process of detecting sarcasm on social media where abundant context information is often unavailable and the group of users communicating with each other are rarely well-acquainted. We have conducted a qualitative study to examine the patterns of users conveying sarcasm on social media. Whereas most sarcasm detection systems deal in word-by-word basis to accomplish their goal, we focused on the holistic sentiment conveyed by the post. We argue that utilization of word-level information will limit the systems performance to the domain of the dataset used to train the system and might not perform well for non-English language. As an endeavor to make our system less dependent on text data, we proposed a multimodal approach for sarcasm detection. We showed the applicability of images and reaction emoticons as other sources of hints about the sentiment of the post. Our research showed the superior results from a multimodal approach when compared to a unimodal approach. Multimodal sarcasm detection systems, as the one presented in this research, with the inclusion of more modes or sources of data might lead to a better sarcasm detection model.

KEYWORDS: social media, qualitative study, image, text, multimodality, sarcasm detection, attention model

A MULTIMODAL APPROACH TO SARCASM DETECTION ON SOCIAL MEDIA

By

Dipto Das

A Master's Thesis
Submitted to The Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Computer Science

August 2019

Approved:

Anthony J. Clark, Ph.D., Thesis Committee Chair

Jamil M. Saquer, Ph.D., Committee Member

Lloyd A. Smith, Ph.D., Committee Member

Julie Masterson, Ph.D., Dean of the Graduate College

ACKNOWLEDGEMENTS

I express my gratitude to the Almighty God for granting me the ability that let me complete my thesis. Without His Grace, it could never be possible to complete this study.

I am thankful to the people in the Department of Computer Science at Missouri State University, especially my thesis committee members Dr. Anthony J Clark, Dr. Jamil M Saquer, Dr. Lloyd A Smith. A special thanks goes to my thesis supervisor Dr. Anthony J. Clark for his guidance, insight, and motivation. His unwavering motivation kept me engaged in this research work. His personal generosity made my time during research enjoyable and stress-free. From the very beginning with agreeing to the thesis topic to the end of my master's thesis research, his kindness and enthusiasm helped me.

I am grateful to my family members: my mother, Dipa Das and my sister, Dipty Das for their constant motivation and support. I am also grateful to my father, late Nirmal Das, because for his dream was one of the things that motivated me to dream of pursuing a degree abroad. I dedicate this thesis to my family members, whose blessing and support helped me throughout the writing of the thesis and life in general. The value of that increases to me with time.

I thank the participants in my research and friends who helped me recruit those helpful persons. I also thank all the wonderful people I met at Missouri State University.

I received numerous help from Md. Forhad Hossain, Dr. Razib Iqbal along with all of Bangladeshi community in Springfield, Missouri over the duration of my master's studies at Missouri State University. I am also thankful the friends I met here in the United States who made Springfield feel like second home to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Social networking sites (SNSs) have large numbers of users all over the world. Some platforms have a specific genre, that means, users discuss and interact with regard to some particular topics, whereas there are some platforms that are more generalized in users' interests. Over the years, social media platforms have strived to incorporate new features as their user-bases grew. Thus, social media's text-based interaction became more multimodal with the inclusion of images, videos, etc. As the number of users on these platforms grew, the interaction became more versatile due to different cultural backgrounds of the people. While at the beginning of SNSs, most users were relatively young, nowadays people of different generations are joining these platforms. This diverse nature of user interaction data on social media platforms attracted Big Data researchers to retrieve information, recognize interaction pattern, and analyze sentiments on these platforms. While sentiment analysis as a field of computer science is fairly developed, in most cases, sentiment analysis is treated as a binary classification problem. Obviously, human sentiment is so complex that it can be helpful to treat in more than just positive and negative categories. Recently, researchers are working on dividing positive sentiments in more fine categories like happiness, surprise, etc. as well as negative sentiments in more fine categories like sadness and anger. Sarcasm as a form of sentiment on the other hand has gained a lot less attention. Existing literature admits that sarcasm detection is more complex than fine-tuning emotion identification. The need for context information makes it more difficult to detect. Thus, failing to identify sarcasm as a part of user interaction on social media can mislead users about other users' thoughts and may initiate misunderstanding and "internet debate." We propose that multimodality can be a source of context information while communicating with diverse people. We also show how a system can be developed to automatically detect sarcasm on social media and propose some design recommendations.

Existing literature in computer science for sarcasm detection can be attributed mostly as

text mining. They are often based on existing literature in linguistics or real-life views of the researchers of sarcasm usage. Linguistics shed light on sarcasm from three points of view. Firstly, sarcasm can be divided into two emotions as a sentiment analysis problem – surface emotion and intended emotion. Here, surface emotion means the sentiment conveyed by the literal meaning of what the person says. And intended emotion means what the person tried to imply and expected the audience to infer. Some studies suggest that in sarcasm, the surface emotion of statement will be positive whereas the intended emotion will be negative. However, some studies argue with such a generalization. They advocate that while conveying sarcasm, the surface emotion and the intended emotion of the statement will be opposite and this argument is congruent to views of the first group of researchers. Secondly, sarcasm, as a part of communication violates the principles of Grice's maxims of cooperative dialogue, namely (1) the maxim of quantity, (2) the maxim of quality, (3) the maxim of relation, and (4) the maxim of manner. While the maxim of quality requires one to be truthful in a dialogue by not giving any misinformation, sarcasm is attributed as a misstatement about the emotion. Again, the maxim of manner mandates that one needs to be concise, orderly, and clear in what one says, sarcasm by its nature creates ambiguity about it. Thirdly, linguists suggest that sarcasm is often accompanied by some cues in popular patterns. For example, they suggest the variation in speech rate and amplitude, non-verbal cues like air quotes are strong indicators of sarcasm. Computer science researchers first recognized sarcasm detection as a research problem in 2006. Tepperman et. al. [4] utilized the third point from linguistics stated above in their study to detect sarcasm. They utilized the phrase "yeah right" as a notation for sarcasm. Many later studies utilized these views from linguistics to infer context, identifying patterns like capitalized texts, quotation marks, etc., detecting opposite sentiments in different parts of larger statements, and so on as different approaches to detect sarcasm. However, these linguistic studies emphasize in-person communication and thus, neither focus on users' interaction on SNSs nor are limited by the nature of interaction on those platforms. As a result, computer scientists mostly exploit the verbal aspects of sarcasm expressed with text, and have

not been utilizing the non-verbal aspects or cues of sarcasm. We argue that mutlimodal data can provide important information in this regard.

First of all, we want to bridge the gap of research in linguistics and computer science for in-person communication and communication over social media platforms, respectively. For that we conducted a qualitative study with participants from one English speaking and one non-English language speaking group. We asked the users when and how they conveyed sarcasm on social media. This helped us understand how they express the verbal and non-verbal cues of sarcasm (e.g., air quotes, variation of speech rate, etc.) within the limitations imposed by the platform's main mode of communication (e.g. text, emoticons, etc.). Our qualitative study suggests the pattern of cues of sarcasm on social media. Some of the themes raised by our participants were plausible with the patterns used by the previous literature (e.g. capitalization, quotation, etc.). However, some other new themes emerged from our observation. These themes emphasized the importance of multimodality in sarcasm detection. We propose that visual cues and visual contents of the images can contain important indications of an image being sarcastic. In some cases, visual cues in an image are enough for an individual to know if the post is sarcastic, for instance, in cases of **memes**. Again, opposing sentiments conveyed by the visual contents of the images and the text captions of the posts might be a form of multimodal sarcastic post. That means, different sentiments in different modes might indicate sarcasm. Besides investigating the format or structure of sarcastic posts on SNSs, we studied the bidirectional dynamics of how sarcasm impacts the popularity of a post in general and how peer sentiment influences users' sarcasm usage patterns. This later investigation allowed us to suggest design implications as part of this research.

For conducting the qualitative study, we interviewed interested participants. We analyzed the interview data with grounded theory approach. After finding the themes, we approached development of a sarcasm detection system from the point of machine learning model. Therefore, we needed dataset for training the models. We found there was a scarcity of labeled datasets for sarcasm detection. Since sarcasm is a highly subjective concept and only the person making the

statement is certain of whether the post is sarcastic or not, we collected self-annotated posts for our dataset. Posts were labeled as sarcastic by the person who posted that content. We collected data from popular social media platforms – Facebook, Twitter, and Flickr. We also utilized some existing datasets [5]. Aside from dataset collection, we also developed the systems model architectures and trained those with the collected dataset. We also evaluated the performance of transfer learning for many existing popular neural network architecture. We presented how the performance of the system improves with the inclusion of multimodality into the system.

Our system shows the superiority of a multimodal approach to a unimodal approach with text data only. Our system is not limited to any particular phrase (e.g. "yeah right") or language specific formatting (e.g. capitalization of words in English) as the system depends not only on text data, but relies on the holistic structure of the social media post. We analyze the relation of sarcasm use with the users' experience on these platforms to suggest future designs. The remainder of this thesis is organized as follows: chapter 2 discusses the related works in the existing literature; chapter 3 gives brief general overviews of the concepts or the building blocks of the system; chapter 4 focuses on the conducted qualitative study and analyzes the data for discovering the themes; the later three chapters discuss how gradual inclusion of modalities impact the system behavior – chapter 5 and chapter 6 discuss only text and only image based approaches for sarcasm detection, respectively, and chapter 7 highlights how multimodal approach to sarcasm detection with text, images, and reaction emoticons improve the system accuracy; the next chapter evaluates the qualitative findings and the machine learning model together for shading light on designs of SNSs, discusses possible threats to the validity of this study, and the future research directions. Finally, we draw the concluding remarks.

2 LITERATURE REVIEW

Existing literature in the field of sarcasm detection comes from several disciplines, including linguistics, psychology, social science, and more recently, computer science. Though, they differ in their goals. For example, studies from psychology and social sciences focus on the "why" and "when" questions–they ask when and why do people use sarcasm, researchers from linguistics and computer science disciplines focus predominantly on the "how" question. Specifically they investigate how do people convey sarcasm and how can it be recognized. Although researchers from computer science and linguistics align on their question, they differ in their objective. Whereas linguists do not typically concern themselves with automatically detecting sarcasm, computer science researchers focus on developing algorithms for detecting sarcasm as well besides understanding the computational model or nature of sarcasm. In this chapter, we will discuss studies from several disciplines, however, we will focus on the works that tried to address the questions like: "How do people convey sarcasm?" and "How to detect sarcasm?"

2.1 Sarcasm Constructs

Most works that study the constructs of sarcasm are from linguistics, psychology, and cognitive science. Gibbs et al. [6] conducted experiments with 256 undergraduate students, where they showed how non-literal interpretations of sarcastic statements are processed by humans before the literal meaning. They said that when a sarcastic statement is made in an in-person conversation, and the audience have access to non-verbal cues besides the verbal statements, the audience translate the statements into the corresponding intended meaning, i.e., non-literal meaning before translating the statements into their surface/literal meaning. They also discussed how sarcasm impacts how long the participants of a conversation remember a particular statement. They highlight the ease of processing and memory for sarcastic utterances. In a collection of several empirical and theoretical works, Gibbs et al. [7] discuss the theory of irony, especially comprehension of sarcasm in verbal form, social contexts, and functions of irony.

Sarcasm detection as a field of computer science can be placed under the field of sentiment analysis, which first drew the attention of computer science researchers in 2006. Tepperman et al. [4] developed the first work in computer science that recognized the problem of sarcasm detection from the perspective of computer science. They experimented with sarcasm recognition using cues like contextual (e.g., acknowledgement, agreement/disagreement), prosodic, and spectral features (e.g. pitch, energy, duration of each word). Given the limited capability of natural language processing at that time, they proposed a very naïve approach of detecting sarcasm from text data. They emphasized on the nature of sarcasm of being associated with several commonly used phrases. In their work, they only searched for the phrase "yeah right" as an indicator of sarcasm.

Several studies have invested effort to define what it means to be "sarcasm". Gonzalez-Ibanez et al. [8] identified the opposite nature of literal and intended meaning of micro-blog posts as sarcasm. According to them, sarcasm is different from positive or negative statements made on social media. It conveys negative sentiment while the literal meaning (also termed as surface sentiment) of the statement is positive and likewise, conveys positive intended sentiment with apparently negative surface meaning. That means, the study by [8] argues that sarcasm has one intended and one surface sentiment that have opposite polarity, i.e., positive surface meaning with negative intended meaning, and vice-versa. For example, in a statement like: *"Thank you for ruining my day."*, the phrase "thank you" is used with criticizing intention (i.e., negative intended meaning), whereas the phrase itself literally expresses gratitude (i.e., positive surface meaning). However, several other studies do not agree with them in this regard. Filatove et al. [9] argue that sarcasm always has positive literal meaning with a negative intended meaning. They also present observations of sarcasm having clear victims in micro-blogging platforms including social media, blogging sites, etc. They discussed sarcasm and irony replaceably in their work. Kreuz et al. [10] from a linguistic perspective agree with the argument of Filatova et al. [9] on sarcasm having always positive literal meaning with negative intended meaning and clear victim deeming the opposite very unlikely.

Clift et al. [11] explained sarcasm as a phenomenon of divergence between the spoken words and their intended meaning with the Traditional Oppositional Model (TOM). However, this model was criticized for ignoring the requirement of these two aspects of meaning happening at the same time. Sperber et al. [12] suggested that audience just process the intended meaning of sarcasm in a model named "Echoic/Interpretation Model". Later building on this model, the "Echoic Reminder Model" was proposed and reemphasized by Kreuz et al. [10] and Colston et al. [13] discussed the role of generally expected situation or social norms. Instead Kumon-Nakamura et al. [14] suggested sarcasm is achieved by mentioning part of expected situation that has occured while some other part was violated. Later Colston et al. [15] in their book, discussed how verbal sarcasm can be viewed as violation of expectation, and the pragmatically insincere or contrary relationship between literal and intended meaning of statements. This is echoed in the studies by [8–10] where we can see sarcasm as violations of Grice's maxims [16] as suggested by studies like [17–19].

Bamman et al. [20] gave importance to context information for the task of sarcasm detection. They tried to capture extra-linguistic information from the context of an utterance of sarcasm on Twitter. According to them, inclusion of properties of author, audience, and the immediate communicative environment can contribute to the sarcasm detection task. Their argument also situates itself in a line with linguistic study by Utsumi et al. [21] who discuss the comprehension of verbal irony for in-person conversational settings. The role of context can also be explained with the expectation of certain social norms as in [10, 13, 15], and thus reestablishes the incident of violating Grice's maxims [16] as we discussed in the previous chapter.

2.2 Sarcasm Dataset Collection

As we all understand, and as the existing literature suggest, context is important for detecting sarcasm [20], [22]. Because without proper context, a single sarcastic post can be treated as a non-sarcastic post and vice-versa. However, it is difficult to understand the context of a so-

cial media content, specially for the ones that come from unknown users on SNS that is common scenario for public posts.

2.2.1 Independently Annotated Datasets. For many data based analysis, having a large collection of annotated data is very helpful. There are different approaches of aggregating such datasets. Whereas for some cases, annotation of data is readily available, for some other tasks, researchers have to annotate the data themselves. The later case is usually adopted when the labels of data is more subjective in nature. It is a very common practice in text mining and computational linguistics community. Because of having greater control over the annotation and data, researchers can use their best judgments. In order to reduce bias, sometimes the annotations are done also by independent annotators other than the researchers. Since sarcasm data is quite subjective in nature, i.e., it is highly dependent on human perception whether a particular statement or post is sarcastic or not, several works in sarcasm detection prepared the datasets with independent annotations.

Swanson et al. [23] utilized crowdsourcing for the task of labeling a sarcasm detection dataset. They reported high reliability among the labels from untrained annotators on Mechanical Turk using common statistical popularity measurements, like Kappa, EM, majority class, etc.

Golbeck et al. [5] in one of their recent works, collected data from different websites that they classified as either satire or fake news. They collected at most five posts from one website to reduce bias to the way of writing on a particular website or any set of particular topics. They handpicked the data and labeled them manually trough discussion among the authors. Being the first dataset of this kind, this work had a small dataset size, providing with baseline measures for future research works.

2.2.2 User Annotated Datasets. The intention of considering context of the post leads us to wonder who has the access to the full context of any particular post. To address this concern, we argue that the user who posts a content on social media has full access to the context information about that particular post. That means, he/she who posts a certain content on social media is likely to be the original creator of that, have full understanding of the context of

the content, and thus, knows the intent of the caption and the hashtags used in the contents well. To denote a post as sarcastic, there are common trends on different social media platforms, e.g. hashtags like #sarcasm on twitter, ending statements with "/s" on reddit, and so on. These declarations of a particular tweet or post to be sarcastic assigned by the user himself/herself are called self-annotations.

According to Gonzalez-Ibanez et al. [8], sarcasm is a positive/negative utterance that transforms the polarity to the opposite of apparent sentiment. They created a large corpus with messages that the message writer himself/herself identified as sarcastic. They compared the sarcasm uttering tweets with those that convey positive or negative without sarcasm. They also reported the impact of lexical and pragmatic factors on machine learning effectiveness for identifying sarcasm in tweets. They also conducted a post-experiment user study that perhaps unsurprisingly, showed none of the machine learning models or human participants perform very well for detecting sarcasm.

Riloff et al. [24] termed the contrasting positive and negative sentiments as parts of same statement as sarcasm. They collected twitter data with #sarcasm hashtags assigned by the users for positive instances of data, and a collection of random data with a hope that most of the later data will not be sarcastic.

Reyes et al. [25] views the increasing use of irony or sarcasm plausible with the process of the online platforms being more social. They exploited ironic tweets from two perspectives: representativeness and relevance. They used user-generated hashtags (e.g. #irony) as labels in their dataset. They constructed a model of irony detection. Assessment of their initial results were largely positive.

The largest collection of a sarcasm dataset was done by Khodak et al. [26]. They collected 1.3 million sarcastic statements from Reddit that were self-annotated – annotated by the author of the statement himself/herself rather than being annotated by an independent annotator. Besides preparing the dataset, they evaluated the corpus for accuracy using three metrics of interest: (1) size, (2) the proportion of sarcastic to non-sarcastic comments, and (3) the rate of false

positives and false negatives. Their work also provided the field with benchmarks for sarcasm detection, and evaluations of baseline methods. Unlike the other dataset collection studies [8], [25], no model learning approach followed their data collection process, rather they highlighted the collection and evaluation of a large dataset as their main contribution.

2.3 Current Sarcasm Detection Methods

2.3.1 Unimodal Approaches. As Filatova et al. [9] and Riloff et al. [24] suggested, sarcasm as the presence of contrasting sentiments – positive and negative in different parts of a single tweet as an indicator of sarcasm. They showed that identifying contrasting contexts using the phrases learned through their proposed bootstrapping algorithm obtained high recall to detect sarcasm, i.e., they could identify most of the positive instances of sarcasm in the dataset.

Many studies later utilized this idea of contrasting sentiments indicating sarcasm. One of the first work that utilized this aforementioned idea was done by Cliche et al. [27]. Following the work done by [9, 24], this work also resorts to the definition of "sarcasm" by Merriam-Webster Dictionary [28] like some other works in this area [29]. This work utilizes data collected from twitter to extract features like n-grams (precisely uni-grams, and bi-grams), sentiment polarities, and topics. They propose a logistic regression and a support vector machine (SVM) based supervised classification algorithm to detect sarcasm. SVM performed better in their experiment. Peng et al. [30] builds upon this work and analyzed the strengths and weaknesses of that considering Cliche et al. [27] as a baseline model.

Drawing light on context information, Bamman et al. [20] attempted to model context information computationally. They proposed using four kinds of features: Tweet features, Author features, Audience features, and Response features. They used binary logistic regression with L2 regularization using ten-fold cross-validation for the sarcasm detection task.

Ghosh et al. [31] proposed another machine learning based approach with manually extracted features. They used a sample of data from the dataset collected by Khodak et al. [26]. However, for the highly skewed nature of the portion of data that they used, they proposed an

SVM based classifier. They used the binary representation for a certain word from the dictionary in a sarcastic statement.

Ghosh et al. [32] also focused on the semantic representation. However, unlike Ghosh et al. [31] that used representation at word level, Ghosh et al. [32] used it at sentence level to get access to more context knowledge. Instead of manual approach of assigning the representations, they used a neural network for this task. Their proposed architecture consisted of a convolutional neural network (CNN), followed by a long short term memory (LSTM), and finally a deep neural network (DNN). Whereas the prior works depended on a predefined set of indicative hashtags, they extended this list of hashtags by using a Latent Semantic Analysis (LSA) based approach.

2.3.2 Multimodal Approaches. Multimodality in sarcasm detection research is comparatively a new idea. Until recently, all the works in this area used only textual features of a content on social media. The idea behind adopting multimodality is that as users of social media platforms have been going beyond text-only to multimodalities (e.g., text, image, audio, video, etc.), when we are considering only text data from the plethora of multimodal data available on SNS, we are throwing out useful information that could help get crucial context information. Hence, some recent works emphasize the idea of utilizing multimodality in this area.

Schifanella et al. [33] is the first work to advocate for multimodality in sarcasm detection studies. The investigated the relationship between textual and visual aspects in multimodal posts. They ran a crowdsourcing task in which they asked users of the website CrowdFlower.com to quantify the extent to which images are perceived as necessary by human annotators. The users of this platform showed positive results for combining modalities to detect sarcasm across various platforms and methods and evaluated the impact of visuals as a source of context very highly. A survey paper by Razali et al. [29] re-emphasized the importance of such multimodal approach.

2.4 Identified Research Gap

Works from times before the rise of social media, most of the qualitative studies on the construct of sarcasm focused on verbal form of it [10–13]. However, little is known how the ways

of expressing sarcasm get changed for limited yet versatile ways of expressing oneself on social media. For example, the non-verbal cues as described in [34] are difficult to express on social media. Existing literature lacks insights about the ways how these non-verbal cues get transformed according to the expressive capability of the SNS platforms.

Again, we feel that the use of topics in text data as in [5, 9, 24] might limit the application of those approaches to a particular domain of text data generated over a certain period of time. Therefore, it is more important to know the "ways" in which sarcasm gets expressed in text data, i.e., how sentiment in overall fluctuates in a sarcastic piece of text data. Thus, knowing the general nature of variation of sentiments in text will be more applicable to text data from more diverse domain, time periods, or languages.

On top of that, the need for multimodality in sarcasm detection being raised during recent times [29, 33], we want to investigate this approach in more details. We think collection of multimodal datasets (i.e. datasets that will have more than text based data), studying role of modes of multimodal data individually, and the gradual performance change of sarcasm detection process with inclusion of more modes can be gaps in this regard.

Thus, in the following chapters, we presented a qualitative study focused on users' sarcasm expression ways on social media platforms, sentiment/emotion based analysis of textual sarcasm data, multimodal approach of sarcasm detection process, and evaluate the process and its performance.

3 HOW HUMANS DETECT SARCASM

The basic idea behind machine learning based systems, or artificial intelligence in general, is mimicking how humans operate. This is particularly evident for our problem, sarcasm detection on social networking sites (SNSs). Therefore, before proceeding to build a system that can detect sarcasm on SNSs, we attempt to understand how humans do the same. Many studies propose systems based on personal experience and word-level definition of "sarcasm" [29, 33]. However, we feel the need of a qualitative study to find out more general themes that are usual with users to detect and express sarcasm on SNSs to build a more effective sarcasm detection model.

3.1 Background

3.1.1 Convenience Sampling. Convenience sampling is a non-probabilistic sampling that aims to contact subjects that are close at hand. The two criteria applicable to this sampling technique are: subjects being easily available and willing to participate. This sampling does not guarantee that a random sample is generated. This approach is often avoided due to its proneness to sampling error. For example, this sampling might recruit only people who share the same beliefs and values as the researcher. This might lead to confirmation bias. However, it is still widely used because it makes data collection easier and more cost effective.

3.1.2 Purposive Sampling. Purposive sampling is another non-probabilistic sampling. It is also known as selective, subjective, or judgemental sampling. As the names suggest, this sampling recruits subjects based on the objective of the study and the population of interest. This sampling gives the opportunity to generalize the study results with respect to the shared characteristics of a target population. If bias is carefully avoided, purposive sampling can help recruit representative subjects. However, it is not practically possible to avoid the bias completely.

3.1.3 Snowball Sampling. Snowball sampling is a widely used non-probabilistic sampling technique in statistics and sociology research [35]. It is well known for increasing the num-

ber of samples, especially when they are hard to find. In fact, it obtained its name from the fact that it helps the sample population grow like a rolling snowball.

Snowball sampling starts with recruiting a small sample of subjects who have the characteristics of the experiment's interest. Then, the recruited subjects help the researchers identify other potential subjects who are generally hidden and hard to locate. For example, if we want to conduct a research on people who started to write code from middle school, the participants might be difficult to find since such people are not very often seen. However, a person who did this targeted behavior in middle school might know some other people who did the same – either with him/her as a group or individually. Recruiting such an acquaintance might help to recruit more people. In this way, the number of participants increase like a rolling snowball.

Recent literature shows the applicability of snowball sampling for virtual social networks [36]. Whereas traditional snowball sampling is prone to bias towards social networks of the early participants, snowball sampling on virtual networks reduce such bias due to social networks' inherent geographic prevalence around the world. It is also more effective with respect to increasing number of subjects. However, it can have a bias towards the majority of age, gender, and taste categories of online users on the focused social network itself.

3.1.4 Grounded Theory. Grounded theory is a popular approach in qualitative data analysis, and it is often used for research when there is no existing theory related to the research questions. It builds theory iteratively from data. Interview questions address the followings:

- *Core phenomenon:* What is the process?
- *Casual conditions:* What influenced the process to occur?
- *Strategies:* What actions were taken in response to the process?
- *Consequences:* What were the outcomes of the strategies?

Initially, it identifies descriptive *open codes*. These are abstract representations of events, objects, interaction, incidents that were seen repeatedly in the data. Then these open codes are

grouped into related sets. These more organized collections of codes are called *axial codes*. In this phase, patterns of the events, objects, and etc. in open codes emerge. Subsequent combinations of axial codes are more thematic that are called *selective codes*. Instead of patterns in earlier stage, relationships among the phenomena is used to build the theory/model.

3.2 Methodology

Our qualitative study started with the goals to (1) understand how users recognize sarcastic contents on social media, with or without context, (2) study what factors impact the ways of how they express sarcasm, and (3) study how users on social media in general response to sarcasm. To achieve these goals, we conducted an interview based qualitative study with social media users situated in Missouri, United States and Dhaka, Bangladesh. Our data collection consisted of semi-structured interviews with 20 participants from these two countries. We received Institutional Review Board (IRB) approval for all study procedures prior to beginning the study (See in Appendix A).

3.2.1 Semi-Structured Interviews. We conducted semi-structured interviews with participants between November and December 2018. The interviews targeted understanding participants' social media using practices and their ways of recognizing as well as conveying sarcasm. The student researcher (23 years old, Male) in this work was born and brought up in Bangladesh, and has been living in United States for more than one year. He speaks both local languages, Bengali and English. Since use of sarcasm is very common on social media, we began by recruiting participants who were active on social media. We adopted a blend of convenience sampling, purposive sampling, and snowball sampling. First, two participants were recruited from the social network of the student researcher by convenience sampling. Second, since the focus area of this research is the social media platform, the student researcher posted the recruitment flyer of this research on social media (See Appendix B). In the flyer, we described the inclusion criteria for our study and gave a high level overview of the objective of the study. We distributed the flyer through departmental email. Second, we used social media itself as a channel for recruiting

participants since most of the users on this platform will inherently satisfy one of the inclusion criteria. We shared the recruitment flyer on the social media. As a result, the subjects of interests in this research could be easily reached through purposive sampling. Third, as previous literature suggest, by keeping the comment section public for tagging improves the response rate [37], we welcomed tagging other potential participants. Again, our participants recruited through convenience sampling in the first phase helped us recruit additional participants. Thus, snowball sampling in both online and in-person social network helped us to recruit potential subjects. We also utilized in-person communication and also recruited participants through word-of-mouth. In total, we recruited 20 participants speaking two different languages from two different countries.

Participation in the study was voluntary. The average completion time of the interviews was around 25 minutes. The interviews were conducted one-on-one. We gave the participants a high level overview of the study objective at the beginning of the interview. We encouraged them to ask any question they might have, and we obtained written consents from participants before the interviews with the informed consent form, as in Appendix C. The consent form was devised keeping it at a high school standard reading level. However, we also summarized the consent form in their native language, Bengali in case of Non-English speakers. We collected the signed consent form and later sent them a copy. Interviews were conducted at a place preferred by each participant, or over Skype and in his/her native language. The interviews were audio-recorded with permission from the participants.

Interviews were semi-structured and guided by a list of topics. The set of questions is included in Appendix D. We collected the participants' demographic information like their age, gender, most recent occupation, highest attended educational level, etc. We asked about their experience about using social media, e.g., with whom they mostly interact with, what kind of contents they usually see in their newsfeed. We then asked questions that sought an understanding of how they recognize and express sarcasm, including their views about overall user response to sarcastic contents on social media.

Participants' responses were recorded anonymously. Each participant's interview record-

ing was saved on a password protected storage with code identifications assigned to them by the interviewer.

3.2.2 Participants Characteristics. Our 20 participants came from two different language speaking communities originated from two different countries and ranged in age from 19 to 34 years. With respect to their social media usage, all of our participants satisfy these following criteria:

- Must have an account with at least one SNS for more than a year.
- Must be an active user on SNS with spending 5-7 hours per week.

Participants possessed a range of socio-economic backgrounds. Five of them are undergraduate students, six are graduate students, six are employed having undergraduate or graduate degrees, and three are currently unemployed. More detailed information about our recruited participants are shown in Table 3.1.

The participants we studied represent two different sets of social media users. The participants recruited from the United States were mostly users of both Twitter and on Facebook. On the other hand, participants collected from Bangladesh were mostly active on Facebook, some of them having accounts on Twitters that they do not use often. Participants from the United States use English in all their social media activities whereas participants from Bangladesh varied in their language use on social media. They used both Bengali and English on social media, as well as a version of Bengali called "Banglish", Bengali words using English alphabet.

3.2.3 Data Collection and Analysis. The data we collected resulted in a total of 283 minutes (4 hours 43 minutes) of audio-recorded interview data and a collection of field notes. The student researcher working in this research transcribed the interviews and translated them to English. These qualitative data were analyzed using an inductive approach. We utilized grounded theory [38] as the inductive method on the interview scripts. Since to the best of our knowledge, there has been no research on theory about users' sarcasm behavior on online platforms, we in the early phase of our study, aimed to have insights/theories about users' sarcasm behavior on

Table 3.1: Demographics of participants (N=20) in the interview on sarcasm use on SNS

ID	Gender	Age	Language
P1	Male	33	English
P2	Male	29	Bengali
P3	Male	21	English
P4	Male	28	English
P5	Female	22	English
P6	Male	22	English
P7	Female	29	English
P8	Male	20	Bengali
P9	Male	31	Bengali
P10	Male	34	Bengali
P11	Male	30	English
P12	Female	22	Bengali
P13	Male	20	English
P14	Male	25	Bengali
P15	Female	24	Bengali
P16	Male	25	Bengali
P17	Male	21	Bengali
P18	Male	25	Bengali
P19	Male	19	English
P20	Male	22	English

social media. Therefore, grounded theory data analysis meets our need. As *core phenomenon*, we are interested to study how users detect sarcastic remarks on social media. We studied what factors initiate the circumstances of a sarcastic conversation to occur or a sarcastic remark to appear as a part of a conversation as the *causal condition*. This leads to our studies of *strategies*, i.e., how users express sarcasm on social media. Then we study what *consequences* or impacts sarcasm has on users' interaction on social media.

After we conducted the interviews, we prepared the transcriptions of the interview sessions. Then we read through the transcriptions several times. We identified parts of the participants' quotes where they discussed their ways of express sarcasm. The example below shows a participant's use of interjections inappropriately to convey sarcasm. We open-coded this response descriptively as "wrong use of interjection". Repeated patterns in users' interaction give rise to axial codes. For example, "opposing sentiments as parts of a single sentence" is a major clue for human users to detect sarcastic contents. For example, "wrong use of interjection" and "associa-

tion of wrong adjectives" are two open codes categorized under "opposing sentiments as parts of a single sentence". The final codes were agreed upon when themes came to a saturation. In selective coding phase, we integrated the emerged axial codes into theoretical models. Our qualitative study resulted in two separate models – (1) sarcasm detection and expression model for SNSs and (2) sarcasm use and non-use model for SNSs.

3.3 Sarcasm Detection and Expression Practices

Before discussing how sarcasm shapes users' responses to a content on social media, it is important to understand how our participants recognize and express sarcasm on social media. Broadly, the subjects whom we interviewed recognized sarcasm in two ways: (1) unusual emotion/sentiment expression style and (2) usual patterns of sarcastic posts.

3.3.1 Unusual Style of Sentiment Expression. The topics that are usually discussed on social media are often subjective human interaction. That means, users discuss their views, give opinions, and express their feelings about a matter. As discussed earlier, a substantial amount of research has been done to analyze the sentiment and emotion of these user generated contents on social media. Usually, a particular content/post generated by a user contains his/her views, and thus the sentiment towards the corresponding topic. However, in case of sarcasm, our participants report that this sentiment in a particular post might seem unusual.

3.3.1.1 Exaggeration of Sentiments. Many of our participants agree that exaggeration of sentiments in text is a sign of a post of being sarcastic. They think in a well-constructed sarcasm, there are two objectives – to point out a flaw of a targeted person (this was previously identified by previous works), and to entertain others if an audience is available which is common in usual social media settings. According to participant P8,

"It does not matter what emotion you are showing, exaggeration of it will automatically make your targeted person confused whether it is sarcasm or not, since it is so common. Your audience will often find it funny, so you get some people on your side at least, even if the person who was your target does not get the sarcasm."

While discussing this context further, an interesting reasoning was posed by our participants. According to them, when one tries to make a general post, the objective is usually to inform, to share opinion that will eventually lead the audience to some direction. However, in posts with sarcasm, the composer has no such motivation rather the sole goal here is to make people laugh that can be done by making the post subjective, as much as possible. We found this reasoning plausible during our quantitative analysis that we will discuss in a later chapter.

3.3.1.2 Opposing Sentiments. In a subjective writing, a person shares his/her positive or negative sentiment. As previous studies have suggested, a sarcastic remark often has a negative intended meaning. Our participants share the same view as the study by Cliche et al. [27]. They say that in a sarcastic post we can expect to observe opposing sentiments as part of the text. This might be evident by their sentence construct: "Wow! This is ugly" (example given by P6); here, the sentiment in the first sentence is positive whereas it is negative for the second sentence. As P7 gave us an example, "Terribly terrific", such phenomena can be observed at word level as well.

3.3.1.3 Wrong Use of Punctuation. All of our participants agree that wrong use of punctuation is a usual clue for identifying a sarcastic post. They say that this clue often occurs in sarcastic remarks as a part of a conversation. Our participant P19 gives his opinion with an example.

"Suppose, you are surprised and want to say "wow", what mark will you use? You will use exclamation mark with that. But "wow" with a period after that just says that you are not much impressed, rather you might be annoyed and are trying to show your annoyance or callousness with a cold wow."

However, they also agree that though it is a usual clue, it is not very reliable clue. They think users generally want to use social media with minimum effort. If they mistakenly use wrong punctuation with a sentence, they often do not care too much to edit the post to correct a single punctuation mark. They might rather explain that it was a mistake and correct later only if someone else pointed out at that wrong punctuation.

3.3.2 Usual Structures/Patterns of Sarcastic Posts. Participants said that they look for clues in different parts of a post. They agree with the prior views. Some participants reported that the users who have been on social media for a certain amount of time (1) exaggeration of usually necessary emotions in writing, (2) popularly used patterns of sarcastic posts that users learn with time, and (3) opposing emotions/sentiments in different parts of a single post.

3.3.2.1 Reference to Recent Objects. Our participants agree on a very interesting aspect of sarcastic contents on social media. They think there is a temporal factor to the pattern of sarcastic posts on SNSs. As our participant P1 said,

"You know when Star Wars is a very popular movie. But when a new Star Wars movie comes you can expect to see a lot of sarcastic comments referencing to famous quotes from the movie. Like, people might try to use "May the force be with you."

We were curious to know whether it is the repetition of what we explored as "reference to iconic object" earlier. Therefore, we asked the participants about this. However, they think these two are related but different factors. P1 clears up this in this way:

"... No, you see, there are obviously some fans who can tell you the movie's name and what happened in a particular scene when they hear a quote. But most people are not like that. They watch, enjoyed, and may re-watch before a new movie in that franchise comes. That's when the craze is revived, and it will make sense to use these reference only at that time. But sure, if I am talking with my friends who, I know, lives in Star Wars like me, hahaha! Then sure! I can use those reference anytime."

P17 shares a different perspective about the temporal factor of sarcastic posts' pattern. He thinks recent events that get popularity online may impact what users refer to for being sarcastic. He thinks the frequency of these references are maximum a little after when the original event got popularity. With time, users are posed more new events that might be referenced for sarcasm, and the earlier ones are not used as many times as when they were first seen; however, regular users might recognize and use those at times. When we asked for example, P17 said,

"Few years ago, there was a live telecast of an interview with general people in Rajshahi

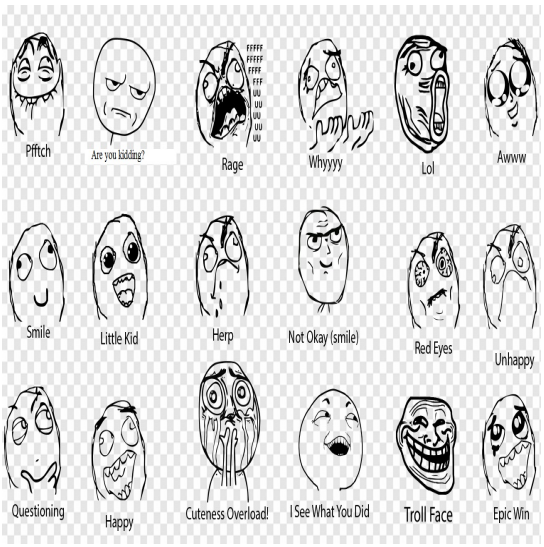
or Rangpur, I don't remember exactly, somewhere in northern Bengal during winter. The reporter asked how the people felt about the winter. So, one of them told that he did not like it and could not work for winter in local dialect, and a particular word in that dialect means something bad in proper Bengali. People in Central Bangladesh made fun about that part of the interview a lot. It became a popular sarcastic clue at that time. Every year when winter comes, you will see some people to refer to that; not as popular as before, but still it's used."

This shows a periodical pattern in temporal factor of sarcastic posts' structure. Several other later participants agreed with him. For example, P18 said it is usual to use some particular reference periodically "every four years during the world cup".

3.3.2.2 Association of Popular Memes/Meme-like Contents. A major clue that our participants reported is association of "meme-like" contents with the posts. Meme is usually an image or short video (sometimes GIF) that is taken directly or with slight variation from some popular media (e.g., TV series, movies, etc.), and spread rapidly among the internet users. For example, as many of our participants mentioned about the presence of photos of Matthew Perry (who played the character of Chandler Bing in popular TV series "Friends") in some special postures (as shown in Figure 3.1(a)) in inset of images help them to identify the sarcastic intention of the post. Discussion with our participants also gave us an idea about other widely used images that are perceived as clues of sarcasm in form of images. Use of hand-drawn meme-faces as shown in Figure 3.1(b), came as another example of such categories of visual cues. Thus, while quite different from each other with respect to the visual representation, all of them depict the same sentiment of "sarcasm" in them.

3.3.2.3 Capitalization. All of our participants agree that capitalization of words in an SNS post denotes emphasized effort from the composer for expressing his/her emotion. As we have discussed earlier, participants agree that extra effort for exaggerating sentiment might be a clue to sarcastic post. Participants also agree that capitalization might also be used to reverse the meaning or sentiment in a sentence. Our participant P13 gave us an example of what he thinks is a popular form of sarcasm of this pattern:

(a) Matthew Perry in his popular posture that work as indication of sarcasm for many participants. Thanks to Participant P14 for providing us with the sample.



(b) Samples of hand-drawn meme faces, collected from: http://tinyurl.com/yyjw36bp

Figure 3.1: Qualitative study participants contributed/suggested samples of images with sarcastic visual cues.

"If I say, the book is SOOOOO good that if you close it once you wouldn't want to open it again. It obviously has opposing sentiments in a single sentence, but when I am using this type of sentence in a conversation, I don't want others to miss that I made a sarcastic remark. So, it makes sense to emphasize to catch their eyes."

In this step, we know how "unusual style of sentiment expression" in a sarcastic post is achieved through a usual pattern of posts.

3.3.2.4 Use of Arcane Style of Writing. We observed an interesting way of conveying sarcasm among our participants from Bangladesh. There are two forms of Bengali written language – *Sadhu* (more formal, used to be in practice till twentieth century), and *Cholito* (less formal, currently is in practice). Both of them use the same fonts, however, vary in their preferred use of words. Most of our participants from Bangladesh agreed that Bengali sarcastic posts on social media are often written in the arcane form. As one of our participants, P12 said,

"You know, no one in general, nowadays write in Sadhu form. So, when you see a piece of text on Facebook that is in Sadhu language, if it is not from some old books or something, you in-

stantly know there is something the person is trying to do. I often find that posts written in Sadhu, are actually sarcastic. At least the person is trying to say something funny, if it's not exactly sarcasm."

In this context, participants P14, P15 presented a related insight. P14 opines that writing in this arcane form is not easy for all as it has not been in practice for a long time. Therefore, it is not often seen in quick sarcasm that comes as reply in a conversation. Rather, it is seen in well-written satire posts that took considerable effort from the writer of that post. Though P15 agrees with P14 about the fact that this clue is not usually seen in sarcastic comment in middle of a conversation, P15 has a different reasoning about this. P15 thinks the reason it is not seen in "quick sarcasm" is less for the extra effort needed, rather more for the fact that most people will not understand the less-used words of this form of writing. According to P15,

"Who do use Facebook nowadays? Mostly young generation. I have seen even school going children to use Facebook. They do not know this writing. Even many people of our age do not know it very well. So, if you write that in middle of conversation, they will either miss the sarcasm or ask for explanation. It will very lame if I have to explain myself after making a sarcasm."

As we can see, though our Bengali speaking participants agree that posts written in arcane form of Bengali writing might be clue for the post to be sarcastic, it is often applicable only for long and satirical posts for very concentrated audience.

3.3.2.5 Wrong Spelling. This pattern of sarcastic posts was very common among our participants from Bangladesh. They said that it is a strong clue of Bengali sarcastic posts that they see on social media. In Bengali, there are some pairs of letters with very close sounds. We showed some examples in Figure 3.2. In these pairs, the former is comparatively soft than the latter one for very similar sound. According to our participants, using the hard sound in place of the soft one, and vice-versa are clues of a piece of text to be sarcastic. However, they agree that users do not do the same with text written in English.

In his context, most of our participants agree that this pattern of sarcastic posts emerged comparatively recently. Though first Bengali keyboard was published in 1988, it was fairly com-

plicated for general users to learn. This limited the use of Bengali language on digital media. In 2014, a phonetic Bengali keyboard named Avro was released. This made it easier for users to write Bengali on computers, and eventually, helped increase the presence of Bengali online. After that, it was possible to distinguish 50 letters of Bengali alphabet easily that could not be done with 26 letters of English alphabet. For example, each Bengali letters pair in Figure 3.2 have only one corresponding letter in English. Since before 2014, most of the Bengali users wrote Bengali using English fonts online, it was not possible to use this hint for conveying sarcasm.

Soft sound	Hard sound	English sound
র	ড়	r
ত	ট	t
দ	ড	d
স	শ	s

Figure 3.2: Examples of pair of soft and hard Bengali sounds for corresponding single English sound. The list is not exhaustive.

Participants P15, P16 raised another concern about this clue to sarcasm. They said, as less educated people are not often aware about the distinction about those sounds, they spell words wrong unknowingly. Therefore, wrong spelling in Bengali text can be thought as a clue to sarcasm only if the post was composed by a person with schooling proper enough to learn spellings of usually used words.

3.3.2.6 Use of Similar Sounding Words. Participants agree that use of similar sounding words having different meanings is a major clue for sarcastic posts on SNSs. They also think that meshup of two words is also often deemed as sarcastic among their audience. The reason they think it as a better clue for sarcasm on social media is that on SNS, posts are written and audience have more time to put attention to details to understand the hint themselves, unlike for in-person communication, it is difficult to put such subtle hint on the go.

3.3.2.7 Reactions and Emojis. Our participants have pointed it out that reaction buttons and emojis often reverse the meaning of a post. They described this dynamics in a bidirectional manner. First, the post composer himself/herself can associate the post with emojis that are often used to joke on the internet. This might change the tone of the post, in other words, make the post sarcastic by creating a difference between surface sentiment and intended sentiment of the post. This aligns with the theme of opposing sentiment that we discussed earlier. As participant P2 said,

"If I see a friend to write something very serious, and put a wink emoji at the end, I'll know this person is being sarcastic about his comment."

Second, all participants agree, in a sarcastic post, the received reactions from the audience is always very mixed. While some of the audience react to the intended meaning after understanding the sarcasm, some might want to play along with the sarcasm. Our participant P2 said,

"Suppose, you posted a sarcastic post about something that annoys you, but you said you loved it or you used a "love" emoji with that. Many of your peers will show annoyance as their reaction if they understand the sarcasm. But many, specially my friends do it, might want to keep the flow going by being positive about it in their reactions and comments. Some might be just totally lost."

Thus, a sarcastic post receives a mix of emojis and reactions both from the composer and the audience that our participants think as a usual pattern of sarcastic posts.

3.4 Sarcasm Use and Non-use on Social Media

We identified four kinds of SNS users with respect to their use of sarcasm. This use comprises two functionalities – detecting sarcasm and expressing sarcasm. The dynamics of sarcasm detection and expression among users is shown in Figure 3.3 with binary levels of abilities and practices. *Non-users* of sarcasm means the users who cannot detect and use sarcasm on social media. In sub-figure (first from left in 3.3), we see both detection and expression capabilities at

"no" level. Mostly new SNS users fall into this category. *Detectors* are users who gain the experience needed to detect sarcasm on SNS, but are not experienced enough to compose sarcastic posts on their own, i.e., their sarcastic posts are often misinterpreted by the audience. The upward trend followed by the attainment of "yes" level in the detection ability in the corresponding sub-figure (second from left in 3.3) shows users gaining the ability to detect sarcasm. The expression capability is still in "no" level in this sub-figure. *Consistent* users are who can detect sarcastic posts, and express sarcasm in their posts without much misinterpretation in most of the cases. In sub-figure (third from left in 3.3), we see both detection and expression capabilities reaching "yes" level. *Discontinued* users are experienced SNS users who can detect sarcasm in most of the cases, and capable of composing such posts, however, chose not to do so for some reasons. Though the detection capability is still in "yes" level in the corresponding sub-figure (first from right in 3.3), the downward slope following the initial "yes" value of expression line denotes the users' choice of not using sarcasm.

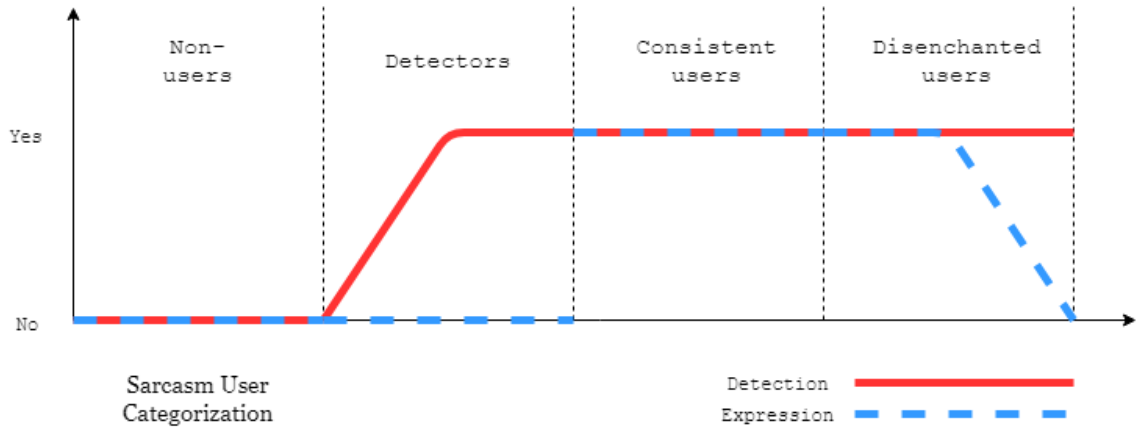


Figure 3.3: Sarcasm users and non-users engagement dynamics

3.4.1 Use of Sarcasm on Social Media. Some of our participants displayed enthusiasm for sarcasm on social media. They think that people on social media in general, should take social media lightly where they can make small jokes about the happenings of their daily lives.

They believe sarcasm is a way to do that. Thus, sarcasm may work as a driving force for making a content popular on SNSs. According to our participant P10, this force works behind popularity beyond online platforms as well. He describes SNS as the place for him to get popularity, and sarcasm as the driving force behind it. As he says,

"I am one of the very first people in Bangladesh who were regularly active on Facebook. There were some groups at that time where I mostly wrote. I think my main strength is that I write about things like politics, or day-to-day life using humor or sarcasm. People like that. That actually made me popular."

Besides, several others of our participants agree that with sarcastic contents that refers to a recent event or that can be understood with little or no context get a lot popularity.

3.4.2 Non-use of Sarcasm on Social Media. Unlike what we discussed earlier, some participants also reported their reasons of non-use of sarcasm on social media. Our participants present mainly two factors in this respect. First, inexperience of using social media might present the users a challenge while understating and conveying sarcasm on social media. Our participants think older people are a large part of this group. Our participant P1 says,

"It often happens that I am being ridiculous with my friends on a sarcastic post, and my aunt comments in a serious tone. Then, I have to explain that we are joking."

Second, previous bad experience of using sarcasm might demotivate a user from using sarcasm on social media. Most of the examples that our participants discussed had a common pattern. They used a sarcastic remark, that was criticized earlier. Or the flow might be opposite – where they were being serious about something, and their audience did not take it in the intended way staying under the hood of sarcasm. Either way, it belittled the intention of the post, and that experience demotivates the use of sarcasm. Our participants P13, P16 show a lot disgust about this. P16 says:

"There are some people who just take everything lightly. If I write about something, and someone gives a "haha" on that it upsets me a lot. I don't know why even Facebook gave this

emoji. ... I often write with my post, I will block whoever gives a "haha" without understanding the post."

3.5 Discussion

Our qualitative study ended in thematic analysis of users' sarcasm behavior on social networks. Our data analysis resulted in two models. First, the sarcasm expression model discusses how users detect and express sarcasm on social media. Second, the sarcasm use-non-use model discusses why users choose to use or not to use sarcasm on social media platforms. While the first model provides valuable insights for building sarcasm detection model/system, the second model might be useful to identify design implications for SNS platforms with respect to users' sarcastic contents sharing.

4 TEXT-BASED APPROACHES TO SARCASM DETECTION

Satire and fake news are both based on misinformation. The difference between them is their motivation. As we have seen during our qualitative study, some of our participants mentioned incidents when they confused satire to be news, particularly fake news. Though existing literature thoroughly investigates how to detect misinformation in digital contents, there has not been much research to identify the motivation behind the origination or propagation of "a particular content". we argue that the way misinformation is conveyed, i.e. the style of storytelling is a good indicator of the motivation and effort of the person(s) behind that misinformation. Our argument is based on the findings from our qualitative study that sarcastic contents on social web has a pattern in their writing – they exaggerate the feelings in the text higher than the usual. we also show how this concept can be used to design a supervised learning model for distinguishing between satire and fake news.

Though fake news detection is a Ill studied field of computer science, to the best of our knowledge, Golbeck et. al. [5] is the only work in existing literature to address the problem of classifying satire and fake news. In their work, they presented a dataset for fake news and satire. They showed applicability of naïve Bayes algorithm to classify satire and fake news from the corresponding texts. However, we found that their approach is highly biased to the buzzwords of the period when the articles of the dataset Ire collected. For example, we found that the dataset contains terms like Obama, Trump, etc. and the naïve Bayes model by [5] uses these terms to distinguish between satire and fake news. However, these terms are very specific to American politics during time around the election of 2016. Thus, this approach looses universality with respect to time.

We argue that since the motivation and the targeted audience of satire and fake news are different, there will be difference in the storytelling approach while propagating these different types of articles. Fake news are shared with a view to deceiving people. This objective of decep-

tion often becomes successful when there is no reliable medium of verifying information and the targeted audience also do not have sufficient data and context information. On the other hand, the motivation behind satire is to criticize someone. The objective of satire is fulfilled when its targeted audience have access to enough context information to understand the basis, i.e. event behind it.

We used the dataset presented by Golbeck et al. [5]. First, we show how preprocessing the data can improve performance of their proposed model. Next, we identify the most influential factors behind their model and evaluate their correlation with the time period of the data collection and found high biasness. we studied how storytelling approach varies with the categories of articles – satire and fake news. Then, we used the variation of tones used in articles to differentiate satire and fake news. Since, storytelling approach is largely independent of any particular time, we argue that our proposed approach is more widely applicable than the approach by Golbeck et al. [5].

The discussion in this section is divided into two parts. First, we identify flaws of the existing approach and showed how performance of the existing model can be improved by using the text data from the articles. Second, we discuss how the approach of conveying message differs from satire to fake news, and propose a supervised learning approach to classify satire and fake news. The rest of this section is organized as follows: the next section discusses related works; then we discussed how the model proposed by [5] can be improved and how this approach might be very specific with respect to the time of publication of the articles; in the later section we studied the difference of approaches of conveying messages according to the motivation that leads misinformation to satire or fake news, and proposed a supervised learning based approach to classify satire and fake news.

4.1 Background

4.1.1 Definitions. We deem it important to first define the terms: fake news and satire. Some prior studies [39, 40] discuss the definition of the terms fake news. According to them,

news satire, news parody, manipulation, fabrication, large scale hoaxes are different kinds of fake news. However, the problem with such definition is that this cannot take the motivation behind the origination/propagation of a content. In our work, we followed the definition by Golbeck et al. [5]. According to them, *fake news* is misinformation that is presented with the motivation to deceive the consumers. They excluded satire from the definition of fake news because of the different motivations. Golbeck et al. [5] did not provide a definition for satire, so, we followed the definition by Merriam-Webster Dictionary [41] that says *satire* is "a literary work holding up human vices and follies to ridicule or scorn; or trenchant wit, irony, or sarcasm used to expose and discredit vice or folly."

4.1.2 IBM Tone Analyzer. IBM Watson Tone Analyzer draws from the works of researchers from psychology theories and linguistic behavior. The correlation between linguistic features of written text and emotional language tones is analyzed to develop each tone dimension.

Psycho-linguists opine that our language expresses more than what we just want to say. Language can provide clues to an individual's personality, thought process, social connections, and emotional states [42, 43]. Moreover, studies show how a user's emotions are perceived by others, and collectively shape that user's online identity [44, 45].

The IBM Tone Analyzer is based on a general-purpose model that is applicable for a large range of users. It uses stacked generalization based ensemble framework – a high level model to combine lower level models to achieve higher predictive accuracy. Features like n-grams (e.g., unigram, bigram, etc.), punctuation, emoticons, greetings (e.g., hi, hello, etc.), gratitude or curse words, sentiment polarities, and etc. are fed into to categorize emotion in language. More details about the approach behind the IBM Tone Analyzer can be found in their official documentation [46].

Analyzing the characteristics of written text, IBM Tone Analyzer can provide us with three types of scores as follows:

1. *Language Scores:* IBM tone analyzer takes three aspects of language of an article as follows: Analytical (the amount of technical substance and reasoning); Confidence (the de-

gree of expression of certainty); and Tentative (the amount of words expressing uncertainty).

2. *Emotion Scores:* IBM tone analyzer calculates the probability of a sentence to express each of the following emotions: angry, joy, fear, disgust, and sadness.

3. *Social Scores:* IBM tone analyzer calculates the likelihood of a sentence to express five personality characteristics as follows: agreeableness, conscientiousness, emotion, extraversion, and openness.

4.1.3 Storytelling. The activity of sharing stories is termed as storytelling. This ties with the social and cultural context of nation, and often serves as a way of preserving and passing on values, entertainment, and information. In recent years, communicating by storytelling is considered to be a more compelling and effective way for managing conflicts, interpreting information or processes, marketing, and so on. Natural language processing research has also focused on analyzing how storytelling approach differs with the context.

4.1.4 Commons Machine Learning Algorithms. In this subsection, we will discuss some common machine learning algorithms that we used in this chapter.

4.1.4.1 SMOTE. SMOTE stands for synthetic minority oversampling technique [47]. In supervised classification problem, the percentage of number of instances in the training set plays an important role in the training of the model. Hence, balanced datasets are always preferred. However, there might be cases when we need to deal with imbalanced datasets. In such cases, model training may overfit with respect to the majority class and give a biased prediction. To remedy such situation, random under-sampling can be an approach that discards samples randomly, and thus may discard some valuable patterns. Another remedy could be random oversampling that might make the model to overfit due to exact replication of data. SMOTE appears to be a useful technique by creating new synthetic observations.

First, SMOTE plots all the minority class observations. Then it calculates the feature vectors of those. Then it takes one observation and its nearest neighbor to calculate their difference.

Then the difference is multiplied by a random number between 0.0 and 1.0. Then, we need to find a synthetic data point on the line segment by adding a random number to the feature vector. This process is repeated until the necessary number of synthetic instances have been generated to make the dataset balanced.

4.1.4.2 Naïve Bayes. Bayesian classifiers are a probabilistic framework for solving classification problems. Bayesian theorem is rooted at conditional probability. Assume, we have n attributes: $(A_1, A_2, ..., A_n)$. Given the values of these attributes, the goal of Bayesian classification is to predict the class C. For that, Bayesian classifiers compute the posterior probability $P(C|A_1, A_2, ..., A_n)$ for all classes C using the Bayesian theorem:

$$P(C_j|A_1, A_2, ..., A_n) = \frac{P(A_1, A_2, ..., A_n|C_j)P(C_j)}{P(A_1, A_2, ..., A_n)}$$

Then, Bayesian classifier gives the C_j as class that maximizes the posterior value, i.e. chooses the C_j that maximizes $P(A_1, A_2, ..., A_n|C_j)P(C_j)$.

When we assume independence among the attributes, the variation of Bayesian classifier is called Naïve Bayes classifier. Thus,

$$P(A_1, A_2, ..., A_n|C_j) = P(A_1|C_j)P(A_2|C_j)...P(A_n|C_j)$$

$P(A_i|C_j)$ is easier to calculate. And thus, class of the data instance can be easily assigned to C_j if $P(C_j) \prod_{i=1}^{n} P(A_i|C_j)$ is maximum.

4.1.4.3 Random Forest. A decision tree (DT) is a tree-like model that comes to a decision by testing on a certain attribute of an object. Decision tree with only one level of attribute checking is called Decision stump. This is considered to be a very weak classifier as it decides the class of an object based on only one attribute, thus it can output with slightly better accuracy than random guessing.

The basic idea behind ensemble algorithms is that the knowledge of crowd is better than the knowledge of a single classifier as long as the members of the crowd are slightly better than

random choices. Random forest is an ensemble algorithm of DT classifiers. Each member DT uses a random set/bag of features. Usually, given total D features, each DT uses $\sqrt{D}$ features randomly. Random choices make the DTs uncorrelated. All DTs usually have same depth. Each DT splits the training data at leaves differently. Prediction for an instance is decided by votes from all DTs.

4.1.4.4 Confusion Matrix. In case of a binary classification, confusion matrix is a table with two rows and two columns, as shown in Table 4.1. In its four cells, it stores the values of true positives, true negatives, false positives, and false negatives. This table is also known as error matrix. It is usually used for supervised learning. In case of unsupervised learning, it is called matching matrix. This table allows more metrics calculation along with accuracy.

Table 4.1: The structure of a confusion matrix for binary classification

		Actual class	
		Positive	Negative
Predicted class	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True Negative (TN)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN}$$

4.2 Investigating a Current System

Here, we use the dataset prepared by Golbeck et. al. [5]. They collected and annotated 203 satirical stories and 283 fake news stories. Their dataset contains collected articles related to

American politics after January 2016. They justified this decision to ensure minimal topic variation in the dataset. They also performed an empirical analysis on the themes of the articles in the dataset and found seven different categories: (1) hyperbolic position against a person or a group, (2) hyperbolic position in favor of a person or a group, (3) discredit a normally credible source, (4) sensationalist crime and violence, (5) racist messaging, (6) paranormal theories, and (7) conspiracy theories. They showed the applicability of multinomial naïve Bayes classifier in the classification context of satire and fake news. Their classifier achieved 79.1% accuracy with ROC area[1] of 0.88. They concluded that this shows a high difference between the type of language in satire and fake news in their dataset.

At first, we used multinomial naïve Bayes classifier proposed by Golbeck et. al. [5] with some changes. Instead of using the text directly, we stemmed (reduced words to their root/base forms; e.g.: working $\rightarrow$ work) the words using Lovins Stemmer algorithm [49]. This reduced the probability of considering the same word differently due to different structures of the sentences. We discarded the stopwords (the words that do not have much significance in word based queries, e.g.: articles) defined by [50]. Including these steps improved the accuracy of the performance to an accuracy of 80.3% with a ROC area of 0.87.

In our study, we investigated how the model makes decision or distinguishes satire from fake news. We find out which words the classifier was using to differentiate between satire and fake news. We used Shannon information gain [51] based attributes evaluation on the word vectors of the article corpus for this purpose. The top 15 words contributing most to classification of satire and fake news are: Obama, report, Donald, good, people, Clinton, Trumps, years, Barack, jobs, States, dress, United, Hillary, and government. Words with the most information gains are shown as wordcloud in Figure 4.1.

Here, we can see that the words that contribute most while using naïve Bayes classifier are mostly proper nouns or part of proper nouns (e.g. United, States) related to recent American politics. The other high information gain yielding words are also closely related to American

[1]ROC area: a representation and interpretation of the area under a receiver operating characteristic (ROC) curve obtained by predictions by the model [48]

Figure 4.1: Wordcloud of the words with high information gain.

politics. Since, the dataset was curated within the specific domain of American politics, it is expected to have many words regarding this as distinguishing terms. However, high information gain of the proper nouns show that the model is highly specific to the terms used in a specific period of time. This can be viewed as a drawback of both the existing naïve Bayes classifier [5] and our improved version.

4.3 Tone as a Way to Differentiate between Satire and Fake News

We hypothesize that the person or group of person who create fake news and satire use different approaches in their content creation or writing. Thus, the tone conveyed in a satire will be different from the tone conveyed in a fake news. Also, it is likely that the trajectory of this

level of sentiments/tones will have different trajectories according to different categories of articles – satire and fake news.

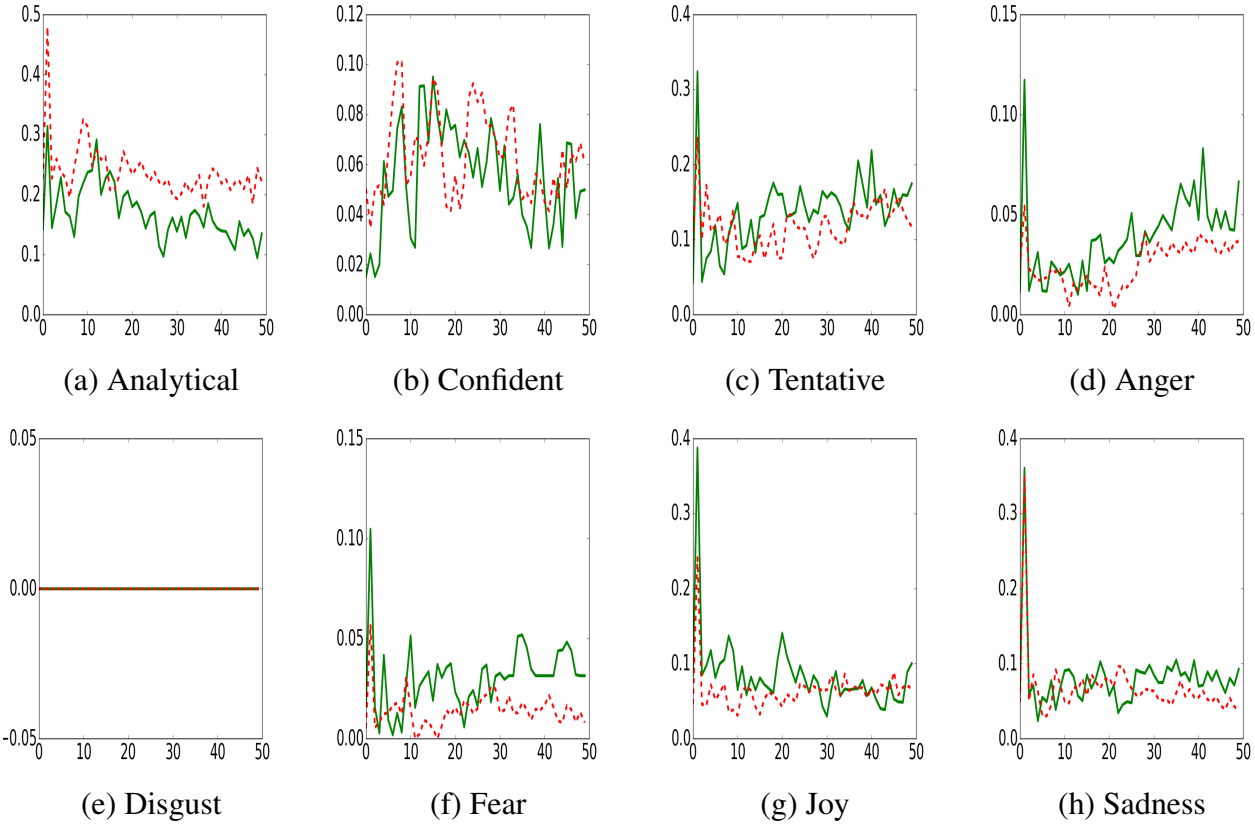
We used the IBM Tone Analyzer to calculate different aspects of each article. It outputs scores (in a scale from 0.0 to 1.0) representing the tone conveyed by corresponding sentence. IBM Tone Analyzer calculates 13 kinds of tone that belong to three different classes, as discussed in 4.1.2.

For constructing narrative trajectories, we followed the algorithm presented by [52]. We calculated these scores for each article in both categories. Then, we used the scores of each sentence in an article to construct the narrative trajectory of that particular article. We considered the scores for a specific tone in an article as a signal S_{raw}. Next, we used a Hanning smoothing window with size = 3, to construct a smooth signal S_{smooth}. Then, we cropped the signal to remove the boundary effects introduced by filtering. Finally, the smoothed and cropped signal S_{crop} is interpolated to have a canonical length of 50 samples. We refer this final signal as the narrative trajectory.

We argue that a satire article would differ from a fake news article in the way of describing an event. For example, since the motivation behind creating a fake news is to make people believe something, the content creator needs to make it look like a real news, hence, be more analytic while writing. Likewise, if a fake news tries to disseminate a conspiracy theory, it will try to convey fear. Whereas a satire needs to be funny to the readers, a fake news obviously will not have such tone in it. We constructed narrative trajectories for all articles in both categories. Then, to verify the applicability of our argument, we calculated the resultant signal of summation of all the signals from the articles in each category.

As we can see, satire articles in the dataset often had different narrative trajectories with slightly different amplitudes than the fake news articles in the dataset. For example, analytical scores for satire articles were not as high as the ones for fake news (Figure 4.2(a)); satire articles' angry tone level was often higher than that of fake news (Figure 4.2(d)) which might indicate the exaggeration of emotion in satire posts and attempt of the fake news to look unbiased like a real

news. Social tone scores had almost no trajectory in their narrative approach, and thus there was not much difference in the signals generated for satire and fake news categories. We also did not observe much difference from the graphs for disgust emotion tone score trajectory and confidence language score trajectory.



(a) Analytical (b) Confident (c) Tentative (d) Anger

(e) Disgust (f) Fear (g) Joy (h) Sadness

Figure 4.2: Comparison between narrative trajectories of satire (green solid line) and fake news (red dashed line) for different tones.

We used data processing steps like stopwords elimination and stemming that improved the performance of the system by a small margin. Whereas naïve Bayes text classifier is limited by the used terms in the articles in the dataset and thus the trained model is likely to be confined to be useful for only specific domain and time period, our proposed approach using tone data extracted from the text is less dependent on exact words of articles and thus is less likely to be confined to any specific domain or time period.

4.4 Classification Based on Tone

Since satire and fake news only differ in motivation, we have to first consider how human users actually recognize satire from fake news. Without access to information about the source of the article (e.g. website that publishes the article might be known for sharing satire), an important clue about the nature of the article can be the storytelling approach of the article. Narrative trajectory based on sentiment is an important indicator of the storytelling patterns of text articles [53–56]. The main idea behind this is that though sentence-wise sentiment scores of an article corresponds to individual reader experience, if we filter/smooth the sentence wise scores for a large amount of text, the variation can indicate narrative style/pattern of the articles of specific category [54]. Existing literature uses several different sentiment analysis approaches, including: Wordnet [56, 57], PCA [55, 58], and the IBM Tone Analyzer [46, 52, 59]. In our work, we used IBM Tone Analyzer because of its wide spectrum of considered sentiments.

We use tone information to classify satire and fake news articles. If we use the tone scores to train the models instead of the text directly, it will make the models less dependent on the exact text data, and thus, less confined to any specific domain or time period.

We argue that the headlines of satire and fake news articles might have relevant sentiment information about the article. Therefore, we calculated the subjectivity and polarity of sentiment conveyed by the headline using TextBlob [60]. We extracted the tone data using IBM Tone Analyzer. We recorded the overall tone data conveyed by the article as document tone data. Then, we calculated sentence-wise tone data using IBM Tone Analyzer. Thus, we obtained features as following: (1) two features from headline: subjectivity and polarity; (2) thirteen tone data (three language tone, five emotion tone, five social tone) for each document; and (3) thirteen summation of tone data for all sentences in the document. We also added the number of sentences in the article as a feature to train our model. In total, it gives 29 features for learning our model.

Another issue with the dataset provided by Golbeck et. al. [5] is that it has 203 satire articles (41.7%) and 283 fake news articles (58.3%). Hence, the dataset is slightly biased. Therefore, we decided to apply SMOTE (Synthetic Minority Over-sampling Technique) [47] on the minority

class satire with 40% oversampling ratio. We used Random Forest classifier for this classification task between satire and fake news. We achieved 75.8% accuracy with ROC area 0.83. Detailed performance results are shown in Table 4.2.

Table 4.2: Performance of classification task with tone data extracted from articles (article text independent approach)

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
Satire	0.729	0.212	0.775	0.729	0.751	0.518	0.827
Fake	0.788	0.271	0.743	0.788	0.765	0.518	0.827
Weighted Avg.	0.758	0.242	0.759	0.758	0.758	0.518	0.827

We achieved comparable performance without using text data unlike the existing work [5]. We argue that if we use tone data along with text data, it will show increased performance in classifying satire and fake news. Like the existing work [5], we also added the theme information with these features. With all these features combined, we achieved 82.5% accuracy with ROC area 0.91. We show the detailed performance results using Random Forest classifier [61] for this classification in Table 4.3. We used Scikit-learn [62] for training the model.

Table 4.3: Performance of classifier model with text, tone, and theme data combined

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
Fake	0.905	0.254	0.782	0.905	0.839	0.660	0.911
Satire	0.746	0.095	0.887	0.746	0.811	0.660	0.911
Weighted Avg.	0.826	0.174	0.834	0.826	0.825	0.660	0.911

We achieved comparable performance using tone-based approach and we showed that combining tone information with text and theme data of the articles can improve the performance of the model by a considerable margin. However, we further investigated the contribution of the features of our model to classify satire and fake news articles using Shannon information gain [51]. Table 4.4 shows top five features in our model with highest information gain. We can see that though feature vectors generated from model are associated with text of articles, tone and

theme based features have highest information gain, and thus can be good features for classifying satire and fake news.

Table 4.4: Five features with topmost information gain values (type of the feature is inside parentheses)

Feature	Information Gain
Conspiracy (theme)	0.1035
Document Joy (tone)	0.0668
Document Analytical (tone)	0.0402
Sentences Analytical (tone)	0.0395
Sensationalist Crime/Violence (theme)	0.0390

4.5 Experiment with non-English Dataset

Sarcasm detection being a comparatively new field in machine learning research, a lot of varieties in datasets are not available yet. For example, to the best of our knowledge, there is no publicly available satire-fake news dataset containing data in non-English language. In our experiment, our next step was to test the performances of the improved existing system and the tone based approach on non-English data. We chose Bengali language for collecting non-English data.

4.5.1 Dataset Collection. We collected 30 satire posts and 30 fake news posts. We collected the satire posts from popular Bengali satire sites like Motikontho[2], Earki[3]. These sites vary in their period of popularity – while the first one was popular formerly and usually writes satires on political events, the second one is a popular satire site as off 2019 and writes satires on different events – both political and non-political. Our collected data reflects this diverse nature with slightly more political posts than the non-political ones. Our collected satire posts vary in the dates of those being posted since as early as 2013 till as recent as 2019. Since we also did not find any fake news dataset in Bengali, we collected those data by ourselves. To be certain about

[2]https://motikontho.wordpress.com/
[3]https://www.earki.com/

whether a post is a fake news we relied on the most popular Bengali fact-checking site Jaachai[4]. This fact-checking site follows a common structure of their posts. They include an example of a widely popular version of the false news before it discusses the true news. An example can be referred to by this link[5]. We collected those examples of fake news in different posts to serve as samples of fake news in our dataset.

4.5.2 Experiment and Results. Since we have a small non-English dataset, we used the entire dataset for testing purpose. First, we trained two models using the dataset by Golbeck et al. [5] – the first one with the naïve Bayes approach used by [5] after adopting the proposed improvements in an earlier section of this chapter, the second one with the tone dependent classification approach described in the previous section. Then, we used the non-English dataset as test data to these models. That means, none of these models has not seen non-English data before. The accuracies of the models on non-English dataset are shown in Table 4.5.

Table 4.5: Performances of the Naïve Bayes and the tone based approaches on non-English (Bengali) dataset.

Model	**Accuracy**
Improved Naïve Bayes Approach	93.33%
Tone Based Approach	61.29%

4.6 Statistical Checking of the Tone Based Approach

As we can see, tone based approach performed much worse than the Naïve Bayes approach on non-English dataset, though the tone based approach was better than later one on English dataset by Golbeck et al. [5]. This directs the next step of our experiment – to check whether the differences between the tone values of satire and fake news are significant enough or it was observed simply due to the particular sampling of the dataset by [5]. To check this, we used t-test on the tone values of about 6% (15 samples from each category) data of the dataset by Golbeck et al. [5].

[4]https://www.jaachai.com/
[5]https://en.jaachai.com/posts/post-1872

Table 4.6 shows the results of t-test for different language and emotion tone values. In t-test, t-values can indicate the effect size. The higher the effect size is, the difference is less likely to be random. This can be shown more easily with p-values. A p-value less than a certain value, e.g. 0.05, means that there is strong evidence that the difference between two samples are strong, and the probability of this difference being out of randomness is less than 0.05.

Table 4.6: t-test result on different language and emotion tone values

Language/Emotion Category	**t-value**	**p-value**
Analytical	0.7816	0.44
Confident	0.2387	0.81
Tentative	0.9603	0.34
Anger	0.8443	0.4
Disgust	0.0	INF
Fear	0.3214	0.75
Joy	0.3044	0.76
Sadness	0.4674	0.64

However, in our case, we can see the p-values are so high that these do not provide strong evidences about the differences not being out of randomness. Though these high p-values are insignificant with respect to satire and fake news categories, close look at those values can be helpful. The values help us understand deeper studies of which language or emotion aspects can be beneficial, in other words, which language or emotion aspects have potentials to be used to differentiate satire and fake news. As we can see, Tentative, Anger, Analytical – these emotion and language classes have smaller p-values than the others. On the other hand, infinite p-value for Disgust emotion class indicates its inability to differentiate satire and fake news.

4.7 Discussion

Our analysis of text data to differentiate satire and fake news has several steps. First, we investigated an existing approach. By adopting some standard natural language preprocessing, we improved the performance of the existing system by some small margin. Then, we used one of our qualitative study's findings. As the study indicated, the storytelling approach or the changes

in tones in sarcasm is different from that in other types of descriptive text data on social networking sites. Narrative trajectories of different language and emotional tones for satires and fake news data from [5] also supported this finding. With this tone-based approach, we achieved comparable performance to the existing approach on the dataset collected by [5].

Since our tone based approach does not use the exact words in the text data samples, we hypothesized that our approach will have better performance on non-English data compared to the existing naïve Bayes-based approach by Golbeck et al. [5]. To test our hypothesis, we collected a small dataset containing 30 data instances of satire and fake news. However, trained both our model and the existing approach model [5] on English dataset, our model could not perform as good as the existing model on non-English dataset. To investigate the reason of our model performing well for English dataset but not so for non-English dataset, we statistically checked whether the difference of tone values in English dataset observed was random or significant enough. Our t-test suggested differences of most of the tone values were due to randomness, and thus might not be very useful.

5 USING VISUAL CUES FROM IMAGES TO DETECT SARCASM

As we discovered in our qualitative study, memes, a special and popular type of images on SNSs, often contain visual cues that indicate whether or not a post is sarcastic. Thus, identifying memes (as in Figure 3.1 can be a step towards detecting sarcasm in image based posts on SNSs.

As a step towards detecting sarcasm using multimedia information, we focus on the visual style of images to express sarcasm. We hypothesize, images that are capable of presenting sarcastic cues have different visual style than the ones that do not. Prior works show that neural networks can be expressed as a powerful non-linear classifier that can map any set of inputs to any sets of desired outputs. We argue that a convolutional neural network architecture-based model can be designed that can categorize images depending on the "sarcastic" and "non-sarcastic" visual cues.

In this chapter, at first, we will give an overview of some background concepts that are necessary to understand the experiment methodologies. Since we are focusing on image data in this chapter, data collection section of this chapter will concern itself with how we collected data from Flickr, an image-based social networking site. Then we discuss our experiment and results.

5.1 Background

We used images from a popular social networking platform for sharing photos called Flickr. Specifically, we trained a convolutional neural network (CNN) to detect sarcasm from visual cues. For data collection, we used the snowball sampling technique.

5.1.1 Image Representation: RGB Color Space. Image representation is basically how pixels are represented on the monitor or screen of a computing device. We can think of pixels as the building blocks of images. In the representation of any digital image, pixels have a numeric representation that indicates its color.

Now, the question is how the numeric representation of these pixels are assigned? There

have been mainly two approaches to assigning these representations. They are additive colors and subtractive colors. In our discussion, we will focus on additive colors. Whereas subtractive approach creates colors by absorption/subtraction from white color, additive approach creates colors by mixing lights/illumination.

Depending on how pixels are using lights, there are many variations of additive color models. However, before proceeding any further, we want to note that none of these color models are perfect and each has its own merits. We will focus on RGB color model in our discussion. It is a tri-stimulus color space using three primary colors: red, green, and blue.

To create colors in RGB model, any combinations of red, green, and blue are superimposed. Adding any two of these colors produces secondary colors: cyan (green + blue), magenta (red + blue), and yellow (red + green). In a version of RGB model, there is an additional component called alpha. However, we will not discuss about that extension.

The RGB color model uses a 24-bit color that is divided into three 8-bit components of red, green, and blue. That means in RGB color model, we can have $2^8 = 256$ distinct values (0 to 255) of each of the component colors. The values of the components indicates the intensities of the corresponding components, i.e., a value of 0 indicates absence of a color component, whereas a value of 255 means the full intensity of that color component. In total, RGB model offers $(2^8)^3$ variations of colors. For example, $R = 255, G = 255, B = 0$ creates the color Yellow. Similarly, values of all R, G, B components being 0 means absence of all color components or no light, creating the darkest color, Black. Likewise, all three R, G, B components having the highest value 255 creates the color White.

The RGB model can not only represent color images but also grayscale images. If all three components of a pixel, or of all pixels in an image are equal, that is basically a grayscale image.

5.1.2 Artificial Neural Networks. Before we learn about artifical neural networks (ANN), we need to know about perceptron. Perceptron is a linear classifier. It may not be optimal for many real-life problems but it is very simple to use. In multidimensional space, linearly

separable classes can be identified by a perceptron by a hyperplane. For two-dimensional space, hyperplane reduces to a line.

A perceptron finds the linear classifier by minimizing a defined cost function. To find the minimum cost, gradient descent is adopted. The perceptron algorithm converges in a finite number of iteration steps to a solution if patterns are linearly separable.

But what happens if the classes are not linearly separable? For example, let us consider the XOR separability for two binary variables. As shown in Figure 5.1, AND, OR classifiers for two binary variables x1 and x2 can be drawn with a single line. However, XOR function output cannot be identified with a single line that is shown with the shaded area in the Figure 5.1.

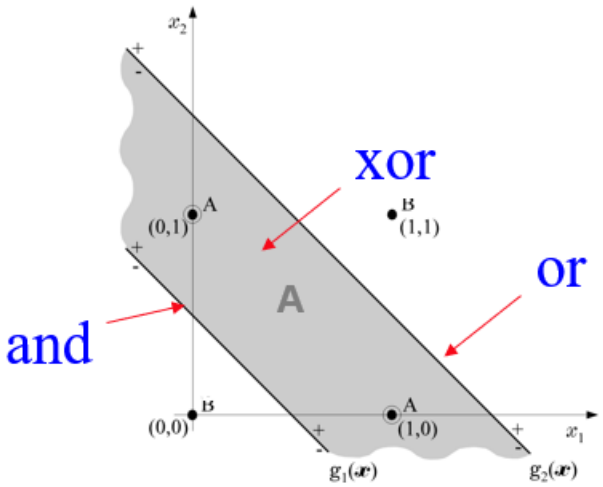


Figure 5.1: XOR function classification with multiple perceptron.

Each of AND and OR can be realized with a perceptron. They perform a mapping according to their corresponding functions as the first phase. The final output mapping for XOR function is done on the transformed value in the first phase. This can be performed via a second line that is realized by another perceptron. Computations of the first phase perform a mapping that transforms the non-linearly separable problem to a linearly separable one. Thus, multi-layer perceptron (MLP) or ANN can transform more complex non-linear problems into linearly classifiable one.

5.1.3 Convolutional Neural Networks. Convolutional neural networks (CNNs) can process large images or video sequences. For that, it automatically generalizes for spatial translations of inputs. It is applicable for any input that can be laid onto a grid.

Convolutional neural network replaces matrix multiplication with convolution keeping everything else same in artificial neural network. These convolutions, specifically the filters in them, can detect patterns, i.e., edges, corners, geometric shapes, etc. When adding a convolution layer to a CNN, we have to specify how many filters it should have. A filter is a small matrix for which we decide the numbers of rows and columns. The values in the cells are assigned randomly. When the layer receives an input matrix, the filter will slide gradually until it has traversed every $r * c$ blocks in the input. This process is known as convolution. By training we actually learn the values we should have in the cells of the filters. Learned values of filters help us detect patterns. As the network goes deeper the pattern recognition capability of it becomes more sophisticated and it can detect objects.

5.1.4 Several Backbone Neural Networks. We used four backbone networks in our experiment. Those are: VGG16, ResNet50, InceptionV3, and Mobilenet. We discuss those briefly in this subsection.

5.1.4.1 VGG16. The visual Geometry Group at the University of Oxford proposed a convolutional neural network in their paper [63]. It can find a wide range of objects in a given image. It accepts RGB images of 224x224 pixels. It was the first runner of ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2014 in the classification task. It uses 16 layers in total comprising of convolution layers (3x3), max-pooling layer(2x2), and fully connected layers. A popular variant of VGG-16 is VGG-19 with 19 layers.

5.1.4.2 ResNet50. Resnet is a short form of Residual Network [64] (schematic shown in Figure 5.2). As the name suggests, this network uses residual learning. When very deep convolutional networks are used, network accuracy gets saturated and degrades. No matter what happens, a deep network should be at least as good as a shallow network. That means we do not rule

out the necessity of deep networks for some problem cases, however, we want the deep network to adjust itself to be a shallow network if the problem demands.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 5.2: Schematic of Resnet having different numbers of layers

He et al. [64] proposes we can use a deep network and train all early weights of the network to be zero. Instead of learning a direct mapping between $x \longrightarrow y$ like $y = H(x)$, we define a residual function $F(x) = H(x) - x$ i.e., $H(x) = F(x) + x$, where F(x) is the output of stacked non-linear layers and x is identity function. If we want to make the network shallow, making $F(x) = 0$, helps us skip a certain layer. Figure 5.3 shows how a residual block is used with weight layers.

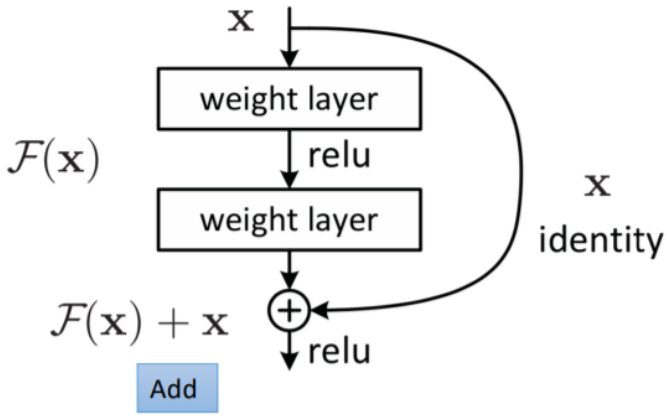


Figure 5.3: Residual block with weight layers and skip connection.

5.1.4.3 Inception V3. The premise of Inception networks is that salient parts of the images can have large variation in size and location. Thus, choosing the right size of kernel can be difficult – larger kernel for globally distributed objects and smaller kernel for locally distributed ones. Therefore, the idea here is that we use multiple sizes of kernels in a single layer making the network go wider. A naïve implementation of inception block looks like as in Figure 5.4.

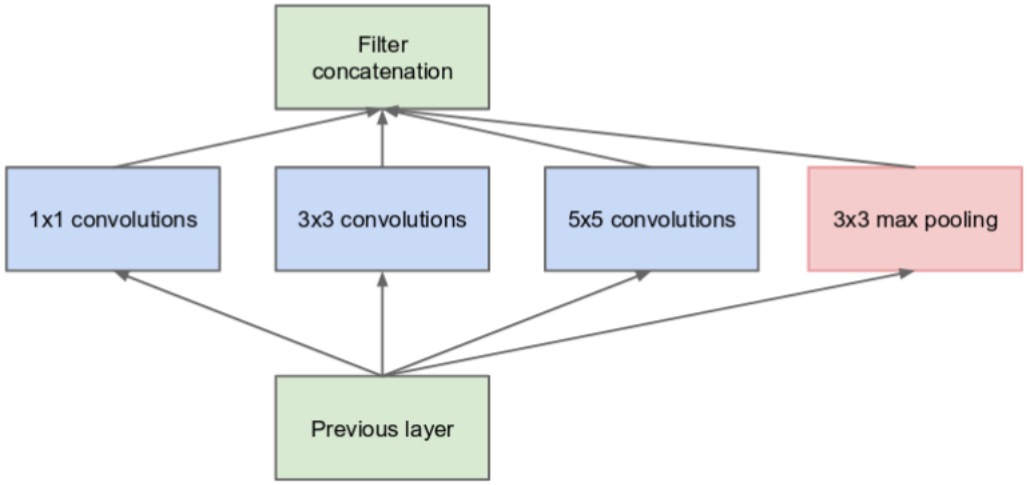


Figure 5.4: A naïve implementation of inception block from Szegedy et al. [1]

Inception V2 and Inception V3 were presented in the same paper [65]. In V2, the representational bottleneck was removed and 5x5 convolution was factorized using two 3x3 convolutions to make it computationally faster. Inception V3, along with the changes in V2, incorporated RMSprop optimizer, factorized 7x7 convolutions, and used batch normalization for auxiliary classifiers.

5.1.4.4 Mobilenet. Mobilenet is an architecture proposed for mobile and embedded based computer vision applications to function with limited computing power. The normal convolution is replaced with depthwise convolution in this architecture. This significantly reduces the number of parameters than the network with normal convolution and same depth. Reducing the number of floating point matrix multiplication, it sacrifies a small amount of accuracy with a network of low complexity.

5.1.5 Transfer Learning. Training a CNN from scratch, it has to learn to identify basic shapes from the very beginning. Therefore, usually networks are seldom trained from scratch. Specially, if the task at hand relates to a previously known task, there is a high probability that both networks–the one for a new task, and the one already trained for a related task, might have to learn to identify similar shapes/patterns. Again, if the size of the dataset is not large enough, the training of the network from scratch might take forever or might overfit the dataset.

For example, if the task at hand is to identify whether an image is sarcastic or not, the network for that might need to learn certain patterns. Though object detection is not directly the task at hand here, the learning for an object detection network might come handy. Thus, our training might take less time to achieve a certain performance if we start with a pretrained model for another task, hoping that it will *transfer* its learning to this new task. This is transfer learning.

For transfer learning, we start with an architecture and its pre-trained weights. We modify the last layers according to the demand of our problem. Finally, we train the entire network on our dataset.

5.2 Dataset Collection

While collecting datasets for sarcasm detection, the first challenge is finding sarcastic posts online and differentiating them from non-sarcastic posts. As we discussed in section 2.2, there are two approaches to collecting sarcasm datasets – independent annotation, and self-annotation. To recognize the importance of context information for identifying sarcasm data as pointed out and emphasized by [20, 26], we used the latter approach as done by Khodak et al. [26] and Reyes et al. [25]. In this phase, we focused on utilizing image based data for sarcasm detection instead of utilizing only text-based data. Whereas Schifanella et al. [33] and Razali et al. [29] discussed that multimodality from text and image based posts can help complement context information for each other, in this phase, we are arguing that images alone can also contain important visual cues that can indicate sarcasm in a post. Therefore, we collected images that were labeled by users who posted those online with hashtags that denote sarcasm.

5.2.1 Deciding on Search Words. As pointed out by Wallace et al. [22], sarcasm is not often used during normal conversation. Therefore, though the amount of posts available on social networking sites is abundant, it is hard to find posts that were intended to be sarcastic. Our initial sets of hashtags for both classes had one word in each. The sarcasm class consisted of only one word – "sarcasm", and for non-sarcasm class consisted of the word – "non-sarcastic". Obviously, these sets of words are too small to consider for large amount of data collection. Therefore, we borrowed from the idea of snowball sampling [35, 36] as described in section 5.1.

Images with sarcastic intent are hard to find. Likewise the hashtags used to indicate sarcastic images are difficult to locate. Therefore, we used this non-probabilistic sampling technique to increase the number of words in the sets of hashtags to consider. For increasing the number of words in the sarcasm class, we consider the other synonyms of the only existing word in this set for now (i.e., "sarcasm") as potential samples of hashtags that might be used with a sarcastic image because those words will have same meaning as the existing word – "sarcasm" and the users can use any of those words at their discretion. We chose the following words: "sarcasm", "sarcastic", "irony", "satire", and "wit" as potential hashtags for images of the sarcasm class.

Filatova et al. [9] pointed out while sarcasm has opposite literal and intended meaning, the literal meaning is often positive, and the intended one is negative with a clear victim. Thus, sarcastic utterances can easily be confused to be positive statements that are made about a person or a group of persons. Positive statements used to admire a person or group of persons can be termed as praise. Thus, images using words like "praise" or its synonyms (e.g. applause) as hashtags can be confused with sarcasm that in reality are not sarcastic. Now, we focus on the fact that sarcasm has opposite literal and intended meaning. This scenario is not desired or seen while conveying information. Thus, we consider the words like "information" and "fact" as member of set of hashtags used with non-sarcastic posts. We also considered different variations of the word: "non-sarcastic", e.g., "nonsarcastic", "not-sarcasm", "notsarcasm". Therefore, the words we consider as hashtags for the non-sarcasm class are: "non-sarcastic", "praise", "applause", "fact", "in-

formation". We assume that none of these words are used sarcastically. If an image has hashtags from both sarcasm and non-sarcasm classes, we discard that image during data collection.

5.2.2 Yahoo Flickr Sarcasm (YFS) Dataset. We collected images from the popular social photo sharing service, Flickr. Previous works in image based emotion classification and social media research have widely used the publicly available image data from this platform [66, 67]. We used the Flickr API service to collect the image data. At first, we collected the image metadata using the words discussed in previous subsection as keywords or search query parameters in both "sarcasm" and "non-sarcasm" classes. As parts of metadata, we collected image IDs and image urls. Then we retrieved the images by their urls returned by the queries and saved them under the keywords that were used for queries. Table 5.1 shows the number of images that we found for each keyword. Images often contained more than one word from the same category as hashtag keywords. We considered such images only once in the table. That means, we saved an image only once in our dataset when the image had more than one word from the same class as keywords. We ensured this by comparing images by their ID in metadata that is unique for any post on Flickr. To respect the privacy of user generated content on social media, we only collected the images and their metadata that are available under creative commons license. Following the study by Gajarla et al. [66], we also sorted our result set by "interestingness" [68]. We queried with the hashtags. We collected images' metadata at first and then we collected images later. This helped us collect data without exhausting the daily query limit of the API. In this way, the metadata information can be utilized directly without further queries for collecting data in a short amount of time.

We collected 1846 images in total. Among these images, 443 images were collected using keywords from the "sarcasm" class as query terms and 1403 images were collected using keywords from the "non-sarcasm" class. We split our data in each class into 90% training data and 10% validation data. This resulted in a training set having 1603 images, among that 399 images belonged to the "sarcasm" class and 1263 images belonged to the "non-sarcasm" class. The validation set consisted of 184 images of that 44 images were from the "sarcasm" class and 140

Table 5.1: Number of images for each keyword individually.

Class	Keywords	Number of Images
Sarcasm	sarcasm	13
	sarcastic	22
	irony	179
	satire	88
	wit	141
		Total: 443
Non-sarcasm	non-sarcastic	2
	praise	437
	applause	253
	fact	35
	information	676
		Total: 1403

images were from the "non-sarcasm" class. We saved all the images in the jpeg format. Our collected images varied in sizes and shapes. The smallest image in our dataset is 8.0 KB and the largest image is 20.0 MB. The varying resolution of the images was ranged from 180x135 to 6600x4514. We refer to our dataset as the Yahoo Flickr Sarcasm (YFS) dataset.

5.2.3 Comparison against Benchmarks. Wallace et al. [22] pointed out that sarcasm occurs very infrequently, and that was also reflected in our data collection. The largest sarcasm dataset so far was done by Khodak et al. [26]. They prepared a text data based sarcasm dataset using posts from subreddit "politics" on Reddit. In their dataset, they reported 23.2% instances in the "sarcasm" category with the rest 76.8% samples in the "non-sarcasm" category. They presented this as a benchmark for sarcasm collection dataset. Work by Walker et al. [69] collected data from the Internet Argument Corpus that had only 12% of data that belong to the "sarcasm" class that does not meet the benchmark set by Khodak et al. [26].

The only existing image-based sarcasm dataset was prepared by Schifanella et al. [33]. It was a balanced dataset having 50% instances in each category. They reported that 91.16% of their data had images in it. Therefore, we can calculate 45.58% posts of their dataset were samples of image based "sarcasm" posts that also satisfies the benchmark by Khodak et al. [26]. However, they did not include any information about availability of their prepared dataset. Our YFS dataset has the following distribution: 24% images belonging to the "sarcasm" class and

76% images in the "non-sarcasm" class. While satisfying the benchmark for sarcasm dataset collection by [26], we also made our dataset publicly available.

5.3 Methodology

5.3.1 Study Design. We collected image data from Flickr and divided that into two categories: "sarcasm" and "non-sarcasm". Thus, we can model our task of sarcasm detection as a binary classification problem. Usually image-based classification problems are tackled in two ways. Many prior works attempted utilizing hand-crafted features in this respect. In image-based emotion detection research, Siersdorfer et al. [67] also took the same approach of proposing features that were manually calculated and extracted. However, being unable to choose good, hand-crafted features poses the risk of building a poor classifier. At the same time, calculating these features manually has large overhead. With the recent popularity of neural networks and their superiority in identifying and extracting hard-to-detect features from images, many recent image based classification works take this approach. In image based sentiment analysis research, Gajarla et al. [66] suggested this approach.

Scifanella et al. [33] advocated for multimodality in sarcasm detection for the first time. They attempted to understand the semantics in the images. They compared that with textual description of the images. They reported an improved performance in sarcasm detection with the inclusion of semantics of the images. We agree with their concept of utilizing images as a source of information for sarcasm detection. However, instead of semantic representation of images, we focus on the visual cues of the images. We argue that sarcastic images posted on social media have different visual cues or representation than the ones that do not have such sarcastic intent. Thus, whereas Schifanella et al. [33] focused on semantics of the images, we are focusing on visual cues or representation ways for detecting sarcastic images. Though visual cues have been used for emotion detection in images by Gajarla et al. [66] and Siersdorfer et al. [67], none of them considered sarcasm as a mode of emotion or sentiment. We propose that visual cues in an image as input to a CNN can detect sarcasm with high accuracy.

At this phase, our methodology has two approaches. First, we trained a CNN to identify sarcastic cues in image. Second, we performed transfer learning from several popular object detection neural network models to see how fine-tuned content identifier networks perform for the task of sarcasm detection.

5.3.2 System Design.

5.3.2.1 Preprocessing. Since the size of our dataset is not large, we performed image augmentation process on our dataset. We passed the images through a shearing factor of 0.2 and a zooming transformation factor of 0.2. The values of these hyperparameters were chosen from our previous experience of working with small datasets. We also performed horizontal flips of the images to increase the number of training samples. As we discussed earlier, images are represented in RGB color space with integer color values of red, green, and blue ranging from 0 to 255. However, this [0, 255] range is too large for our CNN model. Therefore, we down-scaled the color values to [0, 1]. Using image augmentation techniques, we prepared a dataset having 2000 images containing all original images along with some synthetically generated image samples. For preparing a validation dataset of 800 images, we applied the same augmentation methods on the held out validation images.

5.3.2.2 Architecture. A convolutional neural network (CNN) consists of several convolution layers, a sub-sampling step, and then one or more fully connected layers. We used a sequential neural network architecture. At this point, we will define a layer group. Let's call a group of layers consisting of a 2D convolution layer, an activation layer with "relu" function, and a max pooling layer with pool size = (2, 2); and name it "Layer Group A". We repeated this Layer Group A three times consecutively, that was followed by a flatten layer, and two dense layers with "relu" and "sigmoid" activation functions respectively. Figure 5.5 shows the architecture of the sequential CNN architecture used in our experiment.

5.3.2.3 Learning. We followed a mini-batch weight updating scheme for our experiment. That means, instead of updating the weights in neurons after passing each image, we update the weights after passing a batch of images. We tried several different values for batch size

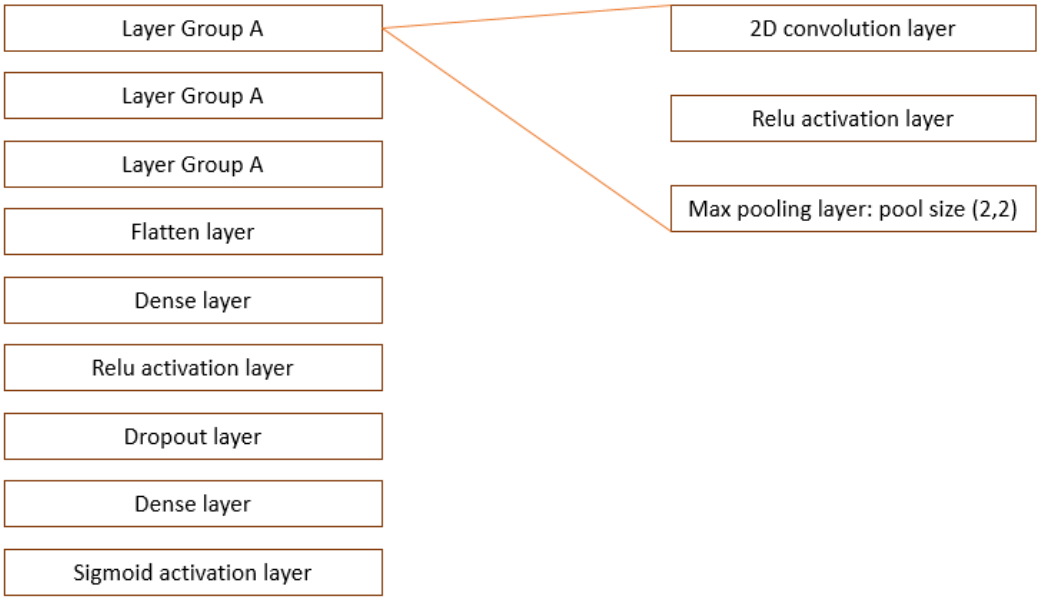


Figure 5.5: Structure of the CNN to learn sarcastic visual cues.

of images to the CNN. Batch size is a hyperparameter in CNN and determining the value of it is an optimization problem. The larger the batch size, the CNN will require more memory to train. On the contrary, a small batch size will require less memory. The choice of batch size for a CNN poses a performance-time trade-off. For our final model, we settled for a batch size of 16. Another important hyperparameter that we had to choose is the number of epochs, i.e. how many times all the training samples go through a cycle consisting of a forward and backward pass. We set the number of epochs to 50, i.e., all our training images pass through the CNN 50 times in a batch size of 16 while updating weights and biases values.

We implemented our CNN using Keras, a wrapper over Tensorflow. We used an input size of 240x240. We chose *binary crossentropy* as the loss function for our model. Next, we find an appropriate learning rate for our CNN. Przelaskowski et al. [70] discussed that image data has a sparse representation. In a later study, Ruder et al. [71] gave an overview of gradient descent optimization algorithms and concluded that for sparse input data, using one of the adaptive learning rate methods is most likely to achieve the best results. A well-suited algorithm for using with sparse data is Adagard, a gradient-based optimization algorithm that adapts the learning rate to

the parameters. We chose to use the RMSprop optimizer, an extension of Adagard that deals with its radically diminishing learning rate.

We achieved an accuracy of 84%. Our dataset is not fully balanced. That means the numbers of images that belong to two categories: "sarcasm" and "non-sarcasm" are not equal. Therefore, we need to choose a metric that ensures proper training for imbalanced dataset. F1 score is a harmonic average of precision and recall. Unlike accuracy that is affected mostly by true positives and true negatives, F1 score reflects all four possible cases – true positives, true negatives, false positives, and false negatives. Thus, F1 score is a better metric for optimizing on an imbalanced data. While training the model again with respect to F1 score, we kept the values of other hyperparameters unchanged. This new training achieved an F1-score of 79%. A high F1-score of the model indicates that the model is not biased towards any category.

5.3.3 Fine-tuning Existing Models. We also looked into the neural network based work on image based sentiment analysis by Gajarla et al. [66]. They experimented using transfer learning on different variations of VGG-ImageNet and ResNet. They reported 73% accuracy for positive and negative senitment detection using visual cues identified by fine tuning the last layers of the aforementioned models.

For fine-tuning existing models to perform the sarcasm detection task, we chose several popular exsiting models, e.g., InceptionV3, MobileNet, ResNet50, and VGG-16. For each of these, we replaced the final layer of the network with a layer having two output nodes so that it can perform a binary classification task of "sarcasm" and "non-sarcasm" detection from images.

5.4 Results

We discuss the performance of our CNN-based model and compare its performance with (1) image semantics based sarcasm detection, (2) image visual cues based sentiment analysis, and (3) transfer learning based sarcasm analysis.

5.4.1 Semantic Based and Our Visual Cues Based Approaches. At first, we compare our visual cues and CNN-based approach with image semantics based approach of sarcasm

detection. To the best of our knowledge, Schifanella et al. [33] is the first, and till our work [72], the only paper that discuss the importance of considering images for sarcasm detection with implementation. They reported 65% and 72% accuracy on their two datasets using visual semantic representations of images. Since they had a dataset having 50:50 ratio between the "sarcasm" and "non-sarcasm" classes, we can consider their accuracy value as F1-score.

Our CNN based classifier trained to classify "sarcasm" and "non-sarcasm" classes from images achieved 84% accuracy and 79% F1-score. Therefore, we can see visual cues based approach achieved superior performance over semantic representation based approach by 16.67% (accuracy metric) and by 9.72% (F1-score metric).

Now, the question is, what kind of visual cues has our model learned? Understanding these visual cues will help us understand what kind of images are posted on social media with sarcastic intent and whether the learning of the model echoes our and our participants' experience on SNSs.

5.4.1.1 Sarcasm. The model learned to categorize images with writings in them as sarcastic images. This aligns with the characteristics of widely used *memes* on social media. Another interesting type of image in this category is the crudely edited photos that are also used as memes. Many of outdoor images of non-human objects were predicted to belong to this category by our model. Besides, our model also learned to identify all hand-drawn cartoons as an indicator of sarcastic images. These visual cues (e.g., crude edit, hand-drawn cartoon, memes, etc.) of sarcastic images also align with our findings from our qualitative study (examples shown in Figure 5.6).

5.4.1.2 Non-sarcasm. The model identified most indoor images as non-sarcastic images. For outdoor images that the model identified as non-sarcastic had a common characteristics, that is, those images were of human beings in outdoor settings. The model seems to learn human faces as a good indicator of non-sarcastic posts that is also supported by some of our participants' opinions. Some examples of images categorized as "non-sarcasm" by our model are shown in Figure 5.7.

(a) Image with text (memes)

(b) Outdoor image of non-human object

(c) Handdrawn image (cartoon)

Figure 5.6: "Sarcasm" Labeled Images

(a) Outdoor image of human

(b) Human face

(c) Indoor image

Figure 5.7: "Non-Sarcasm" Labeled Images

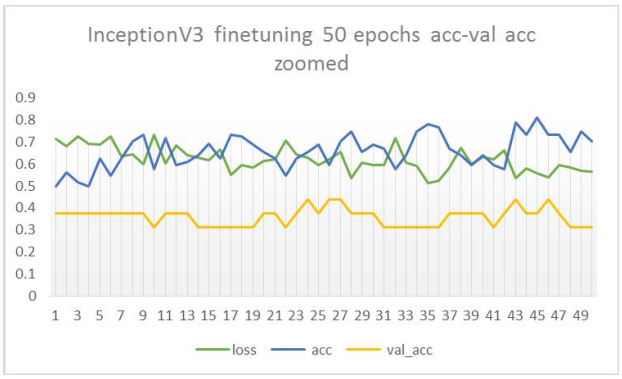
5.4.2 Visual Cues Based Sentiment and Sarcasm Detection Approaches. Gajarla et al. [66] utilized visual cues of images for identifying sentiments. They used transfer learning to accomplish their task. They fine-tuned only the final layers of VGG-Imagenet, VGG-Places205, and Resnet50 and achieved 67.8%, 68.7%, and 73% accuracy respectively.

Our sarcasm detection CNN model with 84% accuracy exceeds the sentiment analyzers' performances. It shows that visual cues is an effective feature not only for binary sentiment analysis but also for sarcasm detection. Thus, we argue that whereas Gajarla et al. [66] classifying sentiment as positive and negative ignored sarcasm as a sentiment, considering sarcasm we can come up with a tri-categories classification problem with "positive", "negative", and "sarcasm"

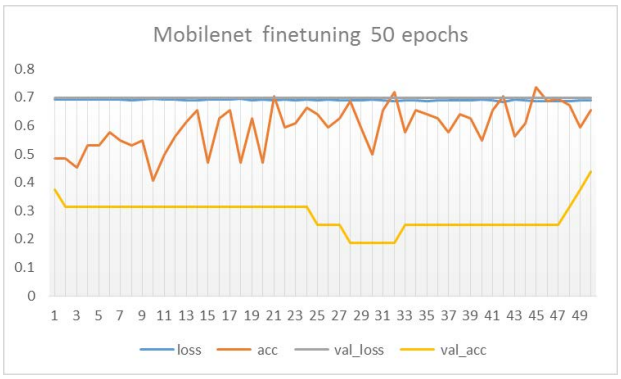
classes; and a possible approach to this problem might be training neural networks on visual cues.

5.4.3 Dedicated Learning and Transfer Learning. We also wanted to see how transfer learning as done by Gajarla et al. [66] compares with dedicated learning of CNN architecture. We chose four existing pre-trained models – InceptionV3, MobileNet, Resnet50, and VGG-16.

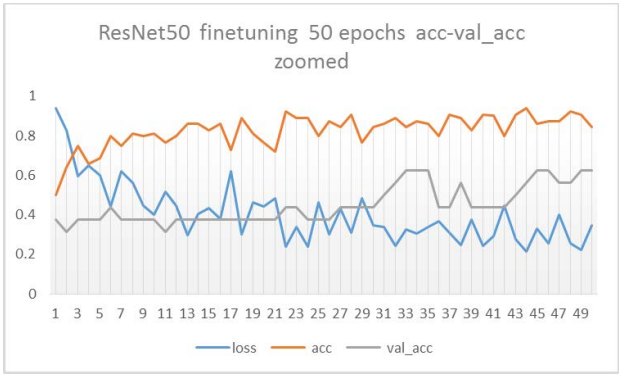
Deciding on the number of epochs for fine-tuning these pre-trained models poses a dilemma of choosing between overfitting model by too many iterations and settling for a poor model by too few iterations. To avoid overfitting, training should be stopped when *val_acc* stops increasing. To get this number of iterations for all four of the aforementioned pre-trained models, at first, we let them fine-tune their final layer for binary sarcasm detection problem with 50 epochs. Figure 5.8 shows their performance of fine-tuning over 50 epochs.



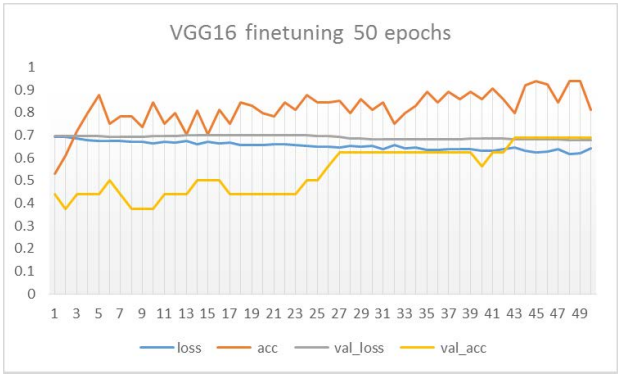
(a) InceptionV3



(b) MobileNet



(c) Resnet50



(d) VGG16

Figure 5.8: Transfer learning on pre-trained models for sarcasm detection.

For avoiding overfitting while transfer learning, we consider the model's accuracy at the epoch before the one when the value of *val_acc* decreased for the first time. For example, while finetuning InceptionV3, its *val_acc* decreased for the first time at 10^{th} epoch; hence, we consider transfer learning of this model upto nine epochs and consider its accuracy at 9^{th} epoch as its performance accuracy.

We get accuracy as follows in Table 5.2. As we can see, transfer learning does not achieve as high accuracy as we achieved with dedicated learning from the scratch, i.e., 84%. However, transfer learning models needs few epochs to obtain comparable performance before running into risk of overfitting.

Table 5.2: Performance of transfer learning models for sarcasm detection

Pre-trained model	**Number of epochs before decreasing "val_acc"**	**Accuracy**
InceptionV3	9	73.44%
Mobilenet	24	66.44%
Resnet50	6	79.69%
VGG16	6	75%

5.5 Discussion

Image-based cues have proved to be a useful feature to detect sarcasm on social networking sites. We collected data from Flickr using snowball sampling based on the tags associated with the images. We trained a network from scratch so that it is dedicated to sarcasm detection. This network performed with 84% accuracy. Then, we tested the merit of transfer learning for our problem scenario. We achieved 79.69% accuracy from transfer learning among the four networks that we fine-tuned. Our contribution in this chapter is the collection of an image-based sarcasm dataset YFS and the trained CNN model for sarcasm detection.

6 A MULTIMODAL APPROACH TO SARCASM DETECTION

Our participants during the qualitative study talked about how they use different kinds of cues in an SNS post to understand whether that post is sarcastic or not. Our study revealed the usefulness of different streams/modes of data for sarcasm detection. Themes emerged from our study extended an existing finding of prior sarcasm detection studies to multimodal level. Existing studies suggested that different sentiments as part of a single SNS post can denote that the post is sarcastic [27]. Our participants opined that in a sarcastic post, the sentiments in different parts of the post can be different. To elaborate on this, a sarcastic post can have a positive caption while the image conveys something very negative and vice-versa. This can also happen with comments along with the post itself – some users might not understand the sarcasm in the post and react opposite to what was expected, some might understand the sarcasm but want to play along with it, while some might react in straight message after understanding the sarcasm. Thus, a multimodal approach has the potential to be a more robust method of sarcasm detection.

In this chapter, we propose a multimodal sarcasm detection approach using text, image, and reaction emoticons. We utilized Facebook as the multimodal SNS platform for our data collection. However, we tried to keep the system generalized enough to be used for any other SNS platforms.

In this chapter, we discuss the structure of a Facebook post, sentiment analysis, auto image caption generation. Then, we discuss how we collected the multimodal dataset, followed by a discussion about the structure of our system. Finally, after model training, we report the results of our multimodal approach for sarcasm detection.

6.1 Background

6.1.1 Structure of a Facebook Post. Each social networking site has its own way of organizing information. Facebook has two types of communication: direct messages and general content posts. Both types of communication support several modes of content, including: text,

image, reaction emoticon, video, and audio. Direct messages are limited to a small number of people whereas general content posts target a range of group sizes from a small group of people to a general public audience. We limit our discussion to general content posts and their structures.

General content posts can be posted by a user in three different scenarios: (1) on a user's own timeline (personalized page for each Facebook profile), (2) in a group (a separate channel for a number of people with recommended maximum size of 5000 users), or (3) as a Facebook page (usually used for advertisement or promotion of any idea, individual or organization). An individual user can add up to 5,000 friends on Facebook with mutual acceptance, i.e., for that he/she needs to send friend requests to others or has to accept friend requests from other users. For a page, other users have to follow that particular page to receive updates from that, hence communication in this case is mostly unidirectional. In a group setting, all members can view and interact with each other irrespective of their connection status with each other as long as they are member of the same group, i.e., they do not need to be friends.

Posts from individual users' timelines can have different privacy settings – public, friends only, friends of friends, etc. However, posts from pages are always public. That means, any post from a page can be viewed by anyone, that includes users of Facebook, and people who do not have an account on Facebook.

As discussed earlier, a post can have text, images, and/or videos as content. There are three main ways to interact with a post: react, comment, and share. Facebook allows six reaction emoticons (as shown in Figure 6.1) to interact with a post with minimal effort to express a user's attitude to one post: like, love, haha, wow, sad, angry – the name of each emoticon by its name expresses what it is intended to express. Users can also comment on a particular post. Comments can also consist of text, images, and/or videos. On Facebook, a user can also share a content with his/her peers as long as it is permitted by the user who originally posted the content.

6.1.2 Sentiment Analysis. Sentiment analysis as a part of natural language processing (NLP) in computer science is well studied in existing literature. Sentiment analysis is "the process of computationally identifying and categorizing opinions expressed in a piece of text,

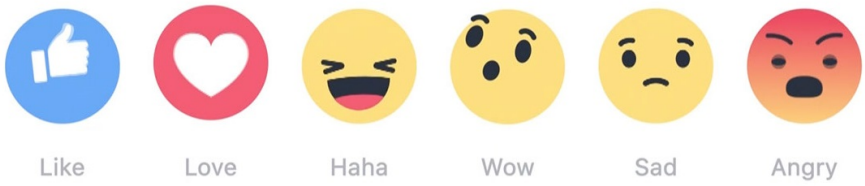


Figure 6.1: Reaction emoticons available on Facebook.

especially in order to determine whether the writer's attitude to particular topic, product etc. is positive, negative, or neutral" [73]. A major field of application for sentiment analysis is social media data. The role of social media and importance of understanding users' sentiment on this platform was repeatedly highlighted during political elections [74, 75], for instance. E-commerce sites are also a major application space for sentiment analysis. Companies on those sites use sentiment analysis to get an overview of customer reviews about their products [76, 77]. Sentiment analysis also goes by different other names, e.g., opinion extraction/mining, sentiment mining, subjectivity analysis, etc.

Sentiment analysis can be termed as the detection of attitudes, i.e., beliefs or dispositions towards something (objects or persons) based on one's emotions or sentiment [78]. Sentiment analysis largely focuses on text sentences or documents to identify such attitudes. The more common practice is to use simple weighted polarity of positive, negative, and neutral as types of attitude.

The baseline algorithm for sentiment analysis was proposed by Pang et al. [79]. It utilizes tokenization, feature extraction (e.g., with unigrams, bigrams, etc.), classification using Naïve Bayes, maximum entropy, and support vector machines (SVMs). Since sentiment analysis relies on the tokens used in the text, at best it can identify the surface or literal sentiment expressed by a text. Therefore, it cannot address the cases of sarcasm where people say something with the opposite intended meaning of what they express with the literal meaning. An example of such case is shown in subsection 8.1.2.

6.1.3 Image Auto-Caption Generation Model. Auto caption generation for images can be described as the problem of describing the objects in an image in natural language based

on the relation among the objects with respect to relative positions and actions. As the Figure 6.2 shows, an auto image caption generation model is expected to output one or more coherent sentences that describe what is happening in a given image.

Input Image	
Possible Caption	A group of people shopping at an outdoor market. There are many vegetables at the fruit stand

Figure 6.2: An example pair of input image and possible caption output. Example taken from Vinyals et al. [2].

In our work, we used the "Show and Tell" auto image caption generation and Neural Image Caption (NIC) model proposed by Vinyals et al. [2]. It is a generative model based on a deep neural network architecture. The architecture has two parts. First, images are passed through a deep convolutional neural network (CNN) that was pre-trained for image classification. They use a CNN at the first phase, replacing the earlier practice of using recurrent neural network (RNN), since it has been convincingly shown over the last few years that CNNs produce better representations for images using a fixed length vector embedding [80]. The encoded results from the last layer of the CNN are fed to a language generating RNN decoder to perform machine translations and generate captions for images.

6.1.4 Common Machine Learning Algorithms. In this chapter, we will use several machine learning algorithms. Among those, random forest and multi layer perceptron were dis-

cussed in the chapter 4 (on our text-only approach). At this point, we will discuss several other common machine learning algorithms briefly.

6.1.4.1 Support Vector Machine. Support vector machine, widely known as SVM, is a commonly used machine learning algorithm. Before ubiquity of fast computing resources, SVM was thought to be one of the best machine learning algorithms. SVM, a supervised learning algorithm, works with an objective of drawing a classifier hyperplane/decision boundary among different classes while keeping the largest possible margins with close to borderline instances on either side of the plane. These margins on either side of the decision boundary are called support vectors.

As depicted in Figure 6.3, both hyperplanes–solid line and dashed separate the instances of both classes correctly. However, it is apparent that B1 will be more likely to perform better than B2 if we are given any new instance of the input data. This is because, B1 has a larger margin than B2. SVM works to optimize these two goals simultaneously–drawing decision boundary, and maximizing margins. SVM can be used for both linear and non-linearly separable classes. In fact, non-linear SVM are done by converting the problem into a linearly separable problem with a method called Kernel trick.

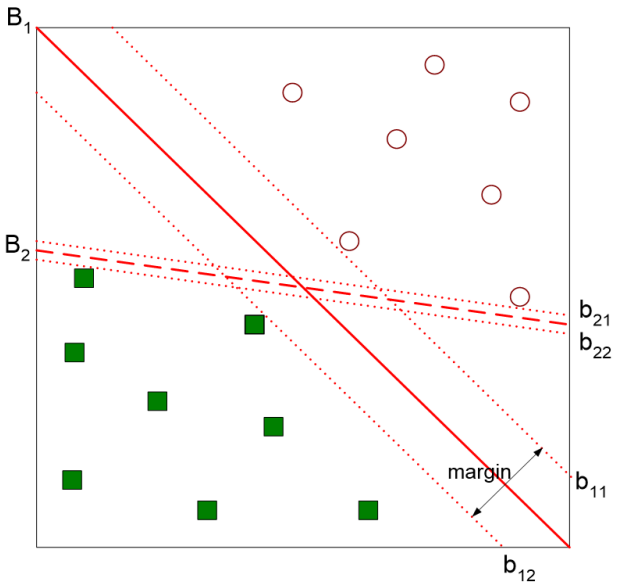


Figure 6.3: Comparison between two SVM decision boundaries in a two-class problem setting. Image taken from [3].

6.1.4.2 Adaboost. As we discussed in 4.1.4.3, a combination of multiple weak classifiers can produce a single strong classifier. In Adaboost, each weak classifier is trained using a random subset of the total training dataset. These subsets can overlap. We can assign weights to the training instances that determines their probability of appearing in the training subset. After training a classifier, the weight of a misclassified example is increased so that the probability of its appearing more frequently in next round of training increases and classifier in that round learns to classify that large portion of examples correctly. After each classifier is trained, it is assigned a weight. A classifier with accuracy equal to random guessing is assigned zero weight, one with higher accuracy is assigned a positive weight, and one with worse accuracy than random guessing gets a negative weight.

6.1.4.3 Gaussian Naïve Bayes. We discussed the basics of Bayesian classifiers and its naïve Bayes variant in 4.1.4.2. We concerned ourselves with discrete values in that subsection. To deal with continuous values, we need to use Gaussian naïve Bayes classifier (GNB).

GNB assumes that the continuous values associated with each class follow a Gaussian distribution. The probability of an instance belonging to class C_k to have continuous attribute value $x = v$ is,

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi{\sigma_k}^2}} e^{\frac{(v-\mu_k)^2}{2{\sigma_k}^2}}$$

where, μ_k is the the mean and ${\sigma_k}^2$ is the variance associated with class C_k.

6.2 Dataset Collection

To the best of our knowledge, there has been no prior work with implementation of multimodal approach to sarcasm detection with a publicly available dataset. Therefore, in order to experiment with multimodal approach for sarcasm detection, our first challenge was to build a multimodal dataset. We aimed to build this multimodal dataset with publicly available data, and make this dataset publicly available (link listed in Appendix E).

6.2.1 Data Source Selection. Now, the question arises that what we can use as the source of our data. We chose to collect data that has multimodal information (e.g. text, image, etc.). Again, the focus application area of this research is the social networking sites. Therefore, we evaluated several social networking sites. Facebook, which has the largest number of users from around the world, allows posts with different modes of data. Many of these posts happen with public view settings. Therefore, in the regard of data collection, we considered Facebook as a good and viable source of data.

The next challenge is to know the differentiation of categories of posts–sarcastic, and non-sarcastic. While independently labeling data by researchers or by crowdsourcing is a common and popular approach for dataset preparation, we used self-annotation on posts. Many previous works used hashtags (e.g. #sarcasm) as indicator of self-annotation of sarcasm on twitter. We extended that using similar words with snowball sampling while collecting our image-based dataset. However, instead of using snowball sampling [35], in this phase, we adopted word-embedding for choosing the potential similar words. We consider the word "sarcasm" and "information" as representative of sarcasm and non-sarcasm classes respectively. We used word2vec model to identify potential synonyms of the word "sarcasm" based on their embeddings or vector representations. Details about this process has been discussed in chapter 4.

Facebook offers a variety of privacy settings for posts at individual users-level, and the users might want to change their privacy setting for a particular post that will not be reflected in a collected dataset. Again, collection of these posts from individual users depend on the connections of researchers' social connections on Facebook. Thus, the value of posts from individual users decays for collecting research oriented datasets. On the other hand, posts from Facebook pages are publicly viewable. However, to put a response to an existing post or to post a new content, one has to have an account with Facebook. Nevertheless, anyone can view these posts. Public posts from Facebook pages have multiple modes of data (e.g., text, images, reaction emoticons, etc.). Thus, these satisfy our criteria for data collection.

We chose several Facebook pages for collecting public contents posted by those. All

pages we selected had at least one million followers. We decided on this threshold because it emphasizes that at least one million users on Facebook endorse that these pages are serving a purpose as their names suggest. For collecting sarcasm class data, we chose the pages that had the word "sarcasm" or any of its synonyms in their names. We also verified the list of such pages' contents for validity of our assumption of those publishing sarcastic contents. We consider such pages as sarcasm related pages. The ten pages we selected for that are following: "Mother of Sarcasm," "Sarcasm," "Sarcasm Society," "Sarcasm Daily," "Sarcasm, Because Killing People is illegal," "Sarcasm World," "I speak sarcasm as a 2nd language," "Sarcasm Hub," "Sarcasm Sodality," and "Sarcasm Meets Humor." Now, we need to decide on the pages that can serve as sources for instances of non-sarcastic posts. It is fair to assume that news related posts are not sarcastic. Though there are some satirical news portals, and some news sources that are not much reliable, we can safely argue that mainstream news media do not spread sarcasm as form of news. Therefore, we considered verified Facebook pages of ten popular mainstream news media as sources of non-sarcastic posts. The pages that we selected are: "The New York Times," "Time," "The Economist," "The Economist, Asia," "The Times of India," "Hindustan Times," "BBC News," "CNN," "The Wall Street Journal," and "Reuters." The threshold of at least one million followers created a good distribution in choice of non-sarcastic pages from different parts of the world that eventually will result in better training.

6.2.2 Collection. We collected the data using Facebook Graph API. Our dataset preparation time was 1 July, 2018 to 3 July, 2018. This time period is after Facebook adopted the GDPR guidelines.

For creating a dataset with mutlimodal data, we collected the description, message, images (if any), reaction emoticons of posts, and comments on those. We did not collect any users' identifying information. We only collected the information/contents that were posted with "public" privacy settings.

At first we collected all data from the date when a page was first created until 1 July, 2018. However, among all the modes of data associated with any post, reaction emoticons are rel-

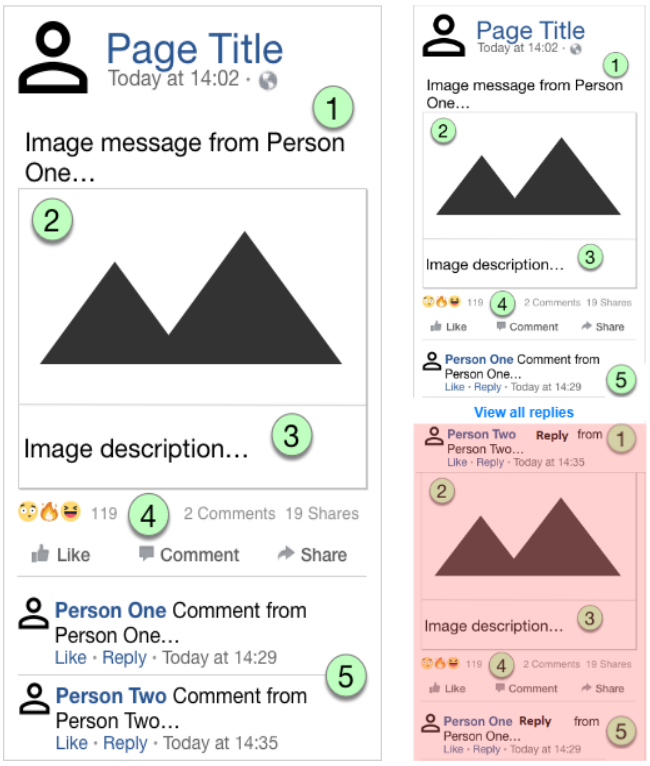
atively new. It was first introduced as a feature in Facebook on February 24, 2016. Thus, posts before that time does not have this information unless a handful group of users decided to re-visit any particular post. Again, it is reasonable to assume that after first launch, users might need some time to get familiar with a new feature. Hence, we consider the rest of February 2016 as a period for users to get familiar with reaction emoticons. Therefore, while preparing the dataset, we preserved the contents posted after February 2016.

In total, we collected 20,120 instances of sarcasm category posts, and 21,230 instances of non-sarcasm category posts. Our dataset with 48.65% sarcastic posts, and 51.35% non-sarcastic posts in it can be identified as a balanced dataset. All of these posts had a set of reaction emoticons, and description/message associated with it while a large portion of those (98.26%) included images.

6.3 Methodology

In our multimodal dataset, there are three modes of data. These are: text, images, and reactions numeric (see Figure 6.4(a)). Now, let us discuss from where these data come from in a Facebook post. As we discussed in the background section in this chapter 6.1, each Facebook post has a description (labeled 3 in Figure 6.4(a)) or a message (labeled 1 in Figure 6.4(a) or both). Usually, a post also receives one or more comments (labeled 5 in Figure 6.4(a)). Though posts without any comments might be seen, it is not the common scenario since comments are the one of the main ways to interact on a public post on Facebook. For most posts, there is an image associated with it (labeled 2 in Figure 6.4(a)). Just like as comments, a post on Facebook usually receives different reaction emoticons from the users who want to interact on the post (labeled 4 in Figure 6.4(a)). Facebook has a pre-defined set of reaction emoticons available on their platform (see Figure 6.1).

Both the original post and comments can be made up with combination of text, images, videos, and reaction emoticons. Facebook also allows users to use react emoticons, comments (called replies), and share (except certain cases) on an existing comments itself. Thus, a com-



(a) (b)

Figure 6.4: (a) Sample of a Facebook post. (1) Message of the post; (2) Image of the post; (3) Description of the post; (4) Count of users' reactions to the post; (5) Users' comments on the post. (b)Symmetric structure of posts and comments. Replies are excluded for making the system work with both post and comments separately.

ment on a post can be thought to be a post itself, and Facebook assigns an unique ID for each post or comment. In our study, we used this concept of symmetric nature of post and comment. If we consider a comment to be a post, replies can be considered as comments on it. Since our objective in this phase is to detect posts that convey sarcastic intent, we wanted the approach to be applicable to comments as well. We excluded the replies (shown with red color shaded area in Figure 6.4(b)) from our computation while working with the original posts so that those can be treated as parts of a comment (second level post), and thus, a symmetric pattern emerges.

We hypothesize that the message/description, image or any other part of the post might not be sarcastic on its own, however all together, they might convey sarcasm. In this phase, we are trying to study whether a post as a whole conveys sarcasm.

6.3.1 Pre-processing Reaction Data in Facebook Posts. A concern is that the reactions received on a post varies with the reach of that post, i.e., to how many users Facebook showed that post. Since the algorithm that Facebook uses to organize end-users' newsfeed is not known, there have been some studies to study users' reasoning in this regard [37, 81]. These studies present some perspectives: reach of a post depends on who posted that online, how old the account from which that content was posted, how much response the previous posts from that account received. Taking these factors into account, we chose to use normalization to eliminate the factor of the age and popularity of the account to post the contents and time of post. We divided the number of each reaction emoticon received by the total number of received reactions on each post to remove the bias created by the post's reach.

6.3.2 Sentiment Analysis of Text Data of Facebook Posts. Existing works suggest that sentiment can be a useful factor in sarcasm detection [9, 20, 24]. We also used sentiment of text data as a component in our multimodal system. For this, we considered two properties: subjectivity and polarity of textual data. Unlike our work in chapter 4, we did not use IBM Tone Analyzer based emotional trajectory approach in this phase because most of the text information (e.g., description, message) are short, and have one to only few sentences that might not be useful to form a wave-like trajectory of tones.

Subjectivity means the characteristics of text to express a user's sentiment, feeling, or opinion. Thus, the subjectivity value of a piece of text represents how much sentiment related information is available with that text. Polarity denotes whether the text information yields positive or negative sentiment. There are several sentiment analysis tools available, for example, TextBlob [60] and Vader [82]. We used TextBlob for determining subjectivity, and polarity. Using this tool, subjectivity is measured in a scale of [0, 1], where a text with subjectivity equals or near to zero does not contain much information about a user's feelings (e.g., names of user's friends tagged as text in comment section of a post). On the other hand, polarity is measured in a scale of [-1, 1], where a text that has a polarity value less than zero is expected to have a negative sentiment, while a greater than zero value of polarity means the text writer user's positive

sentiment. To understand subjectivity and polarity better, we use a comparison between sentiment analysis output for two different statements. Let us assume, our two statements are: text1 = "Bangladesh is a small country", text2 = "Bangladesh is a beautiful country". It is safe to say that text2 is more expressive about one's feelings about a country, Bangladesh. This is also reflected by the output from TextBlob sentiment analysis – text2 having higher subjectivity than text1 ($1.0 > 0.4$), and text2 being more positive statement than text1 (polarity $0.85 >$ polarity 0.25).

Though message, description, caption of image, and comments – all are text based data, there is a difference between comments and the former three. Whereas a post can have at most one for each of these – description, message, and caption of image, a post can have more than one comment. Therefore, for the former three data sources, we can have only one polarity value. We determined the polarity and subjectivity of the text for these. Thus, we got sentiment based features values from textual data. For calculating the features value for all comments in a post overall, we calculated the sum of subjectivity scores, sum of all positive sentiments (when polarity >0), sum of all negative sentiments(when polarity <0) of all comments to obtain them as three individual features.

6.3.3 Utilizing Image Data in Facebook Posts. We used two aspects of image data included in a Facebook post. We used visual cues in images to know the probability of that being sarcastic. Second, we used the semantic representation of the images to compare against the textual information available along with it in the post.

6.3.3.1 Sarcastic Visual Cues Detection. We argued and presented our experiment on how visual cues in images can be used to identify sarcastic images in chapter 5, and in Das et al. [72]. We used our CNN-based sarcasm detection model that can detect sarcasm using an image's visual cues with 84% accuracy. If a post does not have a description or message associated with it, the image becomes the only content posted by the user and a major medium for determining whether the post has a sarcastic intent. We pass images associated with each post through the CNN, developed and trained with our Yahoo Flickr Sarcasm (YFS) dataset, to get the probability of that image to have sarcastic cues in it. We call this value the "CNN score".

6.3.3.2 Auto-Caption Generation. Schifanella et al. [33] discussed the importance of considering visual and textual aspects of contents on social media to detect sarcasm in them. They used semantic representation of the images only. However, a potential flaw of such representation is that it is less capable to express the sentiment expressed by the image. We argue that an image caption can both provide semantic representation and provide us with a hint about the sentiment expressed by the image. As we discussed earlier, sentiment is important information for sarcasm detection.

For automatically generating captions for images, we used the model proposed by Vinyals et al. [2]. We trained the model using the COCO dataset [83]. Our training dataset for this Show-and-Tell image captioning component had 118K images of total size 18 GB, and the validation dataset had 5K images of total size 1 GB. We also used the corresponding annotation data from their website `http://cocodataset.org` of size 1.1 GB and 821 MB respectively. We limited the maximum token count of auto generated caption up to 30.

With the help of this auto-caption generation component, each image now has one model-generated caption. Besides, each image might have a user-assigned caption with it as message/description. We can hypothesize that for a non-sarcastic post, the sentiment in user-assigned caption and the auto-generated caption will be almost same. On the contrary, for a sarcastic post, the user-assigned caption and the auto-generated caption are likely to have sentiments of opposite polarities. For example, let's assume a post has a user-assigned caption: "I had a WONDERFUL day!", and it has an image associated with it of a crying person. We can easily understand that we discussing about an instance of sarcastic post. For the image in this post, the auto-caption generation component is likely to generate a caption like this: "A person is crying". Here, positive sentiment in the user-assigned caption, and negative sentiment in the auto-generated caption have two different polarities that is likely to be a useful hint for classifying sarcasm and non-sarcasm categories of posts.

6.3.4 Model Training. From our collected dataset, we calculated values of sixteen features listed in Table 6.1 for each data instance. For any possible missing value of a feature (e.g.,

unavailable values of auto caption polarity, subjectivity, CNN score if there is no image contained in that post), we used the average value of that feature as the placeholder/representing value. A diagram of our multimodal sarcasm detection feature extraction system is shown in Figure 6.5.

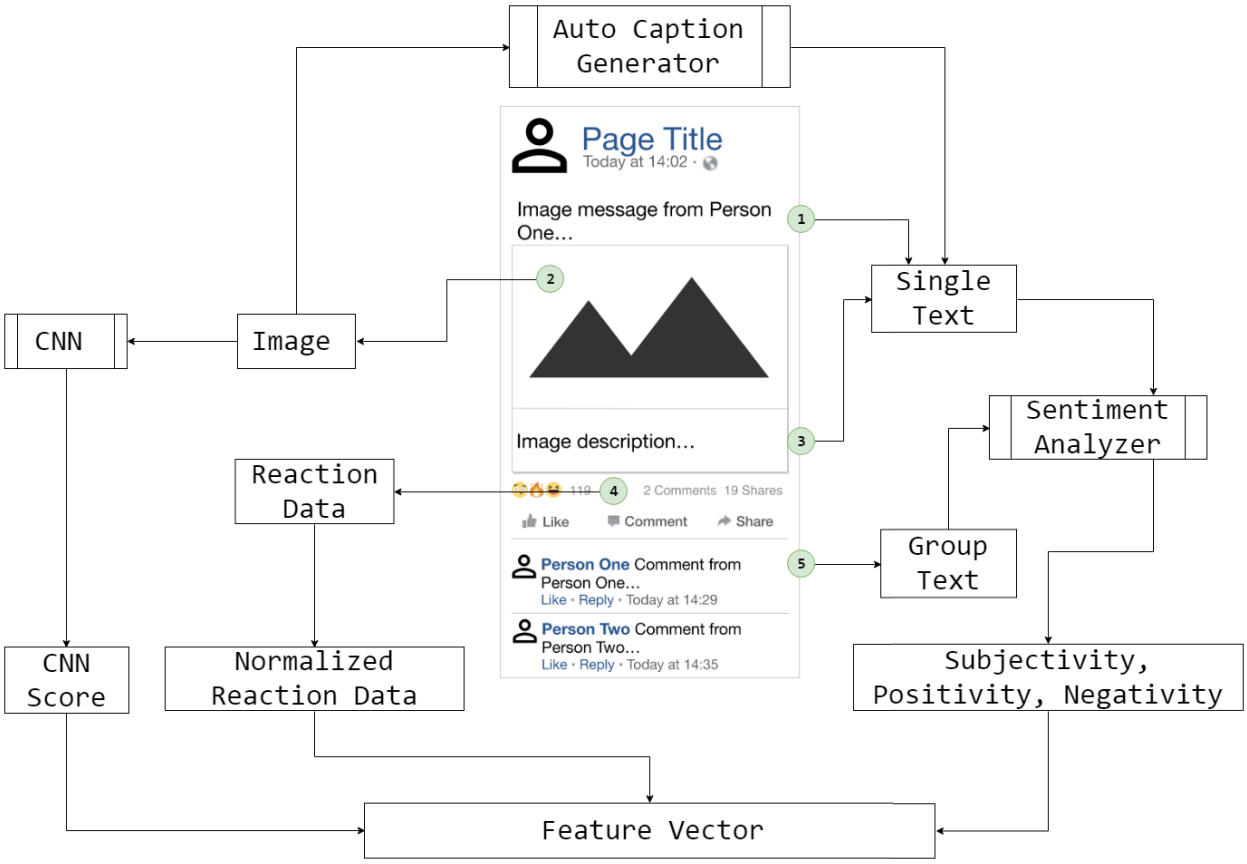


Figure 6.5: Feature value extraction for multimodal SNS post for sarcasm detection.

We used ten-fold cross validation for validating our models. We used five supervised machine learning algorithms as follows: support vector machines (SVM) with linear kernel; two ensemble algorithms: Ada Boost with Decision Tree classifier of depth 1 and Random Forest; Multi Layer Perceptron (MLP); and Gaussian Naïve Bayes. A high level overview of the model training process can be presented as in Figure 6.6.

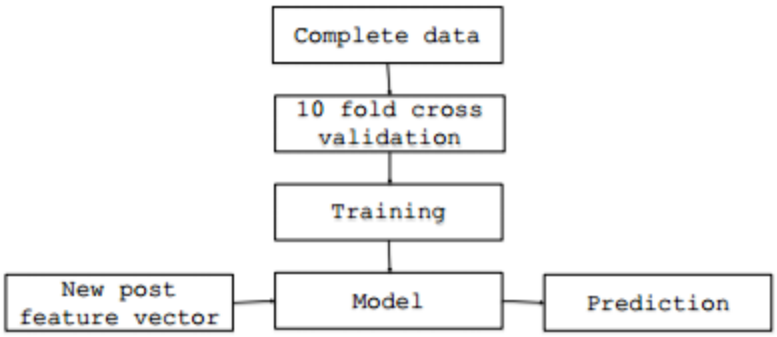


Figure 6.6: Supervised Model Training Process and Usage Diagram

6.4 Results

Evaluation of our system has two parts. Since we discussed a multimodal system based approach to sarcasm detection in this chapter, it is imperative to know which feature contribute to what amount to decide on the classification of the posts, besides the performances of different models. While presenting the performances of different models, we also present reasoning behind the algorithms' performances.

6.4.1 Contribution from Different Features. In total, the system uses sixteen different features to classify a post whether it is sarcastic or non-sarcastic. Among these features, only counts of reaction emoticons (like, love, haha, wow, sad, angry) are specific to the platform/source of our data collection (Facebook). The rest of the features are general to any social networking site. To rank the features according to their contribution, entropy based information gain can be used. The higher the information gain for one feature, the more it might be useful for classification machine learning algorithms [84, 85]. Table 6.1 shows the entropy based information gains for the features used by the system.

Here, we can see that the distributions of reaction emoticons is highly useful while deciding the class of a post on Facebook. High information gain for most of the features' values in that mode of data tells us that posts in sarcasm and non-sarcasm classes have different distributions in the reaction emoticons they receive, while the distribution is mostly similar for posts in the same class. Again, we can see the features sourced from comments have high information gain, i.e., more usefulness to classification of posts. This reemphasizes our observation in qualitative

Table 6.1: Information gains of features

Mode of Data	Feature	Information Gain
Reaction emoticons[1]	angry	0.3217
	haha	0.4904
	like	**0.5534**
	love	0.4275
	sad	0.3328
	wow	0.4493
Image	auto caption polarity	0.0174
	auto caption subjectivity	0.0173
	CNN score	**0.0263**
Text	comments negativity	0.2503
	comments positivity	0.4185
	comments subjectivity	**0.4626**
	description polarity	0.0237
	description subjectivity	0.0253
	message polarity	0.1825
	message subjectivity	0.2044

study that comments help understand the context of the posts while trying to identify if a post is sarcastic or not. In fact, value wise, comments subjectivity has the third to the highest information gain. Again, messages associated with posts have a useful role in this classification task. If we look at the information gain of the image based features, we have a reasonable observation. While CNN score of the images have a moderate usefulness to understand whether a Facebook post is sarcastic or not, auto caption polarity and auto caption subjectivity are not as useful as it. In fact, these two features are the least two useful features in our system. A possible explanation behind this might be that as the image caption generation model is a comparatively new and less developed area, and there is no definitive way to measure the performance of such a caption generation model, this component of our system cannot be guaranteed to be fully optimized. Thus, the feature contributions from this component might not be as useful as the other features used in the system.

6.4.2 Performances of Models. We present the accuracies of the five algorithms mentioned in the previous section in Table 6.2. We also used some stochastic algorithms that rely on some randomization inherently. Therefore, we repeated those algorithms 25 times, and calculated

their mean accuracy to present more reliable measure. We can identify the stochastic algorithms by their non-zero value of standard deviation in the table.

Table 6.2: Applied machine learning algorithms, accuracies with standard deviations

Algorithm	**Accuracy**	**Standard Deviation**
Ada Boost	90.61	0.00
Gaussian Naïve Bayes	73.66	0.00
Multi Layer Perceptron (MLP)	92.06	0.19
Random Forest	93.11	0.196
Support Vector Machine (SVM)	88.39	0.00

In this phase of our study, we used a bag-of-features approach. Usually in such approach, ensemble machine learning algorithms perform well. In this algorithm family, a weak classifier can be trained using each feature in our system. Using the gathered decisions of all weak classifier, we can expect to get a more robust and accurate prediction from a good high level ensemble classifier. This is also reflected by the performances of the two ensemble algorithms we used – Random Forest and Ada Boost. Both of these showed convincing performances ($>$90%) for sarcasm detection. Again, we also evaluated the performance of a multi layer perceptron (MLP) for the task of sarcasm detection. Since we are using a supervised approach for sarcasm detection, and we already have some deep neural network based components in earlier stages of our multimodal system, we used an MLP with a small number of layers and nodes. Despite limited number of layers and nodes, the MLP based approach performed well with $>$90% accuracy. However, the performance of SVM is not as good as the ensemble based and MLP based approaches. As we discussed earlier, naïve Bayes is a family of algorithms and widely used for text classification, and sentiment data analysis. We also saw its applicability in the text-based approach chapter. Since the features used by our multimodal system have continuous values, we chose to use Gaussian Naïve Bayes algorithm. However, among all the algorithms that we used for supervised approach to sarcasm detection, this performed the worst.

6.5 Discussion

Performances of our multimodal models showed the superiority of a multimodal approach over unimodal ones. We achieved highest 93.11% accuracy (std. dev. = 0.196) with random forest classifier using multimodal data [86]. It is higher than accuracies of both of our text-based approaches – tone-based model (82.5%) and improved naïve Bayes based model (75.8%). Multimodal approach has also achieved higher accuracy than only image based approach (84%). Our experimental result supports our qualitative finding.

Since we utilized only three modes of data and these made the sarcasm detection approach more robust, a future direction can be inclusion of more modalities into the system for better performance.

7 RECREATING AND STUDYING THE ATTENTION MODEL OF SARCASM IN VIDEOS

Our participants during the qualitative study agreed that sarcasm helps a particular content get popular on social media. According to them, it works as a cycle – a particular content gets popularity, that content is used as template for other new sarcastic posts, the template gets more popularity, and it goes so on. We also observed similar trends in our image based model training results. Thus, we became interested to study if similar trend exists for videos and what are the objects that are usually looked at in sarcastic videos. In short, we were intrigued to study the attention model of human users in sarcastic videos. We used two deep learning based approaches – regression and semantic segmentation to experiment whether we can replicate the attention model of human user with deep learning. We also studied what are the objects that are looked for while being attentive to sarcastic videos.

In this chapter, at first we discuss some background about deep learning concepts. Then we discuss data collection process and our experiment methodologies. Finally, we reflect on our findings about the attention model in sarcastic videos and usability of deep learning in this context.

7.1 Background

7.1.1 Regression. Regression is a supervised learning approach that tries to estimate a real value for given input. It tries to find out a real valued function between a set of independent variables and a set of (usually one) dependent variable(s). There are three types of regressions.

1. *Simple regression model* attempts to fit a linear regression model with a single independent variable.

2. *Multiple regression model* attempts to predict a dependent variable based on the values of two or more independent variables.

3. *Multi-target regression model* attempts to predict multiple dependent variables on values of a set of independent variables.

7.1.2 Semantic Segmentation. Semantic Segmentation [87] is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. From a classification perspective, a single class is assigned to the whole image. In our case, all the image frames belong to a single class of sarcasm. However, attention wise, the pixels in these frames can be assigned to different classes: gazed (attention given pixels) and non-gazed (not attention given pixels). Semantic segmentation classifies every pixel of the image to one of these classes.

7.2 Dataset Preparation

We collected and prepared our own video-based dataset. Then we recorded the gaze points and generated the corresponding gaze videos. We also processed the video data in ways that would be useful as inputs to neural network-based architectures.

7.2.1 Video Data Collection. We collected 50 short sarcastic video clips from popular TV series: Friends (20), Silicon Valley (10), The Big Bang Theory (10), and Two and a Half Men (10). All the videos were collected from YouTube. All videos were tagged with "sarcasm". As a double checking on the labeling on the data, the student researcher and another student helping in labeling of the data agreed upon whether the video contained sarcasm or if it was only entertaining. To abide by the limitation imposed by the gaze recording software (discussed later in this section), we only used video clips that are up to 1 minute in length. However, we also ensured that the sarcastic incident in the video had enough context information in the video. This created a variation in the lengths of our collected videos ranging from 45 seconds to 1 minute.

7.2.2 Gaze Labeling of the Data. Like many supervised learning approach-based studies, we needed to label our dataset. In our case, these labels were gaze points of a person on the videos. There are two ways to record these gaze points: dedicated hardware based or a webcam based.

The Tobii eye tracker is popular among the affective computing community for its higher precision. However, due to their high cost, many studies utilize webcam-based solutions like Gaze Recorder, Ogama, or Gaze Parser.

In our experiment, we used Gaze Recorder. This software package allows users to configure several settings including degrees of field of view (FOV), resolutions, gender, adaptive/non-adaptive extend time during static scenes, and how much change in frames would be enough to consider a frame as a new one. In this software, one needs to record gaze data in this order: start camera → initialize face → calibrate gaze → play video → record → generate results. Figure 7.1 shows the interface and settings of the Gaze Recorder software.

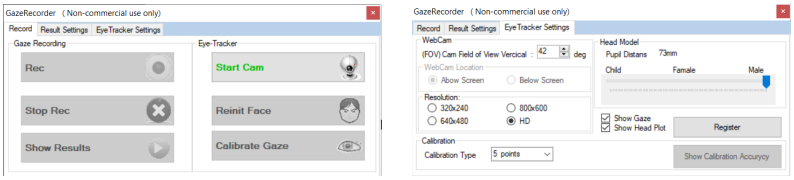


Figure 7.1: Interface and available configurations of Gaze Recorder

After we recorded the gaze points for all videos, we ended up with a collection of pairs of videos with 50 original videos and 50 gaze labeled videos. Figure 7.2 shows an example pair of frames from original and gaze labeled videos.

Figure 7.2: Example pair of frames from original and gaze labeled videos.

7.2.3 Locating the Gaze Point. We located gaze points in the frames of gaze labeled videos by subtraction. For any frame after the first one, subtraction result of each frame and the previous one of that frame gives the gaze point. Though this approach works well for most cases,

this faces issues when there is sudden and drastic changes in video frames. We call the result frame obtained in this way as subtracted gaze frame, as shown in Figure 7.3. In heat map, Red denotes the region where the spectator gazed for a longer time, and green for short period of time.

7.2.4 Preparing Final Dataset. At this stage, original video and the result video with gaze points had different fps for Adaptive FPS due to (a) not enough frame change, and (b) ambient lighting. This resulted in unequal numbers of frames generated from original videos and gaze recorded videos. We discarded frames from the videos with larger fps keeping the ratio between the numbers of read frames from original and gaze video equal to 1.0. Then we passed each subtracted gaze frame for finding the gaze point coordinates.

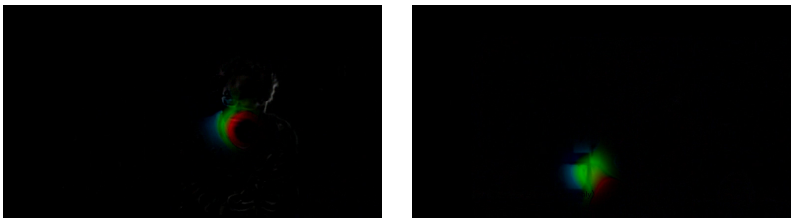
In end dataset, we have 31,307 frames in total. Each frame has a size 1536 x 864. The frames are in RGB having values ranging 0-255. We could not scale the RGB values in [0, 1] range because of memory constraints of the used GPU since usual OpenCV unsigned integer frames have a size of 1 byte per pixel per channel whereas converting frames into scale of 0-1, i.e. float requires 4 bytes per pixel per channel. Thus, this increases the memory requirement by four times.

7.3 Methodologies

7.3.1 Regression Based Approach.

7.3.1.1 Calculating the Gaze Point Coordinates. To find the gaze center, we first try to use a popular OpenCV HoughCircle function [88]. It takes an image as an input and return all the existing circle in the image. However, when we try with our image it does not return any circle. After careful analysis, we find out few issues that are the more likely reasons for HoughCircle function not returning any circles. First, most of the input images do not contain proper circular shape as shown in Figure 7.3(a). They are often half circular shape. Second, different color circle overlapped with one another. Third, image quality was poor.

Next we try with depth first search for finding out circle. But it failed when the image contains circles that have discontinues RGB value as shown in Figure 7.3(b).



(a) (b)

Figure 7.3: (a) Example of half circle shaped gaze points that could not be detected by HoughCircle function (b) Example of discontinuous RGB areas that could not be located correctly with DFS.

Finally, we come up with our own approach for finding circle in the image. First, we figure out which color circle we are searching. After RGB value calibration, we understand that all the images contain three colored color. Their RGB values are following: Red (R >60, G <10, B <10), Green (G >60, R <10, B <10), Blue (B >60, G <10, R <10). We come up with a simple approach called first and last point finding. We scan each image row by row and keep the first and last point of the desired color. Then using those two points as a diameter of a circle we draw circle. Our result shows that our algorithm can detect circle with a very good accuracy.

7.3.1.2 Regression Network. To convert a traditional image classification CNN to a image regression CNN, we needed to do the followings:

- Remove the fully-connected softmax classifier layer typically used for classification
- Replace it with a fully-connected layer with a single node along with a linear activation function.
- Train the model with a continuous value prediction loss function such as mean squared error, mean absolute error, mean absolute percentage error, etc. We used root mean square in our experiment, customizing it to handle two variables (x, y) as parameters.

At first, we started with a network written from scratch. The inputs to the network were original video frames and gaze center points coordinates. We used standard data augmentation techniques like shearing, zooming, etc. We optimized the network using Adam with learning

rate = 0.003, decay = 1e-3/200. We trained the network on NVIDIA 1060 using 20 epochs. However, we did not achieve much promising values of RMSE metric. Therefore, we turned to transfer learning. We used VGG-16 as the backbone network with their trained weights on ImageNet dataset [89]. After 20 epochs, we achieved performance as in Figure 7.4 and Table 7.1:

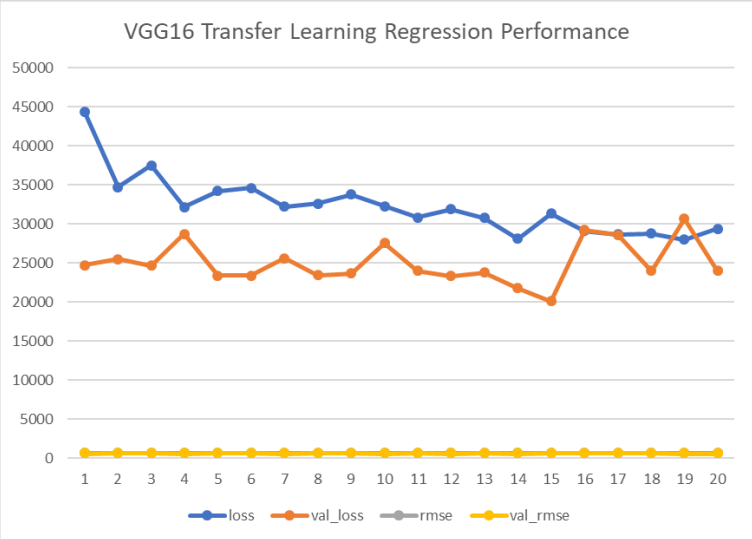


Figure 7.4: Performance graph of transfer learning with VGG-16 for recreating the sarcasm attention model.

Table 7.1: Performance of regression approach for recreating attention model of sarcasm

metric	**score**
loss	29406.0408
rmse	666.9297

As evident by the large value of RMSE, we can understand that the regression based approach might not be well suited for recreating attention model of sarcasm. We also looked at the changes of the rmse and loss value with increasing epochs. The values were not decreasing with epochs by any considerable amount, if they were improving at all. We took this as an indication that allowing the network for more epochs might not help us to achieve better performance.

7.3.2 Semantic Segmentation Based Approach. While locating gaze points by subtracting subsequent frames, we removed the non-gazed areas of frames as background. The gazed regions of frames can be thought as one segment, and the non-gazed regions as another. We used black and white colors for binary labeling of pixels to have semantic segmented images.

We used 25,308 frames from first 40 videos in our dataset for training and 5,999 frames from rest 10 videos for testing. Besides, we also used standard data augmentation methods.

We started with a pre-written U-net from a github repository[1]. The inputs to the network were original video frames and binary colored segmented video frames. We optimized the network using Adam with learning rate 0.004. We used binary cross entropy loss function and accuracy as metric for training the network. At first, we trained the network with 20 epochs and 300 steps per epoch. This did not give us satisfactory performance. The prediction for all test images contained random black and white pixels resulting noisy gray images as outputs. Then we increased the epoch to 50. This helped improve our performance. After 50 epochs, our results are as in Figure 7.5 and Table 7.2:

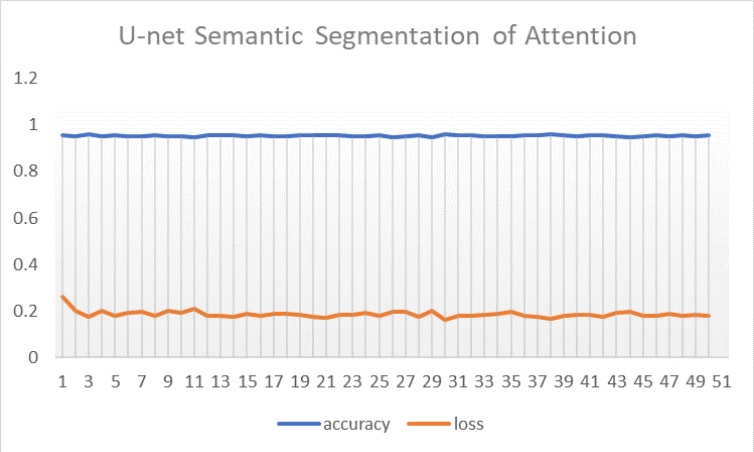


Figure 7.5: Performance graph of semantic segmentation approach for recreating the attention model of sarcasm in videos.

Though metric-wise, these results seem promising, when we looked at the prediction by

[1]https://github.com/zhixuhao/unet

Table 7.2: Performance of semantic segmentation based approach for recreating attention model of sarcasm

metric	**score**
loss	0.1781
accuracy	0.9542

the network as images, as in Figure 7.6, we can see the networks performance was not enough to pinpoint the gaze point. Rather it identifies a larger area as the attention area that includes both the original gaze point area and also some non-gazed or non-attention gained background areas.

(a) (b) (c)

Figure 7.6: Performance of semantic segmentation approach as images (a) original output (b) inverted output (c) inverted output with increased contrast for better view.

7.3.3 Object Location and Distance Based Approach. We did not obtain promising performance from our regression based approach. Though the semantic segmentation based approach achieved numerically promising result, it was not easily comprehensible and also has room for improvement. At this stage, we modeled the experiment in a different way. We wanted to see what objects are often looked for in a sarcastic video. In other words, what are the objects that are often close to the gaze points on the videos.

We passed the original frames in the dataset to a YOLO object detection model [90]. We identified the objects in the frames with their corresponding locations. YOLO can detect 80 different objects. If there is no object in a frame that can be detected by a YOLO network, we used a default object “unidentified” with default location (0, 0). If the frame had multiple objects that YOLO model can identify, we saved all of the objects and their locations.

We calculated the center coordinates of gaze points as described in our regression based

approach subsection. Then, we calculated which object's location is the closest to the center of gaze point for a particular frame. We showed the top five objects that are seen near the gaze points in sarcastic videos, in Table 7.3.

Table 7.3: Top five objects that were closest to the gaze center points

Object	**In how many frames it was the closest to the gaze center**
Person	25,662
Unidentified	1,247
Refrigerator	704
Sofa	535
Tie	412

7.4 Discussion

Our experiment with deep learning in attempt to understand the attention model of users in sarcastic videos has two folds – first, we wanted to recreate the attention model with regression and semantic segmentation; second, we studied what objects are often looked at in sarcastic videos. Our experiment suggests that semantic segmentation has more potential to recreate the attention model of sarcasm than regression based approach. Our experiment with attention model in videos also showed contradictory results to our image based sarcasm detection approach. While our image based sarcasm detection approach showed human as a non-sarcasm object, our attention model based experiment result says, humans are the centers of gaze in most sarcastic video frames. However, there is a threat to the experiment conducted in this phase. We used video clips from directed TV series that might be very different from the real life occurrences of sarcasm that are more likely to be seen on social media. This might be the reason of the contradiction between the results from our two experiments.

8 CONCLUSION

Social networking sites (SNSs) are popular among users of different age, nationality, cultures, and languages. The availability of large amounts of user interaction data has also boosted sentiment analysis research. Despite major breakthroughs in affective computing, sarcasm is a less studied area. In this thesis, we focused on sarcasm detection and satire in several domains. Unlike the previous works in sarcasm detection on social media that used only one mode of data, mostly text, we experimented on the potentials of several modes of data, e.g. text, image, and then came up with a multimodal approach to detect sarcasm on SNSs. Our multimodal approach is backed up with a qualitative study with a group of participants recruited from two different countries speaking two different languages. Our study shows the superior performance of multimodal model over traditional unimodal approaches for sarcasm detection. We made our code publicly available as well (See in Appendix F).

8.1 Design Implications

It is imperative for any human-computer interaction research to suggest design implications from the findings of the study. With our problem of sarcasm detection on online platforms, we discuss design implications in this section. SNS developers can consider these to incorporate into their platforms to improve user experience.

8.1.1 Social Networking Sites Design. SNSs by their very nature utilize persuasive design to engage more and more users on their platforms. In order to understand effective persuasive designs, we need to understand human behavior model. We discuss the persuasive nature of SNSs using a seminal work, Fogg Behavior Model (FBM) [91].

According to FBM, to persuade a user do a target behavior, he/she must (1) have sufficient motivation, (2) have enough ability, and (3) be triggered to perform it. We can assume that a large number of users of SNSs in general have sufficient motivation to interact with other users because interaction is the fundamental reason of users joining an SNS platform. However, the

ability of a user to interact with others depends on various factors. For example, there are some basic ability needs from the users like being able to read and write in English (or in his/her language if it is supported by that particular platform). Besides there are some abilities that users can acquire only over time by using a platform as different SNS platforms organizes their features differently according to the demand of their user-base.

For sarcasm detection on social media, the motivation for users can be described as interacting with users. However, since there are users on these platforms who are not always familiar with each other, and yet communicate with each other through posts, reactions, and comments, a difference of ability among them to understand others' posts and intentions of those posts arises due to their variation of experience–both online and real-life (e.g., cultural, social, national, etc.). Our qualitative study suggests that people unfamiliar with each other often misunderstand sarcasm, especially if they are from different countries and cultural backgrounds. Again, both our qualitative study and Phillips et al. [92] suggest that older adults have difficulty in understanding sarcasm. Older adults are often new users to SNS platforms, and so they lack experience with these. Thus, we can say they might have difficulty to distinguish between sarcasm and general statements in an SNS post.

According to FBM, there are three kinds of triggers called spark, facilitator, and signal. Among those, facilitator is highly appropriate in persuasive design when users have enough motivation but lack ability. In our case, the inexperienced users have motivation to interact on social media but they might not have ability to recognize sarcastic content. An effective facilitator trigger tells users that the target behavior is easy to achieve, i.e., users will not need any further resources or ability. This facilitator can be in form of a text, video, graphics, etc. We suggest that if we can offer a feature on SNS platforms to suggest users whether a post might be sarcastic or not, it will help inexperienced users to be cautious about interaction on the post, and also help them to get accustomed with the platform they are using. However, for experienced social media users it might seem to be an overhead. Hence, the option to enable/disable this feature at user's own discretion should be available.

8.1.2 Natural Language Processing, Understanding, and Generation Tasks.

Natural language processing (NLP) means when an ML model converts unstructured text input data to structured data. Computers being able to understand captured textual/statistical data is termed as natural language understanding (NLU). When ML models can convert structured data into text and write, i.e., generate information in human language is called natural language generation (NLG).

We used many NLP techniques in our study. We will discuss how our findings can be used to complement NLG tasks. Therefore, we believe a little further elaboration of NLG task is in order. NLG in fact depends on NLU and NLP. NLG in application is well-known and well-used for following two purposes: automation of content generation and data delivery in an expected format. For example, a widely seen application of NLG is to produce textual weather forecast reports from input weather data. This can reduce human involvement for trivial tasks.

The scenario that we are going to discuss for NLG is related to auto-replier or chatbot. Chatbots or chatbot-like systems have gained much popularity in online platforms. For example, chatbot-like systems in social media apps suggest some potential replies based on immediate previous sections of a conversation. Again, chatbots deployed in e-commerce sites can take care of replying to trivial positive/negative customer reviews. Let's consider the following two scenarios in Table 8.1. As we can see, the auto-replier system tries to understand the overall general sentiment in the customer review using NLP techniques like sentiment analysis. Then, in NLG phase, if the sentiment in the review is positive, it replies with a general thank-you statement. On the other hand, if the review is negative, the system generates an apologetic reply.

Now, let us consider the following review in Table 8.2. As we can see, this review consists of a text review saying "Thank you for your WONDERFUL service!", and an image of a broken luggage wheel. We can easily understand that this is, in fact, a negative review that was presented sarcastically. If we consider only the text part of this review, using common NLP sentiment analysis, we will be misled to understand the sentiment. For example, when we used TextBlob [60] sentiment analyzer, it showed the following output.

Table 8.1: Sample positive and negative reviews, and replies from chatbot-based auto-replier system.

Sample positive review:	I am satisfied with the service. I received the product within two days with ordinary shipment. They offered a great price. Definitely, I'd recommend ABC company to my friends.	Sample negative review:	Terrible! The product was broken when I received it. After ordering, I also found another site that is offering less price for the same item. Use this company only if you do not love your money.
Sample auto-reply:	Thank you for your kind words.	Sample auto-reply:	We are sorry for your inconvenience. Our support team will be in touch with you for helping you with our service.

```
1 from textblob import TextBlob
2 text1 = TextBlob("Thank you for your WONDERFUL service!")
3 print(text1.sentiment)
4 # output: Sentiment(polarity=1.0, subjectivity=1.0)
```

That means, a traditional sentiment analyzer is highly confident that this review has much subjective information (i.e., information about feelings) in it. This is true for most commercial product reviews. However, the traditional TextBlob sentiment analyzer is fully confident that this is a positive review. This could have been true if we did not have context information, i.e., information about the product's actual quality from the image. That means, despite context information being available through image, if an auto-replier generates an auto reply like "Thank you.", that will not be an appropriate response for this multimodal (text + image) sarcastic review. Thus, we can safely imagine a potential application of our multimodal approach to sarcasm detection in such cases.

Table 8.2: Inappropriate response from auto-replier for a multimodal sarcastic review.

Sample sarcastic review:	Thank you for your WONDERFUL service! 
Sample auto-reply:	Thank you for your kind review.

8.2 Threats to Validity

In any human-centric system that is based on observation of human behavior, it is important to consider potential threats to validity. Such considerations can increase confidence in both the study and the resulting system. This also helps identify and reemphasize the strengths and weaknesses of the system recognizing the limitations, and corner cases. That means, consideration of threats to validity reemphasizes the scope of the study, redraws the application boundaries of the system, and paves the ways to future works.

There have been a significant amount of work on what aspects researchers should consider while evaluating threats to validity of a study. Among them, works by Wohlin et al. [93] and Juristo et al. [94] are considered to be semenal works. In our work, we follow the framework proposed by Wohlin et al. [93]. According to them, there are four main types of threats to validity of human-centric software studies: conclusion, internal, external, and construct validity.

8.2.1 Conclusion Validity. This validity concerns how sure we can be that the treatment we used in the experiment is related to our observed outcome. In our study, the research protocols we used are fairly usual for similar research works. At first to study a previously unexplored problem, we turned to a qualitative study with human users to come up with theories about the sarcasm detection and expression incidents on social media. Then, we collected data from social media platforms. We followed standard data preprocessing, model training and val-

idation methodologies. Thus, we can rule out the probability of the study being associated with conclusion validity threat.

8.2.2 Internal Validity. Internal validity focuses on how sure we can be that experiment actually caused the outcome. For our qualitative phase of the study, we used grounded theory approach for data analysis. Grounded theory builds themes only from the data collected in the study instead of relying on existing theories or intuition based hypotheses. Thus, we can ensure that the themes emerged at qualitative phase of our study were directly caused by the data collected during the experiment. The only way the qualitative phase of the study having internal threat validity is the grounded theory based data analysis being subconsciously impacted by our personal experience. We tried to keep that to a minimal level. Again, for the big data based model development phase, we collected the data from social media and used them as training and testing datasets. The one-to-one relationship between the inputs and outputs of the models in our study nullifies the internal validity threat.

8.2.3 Construct Validity. Construct validity evaluates the relationship between the theory behind the experiment and the observation. With a view to avoiding this validity threat, we designed our research approach accordingly. Instead of relying on hypothesis based on intuitions to design the sarcasm detection model, we chose to proceed with a qualitative study at first. Our big data based experiment for sarcasm detection on social media and the models resulted by that– unimodal ones (text, image) and the multimodal one are designed according to the themes that emerged from the preceding qualitative study. We ensured the participants of our study are social media users of the major platforms from where we collected the data for training our models. Thus, we can safely conclude that we addressed the construct validity concern.

8.2.4 External Validity. External validity draws the application boundary of the system. In our study, we aimed to develop a sarcasm detection system for contents on social media. As we have discussed earlier, for in-person communication, people have access to various nonverbal cues and context information that becomes unavailable or narrowed down for expressing with a small number of modes of data. This objective of our study determined our participant re-

cruitment process of social media users, data collection from social medias, and over all, drew the boundary of application of the system. Though currently our system is designed to accept data from several popular social media like Facebook, Flickr, Twitter, at this phase, we concern ourselves with whether we can generalize the results for other social media,

8.3 Future Works

Evaluations of the potentials threats to validity of our study inspires some new directions for future works of this study.

8.3.1 Generalization to non-English texts. As our qualitative study suggested, there are ways to convey sarcastic cues with text that are applicable to particular languages. In our text based analysis section, we utilized the non-English data by translation. Though it allowed us to validate our model's and existing model's performance on a more challenging data, we could not use the language specific features and cues to improve the model since our model was fed English translation as input. Exploration of ways to incorporate the language and alphabet specific cues to improve sarcasm detection can be a future direction.

Our multimodal model relies on the sentiment of the texts as well. To the best of our knowledge, sentiment analysis for non-English language is not well developed. Inclusion of working non-English sentiment analyzer will help the multimodal sarcasm detection become more generalized.

8.3.2 Utilizing High Level Features of Images. In our work, we utilized image as a source of sarcastic cues. However, we used two types of features of images. First, we investigated how the visual representation style of images can hint about sarcasm conveyed by a post. Second, we generated captions for images to get both sentiment and semantic information about the images. However, as our participants revealed during the qualitative study, the presence of particular objects or persons in the image or as inset of the image can be a vital cue for an image to contain sarcastic content. Closer study to identify such high level features of images can be helpful to detect sarcasm on social media.

8.3.3 Inclusion of More Modalities. In our work, we showed that multimodal approach is better than unimodal approach for sarcasm detection. We utilized text, image, and emoticons as major modalities in our system. Though we used videos for our attention model phase of study, we only used the image channel of video. We did not explore the audio channel as a modality in our study. It might be interesting to study how audio itself as well as the image channel as part of video contributes to sarcasm detection or identify the attention model of sarcasm. Since audio enabled videos can represent more human-like communication on social media, inclusion of it is expected to improve the performance of sarcasm detection.

8.3.4 Deployment at User Level. A major outcome of our study is the multimodal model of sarcasm detection. This model is constructed with generalization to major social networking sites in mind. However, since we developed the model for research purpose, it is not exactly ready for end users. To make it useful for real end users, we can make a web browser extension that will identify a post's probability of being sarcastic. This web browser based solution may pose privacy concerns. For this reason, a website based solution that can do the same probability calculation as browser extension given a link to a social media post might be preferred. Our developed model can serve as the backend engine for both of these end-users solutions.

REFERENCES

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[3] S. Theodoridis and K. Koutroumbas, *Pattern Recognition & Matlab Intro*. Academic Press, Inc., 2010.

[4] J. Tepperman, D. Traum, and S. Narayanan, "" yeah right": Sarcasm recognition for spoken dialogue systems," in *Ninth International Conference on Spoken Language Processing*, 2006.

[5] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, *et al.*, "Fake news vs satire: A dataset and analysis," in *Proceedings of the 10th ACM Conference on Web Science*, pp. 17–21, ACM, 2018.

[6] R. W. Gibbs, "On the psycholinguistics of sarcasm.," *Journal of Experimental Psychology: General*, vol. 115, no. 1, p. 3, 1986.

[7] R. W. Gibbs Jr and H. L. Colston, *Irony in language and thought: A cognitive science reader*. Routledge, 2007.

[8] R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pp. 581–586, Association for Computational Linguistics, 2011.

[9] E. Filatova, "Irony and sarcasm: Corpus generation and analysis using crowdsourcing.," in *LREC*, pp. 392–398, Citeseer, 2012.

[10] R. J. Kreuz and S. Glucksberg, "How to be sarcastic: The echoic reminder theory of verbal irony.," *Journal of experimental psychology: General*, vol. 118, no. 4, p. 374, 1989.

[11] R. Clift, "Irony in conversation," *Language in society*, vol. 28, no. 4, pp. 523–553, 1999.

[12] D. Sperber and D. Wilson, *Relevance: Communication and cognition*, vol. 142. Harvard University Press Cambridge, MA, 1986.

[13] H. L. Colston, "On necessary conditions for verbal irony comprehension," *Pragmatics & Cognition*, vol. 8, no. 2, pp. 277–324, 2000.

[14] S. Kumon-Nakamura, S. Glucksberg, and M. Brown, "How about another piece of pie: The allusional pretense theory of discourse irony.," *Journal of Experimental Psychology: General*, vol. 124, no. 1, p. 3, 1995.

[15] H. Colston and R. Gibbs, "A brief history of irony," *Irony in language and thought: A cognitive science reader*, pp. 3–21, 2007.

[16] "Grice's maxims." https://www.sas.upenn.edu/ haroldfs/dravling/grice.html, 2018. Accessed: May 19, 2018.

[17] S. Attardo, "The violation of grices maxims in jokes," in *Annual Meeting of the Berkeley Linguistics Society*, vol. 16, pp. 355–362, 1990.

[18] A. R. Myers, "Toward a definition of irony," *Studies in language variation: semantics, syntax, phonology, pragmatics, social situations, ethnographic approaches*, pp. 171–183, 1977.

[19] D. Sperber, "Verbal irony: Pretense or echoic mention?," *American Psychological Association*, 1984.

[20] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter.," in *ICWSM*, pp. 574–577, 2015.

[21] A. Utsumi, "Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony," *Journal of Pragmatics*, vol. 32, no. 12, pp. 1777–1806, 2000.

[22] B. C. Wallace, L. Kertz, E. Charniak, *et al.*, "Humans require context to infer ironic intent (so computers probably do, too)," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 512–516, 2014.

[23] R. Swanson, S. Lukin, L. Eisenberg, T. C. Corcoran, and M. A. Walker, "Getting reliable annotations for sarcasm in online dialogues," *arXiv preprint arXiv:1709.01042*, 2017.

[24] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714, 2013.

[25] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in twitter," *Language resources and evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[26] M. Khodak, N. Saunshi, and K. Vodrahalli, "A large self-annotated corpus for sarcasm," *arXiv preprint arXiv:1704.05579*, 2017.

[27] M. Cliche, "The sarcasm detector." http://www.thesarcasmdetector.com/, 2014. Accessed: May 19, 2018.

[28] "Definition of sarcasm by merriam-webster." https://www.merriam-webster.com/dictionary/sarcasm, 2018. Accessed: May 19, 2018.

[29] M. S. Razali, A. A. Halin, N. M. Norowi, and S. C. Doraisamy, "The importance of multimodality in sarcasm detection for sentiment analysis," in *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pp. 56–60, IEEE, 2017.

[30] C.-C. Peng, M. Lakis, and J. W. Pan, "Detecting sarcasm in text," *cs229.stanford.edu/proj2015/044_report.pdf*, 2015.

[31] T. Ghosh, "Sarcasm detection - machine learning perspective," *Analytics Experience*, 2016.

[32] A. Ghosh and T. Veale, "Fracking sarcasm using neural network," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 161–169, 2016.

[33] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 1136–1145, ACM, 2016.

[34] H. S. Cheang and M. D. Pell, "The sound of sarcasm," *Speech communication*, vol. 50, no. 5, pp. 366–381, 2008.

[35] L. A. Goodman, "Snowball sampling," *The annals of mathematical statistics*, pp. 148–170, 1961.

[36] F. Baltar and I. Brunet, "Social research 2.0: virtual snowball sampling method using facebook," *Internet research*, vol. 22, no. 1, pp. 57–74, 2012.

[37] H. S. Ferdous, D. Das, and F. M. Choudhury, "Social media question asking (smqa): Whom do we tag and why?," in *In Proc. OzCHI'18*, ACM, 2018.

[38] K. Charmaz and L. L. Belgrave, "Grounded theory," *The Blackwell encyclopedia of sociology*, 2007.

[39] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: three types of fakes," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, p. 83, American Society for Information Science, 2015.

[40] E. C. Tandoc Jr, Z. W. Lim, and R. Ling, "Defining "fake news" a typology of scholarly definitions," *Digital Journalism*, vol. 6, no. 2, pp. 137–153, 2018.

[41] Merriam-Webster Dictionary, "Satire Definition." `https://www.merriam-webster.com/dictionary/satire`, n.a. Online; accessed 25 September 2018.

[42] L. Gou, M. X. Zhou, and H. Yang, "Knowme and shareme: understanding automatically discovered personality traits from social media and user sharing preferences," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 955–964, ACM, 2014.

[43] J. Zhao, L. Gou, F. Wang, and M. Zhou, "Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 203–212, IEEE, 2014.

[44] K. Byron, "Carrying too heavy a load? the communication and miscommunication of emotion by email," 2008.

[45] J. M. DiMicco and D. R. Millen, "Identity management: multiple presentations of self in facebook," in *Proceedings of the 2007 international ACM conference on Supporting group work*, pp. 383–386, ACM, 2007.

[46] IBM, "The science behind the service." `https://console.bluemix.net/docs/services/tone-analyzer/science.html\#the-science-behind-the-service`, 2017. Online; accessed 29 September 2018.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[48] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[49] J. B. Lovins, "Development of a stemming algorithm," *Mech. Translat. & Comp. Linguistics*, vol. 11, no. 1-2, pp. 22–31, 1968.

[50] A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www.cs.cmu.edu/ mccallum/bow, 1996.

[51] C. E. Shannon, "A note on the concept of entropy," *Bell System Tech. J*, vol. 27, no. 3, pp. 379–423, 1948.

[52] M. I. Tanveer, S. Samrose, R. A. Baten, and M. E. Hoque, "Awe the audience: How the narrative trajectories affect audience perception in public speaking," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 24, ACM, 2018.

[53] K. Vonnegut, *Palm Sunday: an autobiographical collage*. Dial Press, 1999.

[54] J. Gao, M. L. Jockers, J. Laudun, and T. Tangherlini, "A multiscale theory for the dynamical evolution of sentiment in novels," in *Behavioral, Economic and Socio-cultural Computing (BESC), 2016 International Conference on*, pp. 1–4, IEEE, 2016.

[55] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Science*, vol. 5, no. 1, p. 31, 2016.

[56] S. Samothrakis and M. Fasli, "Emotional sentence annotation helps predict fiction genre," *PloS one*, vol. 10, no. 11, p. e0141922, 2015.

[57] C. Strapparava, A. Valitutti, *et al.*, "Wordnet affect: an affective extension of wordnet.," in *Lrec*, vol. 4, pp. 1083–1086, Citeseer, 2004.

[58] C. M. Bishop, *Pattern recognition and Machine Learning*. Springer, 2006.

[59] IBM, "Tone analyzer, understand emotions and communication style in text." `https://www.ibm.com/watson/services/tone-analyzer/`, 2017. Online; accessed 29 September 2018.

[60] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey, *et al.*, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.

[61] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[66] V. Gajarla and A. Gupta, "Emotion detection and sentiment analysis of images," *Georgia Institute of Technology*, 2015.

[67] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 715–718, ACM, 2010.

[68] "Flickr: Explore interesting contens around flickr." https://www.flickr.com/explore/interesting/, 2005. Accessed: May 19, 2018.

[69] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King, "A corpus for research on deliberation and debate.," in *LREC*, pp. 812–817, 2012.

[70] A. Przelaskowski, "The role of sparse data representation in semantic image understanding," in *International Conference on Computer Vision and Graphics*, pp. 69–80, Springer, 2010.

[71] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[72] D. Das and A. J. Clark, "Sarcasm detection on flickr using a cnn," in *2018 International Conference on Computing and Big Data (ICCBD)*, (Charleston, South Carolina, USA), 9 2018.

[73] "Dictionary - google search." https://www.google.com/search?q=Dictionary#dobs=sentiment. Accessed: June 13, 2018.

[74] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*, pp. 115–120, Association for Computational Linguistics, 2012.

[75] M. Choy, M. L. Cheong, M. N. Laik, and K. P. Shung, "A sentiment analysis of singapore presidential election 2011 using twitter data with census correction," *arXiv preprint arXiv:1108.5520*, 2011.

[76] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, no. 6, pp. 282–292, 2012.

[77] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, pp. 415–463, Springer, 2012.

[78] "Sentiment analysis." https://web.stanford.edu/class/cs124/lec/sentiment.pdf. Accessed: March 12, 2019.

[79] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[80] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[81] E. Rader and R. Gray, "Understanding user beliefs about algorithmic curation in the facebook news feed," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 173–182, ACM, 2015.

[82] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[83] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[84] B. Sui, *Information gain feature selection based on feature interactions*. PhD thesis, 2013.

[85] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, "Feature selection using information gain for improved structural-based alert correlation," *PloS one*, vol. 11, no. 11, p. e0166017, 2016.

[86] D. Das and A. J. Clark, "Sarcasm detection on facebook: A supervised learning approach," in *20th ACM International Conference on Multimodal Interaction (ICMI)*, (Boulder, Colorado, USA), 10 2018.

[87] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[88] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[89] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[90] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[91] B. J. Fogg, "A behavior model for persuasive design," in *Proceedings of the 4th international Conference on Persuasive Technology*, p. 40, ACM, 2009.

[92] L. H. Phillips, R. Allen, R. Bull, A. Hering, M. Kliegel, and S. Channon, "Older adults have difficulty in decoding sarcasm.," *Developmental psychology*, vol. 51, no. 12, p. 1840, 2015.

[93] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[94] N. Juristo and A. M. Moreno, *Basics of software engineering experimentation*. Springer Science & Business Media, 2013.

APPENDICES

Appendix A. IRB Approval Letter

To:

Anthony Clark

Computer Science

RE: Notice of IRB Approval

Submission Type: Initial

Study #: IRB-FY2018-700

Study Title: Sarcasm Detection on Social Media

Decision: Approved

Approval Date: November 9, 2018

Expiration Date: November 9, 2019

This submission has been approved by the Missouri State University Institutional Review Board (IRB) for the period indicated.

Federal regulations require that all research be reviewed at least annually. It is the Principal Investigators responsibility to submit for renewal and obtain approval before the expiration date. You may not continue any research activity beyond the expiration date without IRB approval. Failure to receive approval for continuation before the expiration date will result in automatic termination of the approval for this study on the expiration date.

You are required to obtain IRB approval for any changes to any aspect of this study before

they can be implemented. Should any adverse event or unanticipated problem involving risks to subjects or others occur it must be reported immediately to the IRB.

This study was reviewed in accordance with federal regulations governing human subjects research, including those found at 45 CFR 46 (Common Rule), 45 CFR 164 (HIPAA), 21 CFR 50 & 56 (FDA), and 40 CFR 26 (EPA), where applicable.

Researchers Associated with this Project:

PI: Anthony Clark

Co-PI:

Primary Contact: Dipto Das

Other Investigators: Dipto Das, Anthony Clark

Appendix B. Recruitment Flyer

Missouri State University

Sarcasm Detection on Social Media

Participants Wanted for a Research Study

This research aims to find out the patterns of contents that users post on social media with sarcastic intents. We will ask the participants how they structure the posts they share on social media by which they want to convey a sarcastic message and how they detect sarcasm in the contents shared by other users both in cases when they do and do not have enough context information. This research will not collect any identifiable information from the participants. The information collected from the participants will only be used by the researchers for the sole purpose of research. The research does not have any commercial objective or goal.

The participants of this study are expected to be active users (using for at least 5-7 hours per week) of social media (e.g., Facebook, Twitter, and etc.). The participants will be required to attend one session of interview with the following researchers of about 30 minutes. Participation in this study is voluntary i.e., a participants can leave the interview at any point if he/she wants.

To learn more about this research, you can contact the following person:

Dipto Das,

Graduate student,

ARCS Lab, Computer Science (CSC),

Missouri State University

Email: dipto175@live.missouristate.edu

This research is conducted under the direction of Dr. Anthony J. Clark, Computer Science, Missouri State University.

Appendix C. Inform Consent Form

Informed Consent

Sarcasm Detection on Social Media

TITLE OF STUDY

Sarcasm Detection on Social Media

PRINCIPAL INVESTIGATOR

Dr. Anthony Clark

Computer Science Department

Cheek 307

(417) 836 - 5438

AnthonyClark@MissouriState.edu

PURPOSE OF STUDY

You are being asked to take part in a research study. Before you decide to participate in this study, it is important that you understand why the research is being done and what it will involve. Please read the following information carefully. Please ask the researcher if there is anything that is not clear or if you need more information.

The purposes of this study are to discover the structure of sarcastic posts on social networking sites and the methods users use to identify sarcasm on such platforms.

STUDY PROCEDURES

You will be asked questions about your behavior on social networking sites, particularly as it relates to sarcasm. For example, how do you identify that a post on a social networking site is sarcastic; how do you understand the differences between a sarcastic and a non-sarcastic post; and how do you construct your posts or contents when you want them to convey sarcastic sentiment?

You should be prepared to participate in an interview for approximately 30 minutes. It is expected that you will only need to participate one such session.

Interviews will be audio-recorded so that researchers can refer back to conversations that took place. Researchers will compare responses among participants to identify common themes of sarcastic posts. Researchers are also interested in a identifying a general method for identifying sarcasm from non-sarcasm on social media sites.

RISKS

You are unlikely to experience any risks.

You may decline to answer any or all questions and you may terminate your involvement at any time.

BENEFITS

There will be no direct benefit to you for your participation in this study. However, we hope that the information obtained from this study will help build automated systems to detect sarcasm on social media sites. This research will provide insights to researches in data science, recommender systems, and linguistics.

CONFIDENTIALITY

Your responses to the interview will be anonymous. Please do not reveal any identifying information about yourself during the interview. Every effort will be made by the researcher to preserve your confidentiality including the following:

- Assigning code names/numbers for participants that will be used on all research notes and documents,
- Keeping notes, interview transcriptions, and any other identifying participant information in a locked file cabinet in the personal possession of the researcher, and
- Storing digital recordings of interviews on a password protected computer.

Participant data will be kept confidential except in cases where the researcher is legally obligated to report specific incidents. These incidents include, but may not be limited to, incidents of abuse and suicide risk.

CONTACT INFORMATION

If you have questions at any time about this study, or you experience adverse effects as the result of participating in this study, you may contact the researcher whose contact information is provided on the first page. If you have questions regarding your rights as a research participant, or if problems arise which you do not feel you can discuss with the Primary Investigator, please contact the Institutional Review Board at (417) 836-8362, ext. 8991.

VOLUNTARY PARTICIPATION

Your participation in this study is voluntary. It is up to you to decide whether or not to take part in this study. If you decide to take part in this study, you will be asked to sign a consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason. Withdrawing from this study will not affect the relationship you have, if any, with the researcher. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

CONSENT

I have read, and I understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without cost. I understand that I will be given a copy of this consent form. I voluntarily agree to take part in this study.

Participant's signature _____ Date _____

Investigator's signature_____ Date _____

Appendix D. Questionnaire for Sarcasm Detection on Social Media Project

Demographic Questions

1. How old are you?
2. What is the level of highest education you attended?
3. What is your occupation?

General questions about participant's experience with Social Media

1. What social media platforms do you currently use? (e.g., Facebook, Twitter, Instagram, etc.)
2. What other social media platforms have you heard of, but do not currently use?
3. Why dont you use these other platforms?
4. How long have you been using social media?
5. What kind of content do you prefer on social media?
6. From whom do you prefer to receive posts on social media? (e.g., close friends, family, acquaintances, pages, etc.)
7. How much time do you spend regularly on social media?
8. With whom do you mostly interact on social media?

Questions about Sarcasm Detection Approach on Social Media

1. How frequently do you see posts that you think are sarcastic?
2. How do you recognize that a post is sarcastic?
3. How important do you think context (e.g., prior conversations, images, etc.) is for detecting sarcasm?

4. If enough context information is not available, how do you differentiate sarcastic posts from non-sarcastic posts?

5. How does sarcasm effect the popularity/reachability of a post?

6. How do you convey sarcasm with your posts?

7. How are sarcastic posts received by others?

8. How do you react to sarcastic posts?

9. How do other users react to sarcastic posts?

Appendix E. Datasets

1. Yahoo Flickr Sarcasm (YFS) Dataset:

   `http://bit.ly/yfsdataset`

2. Bengali Satire-Fake News Dataset:

   `http://bit.ly/bengalisatirefakenewsdataset`

3. Multimodal Sarcasm Dataset:

   `http://bit.ly/multimodalsarcasmdataset`

4. Sarcasm Video Attention Dataset:

   `http://bit.ly/sarcasmvideoattention`

Appendix F. Codes

1. Sarcasm detection on Flickr using a CNN:

   `https://github.com/DiptoDas8/imagenet.git`

2. A multimodal (image, text, emoticons)-based approach to sarcasm detection on Facebook:

   `https://github.com/DiptoDas8/imagetext.git`

3. Sarcasm detection in text with narrative trajectory of tones:

   `https://github.com/DiptoDas8/sarcasmtone.git`

4. Analyzing the attention model of sarcasm in videos:

   `https://github.com/DiptoDas8/sarcasm-attention.git`