Harassment Detection on Twitter using Conversations

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

By

Venkatesh Edupuganti
B.Tech., Andhra University, 2014

2017
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

December 6, 2017

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY VENKATESH EDUPUGANTI ENTITLED HARASSMENT DETECTION ON TWITTER USING CONVERSATIONS BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

_____

Krishnaprasad Thirunarayan, Ph.D.
Thesis Director

_____

Mateen M. Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

_____

Krishnaprasad Thirunarayan, Ph.D.

_____

Amit P. Sheth, Ph.D.

_____

Valerie L. Shalin, Ph.D.

_____

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Edupuganti, Venkatesh. M.S., Department of Computer Science and Engineering, Wright State University, 2017. Harassment Detection on Twitter using Conversations.

Social media has brought people closer than ever before, but the use of social media has also brought with it a risk of online harassment. Such harassment can have a serious impact on a person such as causing low self-esteem and depression. The past research on detecting harassment on social media is primarily based on the content of messages exchanged on social media. The lack of context when relying on a single social media post can result in a high degree of false alarms.

In this study, I focus on the reliable detection of harassment on Twitter by better understanding the context in which a pair of users is exchanging messages, thereby improving precision. Specifically, I use a comprehensive set of features involving content, profiles of users exchanging messages, and the sequence of messages. By analyzing the conversation between users and features such as change of behavior during their conversation, length of conversation and frequency of curse words, I find that the detection of harassment can be improved significantly over merely using content features and user profile information. Experimental results demonstrate that the comprehensive set of features I use in my supervised machine learning classifier achieves F-score of 88.2 and Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) of 94.3.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I would like to express sincere gratitude to my advisor, mentor, and life coach, Dr. Krishnaprasad Thirunarayan, for his immense support throughout my Master's degree. I am always thankful to him for providing me with this amazing opportunity with incredible support. He always takes time out from his busy schedule to keep me in the right direction by his valuable guidance. His astute advice helped me through my journey here, technical, and otherwise. I am sure that lessons learned from him will significantly help in my future.

I would also like to thank Dr. Sheth and Dr. Shalin, for agreeing to be on the thesis committee and their valuable feedback. Suggestions and constructive criticism from them helped me to think differently and improved my thesis.

I owe thanks to Dr. Sumanth Kulkarni and Monireh Ebrahimi for valuable guidance during my initial days of research.

This thesis is based upon the work supported by the National Science Foundation (NSF) through Grant Award No. CNS 1513721 titled *TWC SBE: Medium: Context-Aware Harassment Detection on Social Media.*

I would like to thank Roopteja, Ramya, Sriramya, Ankita, Swati, Shruti, Lakshika, Vaikunth, Revathy, Pavan, and Rajeswari – I couldn't have done it without you guys. I would also like to thank whole Kno.e.sis family. I learned a lot and made great memories.

This acknowledgment would be incomplete without thanking my family. I would like to thank my dad, my mom, and Teja for encouraging me and standing by my side throughout my life.

Dedicated to

my father Srinivasa Rao, my mother Anupama, and my brother Durga Ravi Teja.

1. Introduction

Social media has become a popular medium for communication among people, and for companies and organizations to reach their audience. People share their thoughts, feelings, experiences, and details about their day-to-day activities on social media[1]. For example, in 2017 Twitter had 328 million monthly active users[2] and 500 million tweets per day[3]. 29.2% of United States social media users use the Twitter platform, and 83% of the world's leaders have Twitter accounts[4]. According to the 2016 Pew Internet & American Life Project report[5], 24 percent of the online adults (roughly one-quarter) use Twitter as shown in Figure 1, and this proportion is almost unchanged from the previous 2015. Twitter is one of the most popular social media among highly educated people with 29% of them having college degrees[5]. Furthermore, the use of social media among adults has increased from 5% to 69% in the span of 11 years from 2005 to 2016[6] as shown in Figure 2.

[1] https://en.wikipedia.org/wiki/Social_media
[2] Q217_Selected_Company_Metrics_and_Financials
[3] https://www.omnicoreagency.com/twitter-statistics
[4] https://www.brandwatch.com/blog/44-twitter-stats-2016/
[5] http://www.pewinternet.org/2016/11/11/social-media-update-2016/
[6] http://www.pewinternet.org/fact-sheet/social-media/

Despite the ease of content creation, communication and improved connectivity among people, online communication has also brought with it additional problems. Bullies can more easily reach and torment a much wider vulnerable population as compared to traditional means. Bullying is a form of unprovoked aggression that can create an intimidating or hostile environment for an individual or group of individuals. Traditional bullying is often done face-to-face and at locations such as work or school, whereas cyberbullying can be done anonymously and anywhere. This can cause a big impact on a person's life because information can spread rapidly through the user network, crossing physical boundaries and reaching a much wider audience.

24% of online adults (21% of all Americans) use Twitter

% of online adults who use Twitter

All online adults	24%
Men	24
Women	25
18-29	36
30-49	23
50-64	21
65+	10
High school degree or less	20
Some college	25
College+	29
Less than $30K/year	23
$30K-$49,999	18
$50K-$74,999	28
$75,000+	30
Urban	26
Suburban	24
Rural	24

Note: Race/ethnicity breaks not shown due to sample size.
Source: Survey conducted March 7-April 4, 2016.
"Social Media Update 2016"

PEW RESEARCH CENTER

Figure 1. Demographic statistics about Twitter users[5].

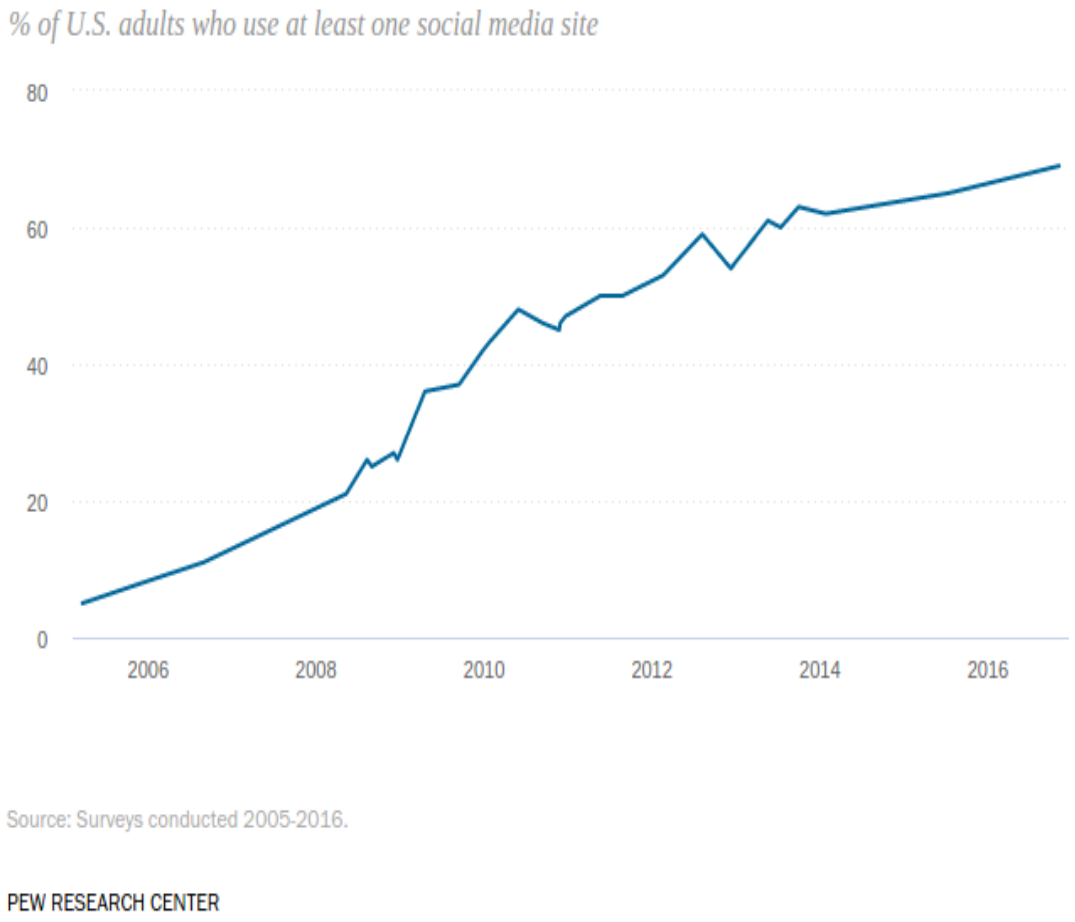


Figure 2. Percentage growth in social media sites used in the United States[6].

According to a 2017 Pew Internet & American Life Project survey report[7], 41% of Americans have personally encountered online harassment, and 66% have witnessed online harassment behaviors directed at others. The report also states that after seeing harassment 27 percent of the participating Americans chose not to post anything online. Furthermore, 62% of the participating public has

[7] http://www.pewinternet.org/2017/07/11/online-harassment-2017/

experienced online harassment in some or the other form[7]. Another report on Online Harassment, Digital Abuse, Cyberstalking in America in 2016[8] states that 70% of young adults have been the target of online harassment. The report also states that every one out of four victims has disconnected themselves from social media, internet, or their smartphones to avoid getting harassed.

Twitter is one of the top social networking sites that has shown a recent, significant increase in cyberbullying, according to BuzzFeed, which has surveyed 2700 twitter users[9]. Twitter has been trying to respond to such issues but has not done so satisfactorily or has been unable to do so successfully. According to BuzzFeed[9], 46.4% of users said that Twitter does nothing after an abusive tweet is reported and 28.2% of users said they never heard back after they reported abuse on Twitter. Only 2.6% of users said that Twitter had deleted the offensive accounts and 18.2% of users got a reply from Twitter that it does not violate their policy and terms of service as shown in Figure 3(a). 37.8% of users reported that Twitter took more than 48 hours to take action against harasser after reporting and 8.2% users reported that Twitter took less than 4 hours as shown in Figure 3(b).

[8] https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf
[9] https://www.buzzfeed.com/charliewarzel/after-reporting-abuse-many-twitter-users-hear-silence-or-wor?utm_term=.egBgRNJwJ#.buZGldJoJ

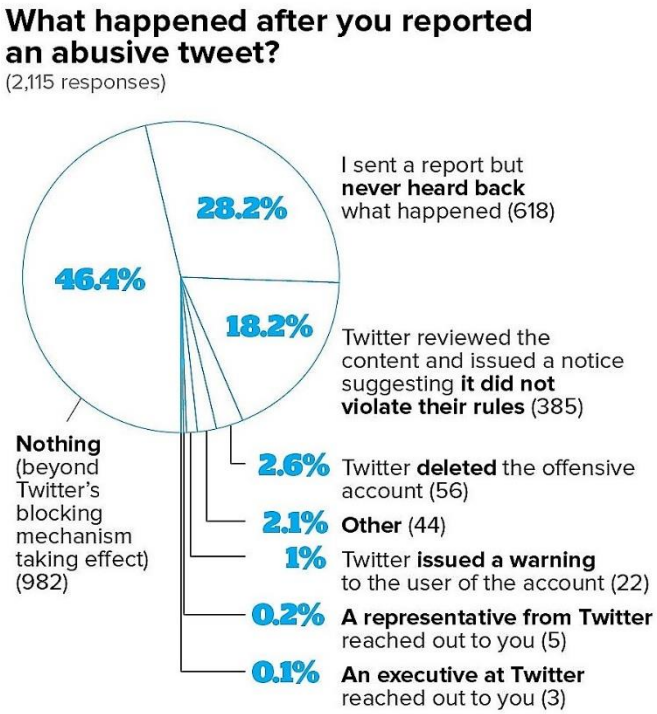


Figure 3(a)

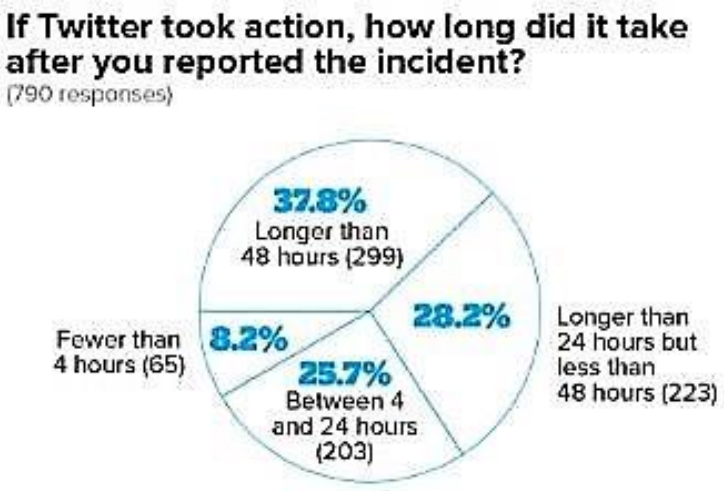


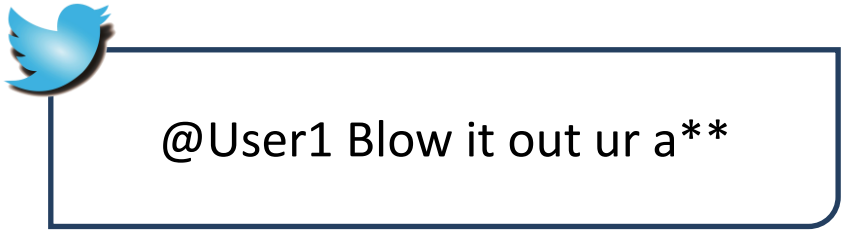
Figure 3(b)

Figure 3. How Twitter responds to harassers[9].

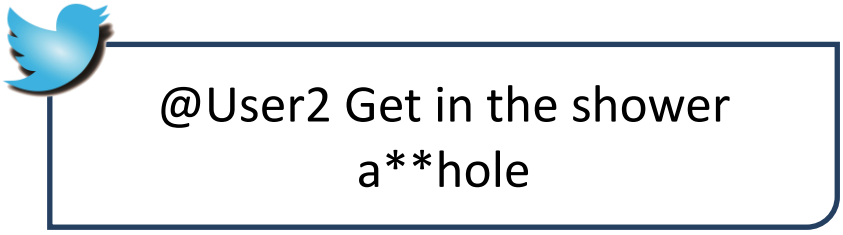
A key challenge to reducing online harassment is to detect harassment at scale reliably and efficiently. This requires addressing the big data challenges where it is required to handle a high volume of tweets from a variety of sources and processing them quickly and reliably.

1.1. Current Research

Ongoing research on harassment detection on Twitter primarily examines the text of an individual message and user information to identify harassing tweets. Researchers use cues such as the presence of demeaning, insulting or obscene words in a tweet to flag tweets that are harassing. Yet, although many harassing tweets share these traits, these same words also appear in tweets that are shared among friends as friendly banter, or as a defense against harassment or as light-hearted scuffle among friends. In such cases, though the tweets contain features commonly seen in harassing tweets such as curse words, these are not harassing. Such tweets cause false positives and reduce the precision of current algorithms. My key observation is that, instead of processing individual tweets in isolation to detect harassment, it is much more effective to understand the context of a tweet by situating it with respect to the conversation in which it appears.



4(a) Harassing Tweet.



4(b) Non- Harassing Tweet

Figure 4. Tweets on Twitter containing demeaning, insulting or obscene words.

@user1	: you're the love of my life and he's my boyf
@user1	: so ur not the loml??? :/
@user2	I thought I was but u have a boyf :/
@user2	get in the shower asshole
@user1	: actually nvm i don't need a Harry Styles when I have you
@user1	: I'll carry you to the bathroom
@user1	: probably my best pic this should be your lock screen
@user1	: but babe
@user2	but baby
@user2	why ya gotta have a job babe
@user1	: i gotta save for the ring i plan on putting on ur finger babe
@user2	m in love with your gay ass
@user2	that's you baby
@user1	: don't give me that face
@user2	I LOVE you bitch
@user2	I ain't never gonna stop loving you BITXH
@user1	: I love you hoe ass
@user2	I love you more gay ass

Figure 5. The conversation between users surrounding Figure 4(b).

Both tweets in Figure 4 are classified as harassing tweets when processed individually and in isolation. But the tweet in Figure 4(b) is not a harassing tweet when I consider the conversation between the two users shown in Figure 5. Considering the prior conversation between users provides the relevant context, and with the help of user profile information, can reduce false positives.

Below I demonstrate that misclassification is remedied by including additional contextual knowledge about the tweets to the classifier. Specifically, the features gleaned from the user's profile and the conversations between a pair of users can provide the necessary context to improve classifier performance.

The challenges I faced while building a reliable classifier for harassment detection using conversations are:

1. *Data collection:* Twitter provides only one percent of its streaming data through twitter streaming API out of which I need to collect the tweets that are suitable for my research.

2. *Annotation:* While annotating the tweet as harassing or not, I used all the tweets exchanged between the pair of the user (conversation) to understand the actual context of the tweet and relationship between the users. Annotating a tweet based on the conversation-between the users is labor-intensive and is discussed in Section 3.3.1.

3. *Feature selection:* Selecting meaningful and significant features which can help the classifier to detect harassing tweets is non-trivial. The features used by most of the previous studies do not yield good results on my dataset. More details about features and their rationale can be found Chapter 4.

4. *Overcoming class imbalance*: Making sure that the classifier can reliably classify the tweets without bias in the face of a dataset that has a predominance of non-harassing tweets is challenging. I used SMOTE approach to overcome the class imbalance issue. This is explained in detail in Chapter 5.

1.2. Outline

The rest of the thesis is structured as follows: In Chapter 2, I review details of the related prior harassment detection research. In Chapter 3, I formally define my approach to harassment detection on Twitter. Then, I discuss the overall framework of my Harassment detection using conversations and explain in detail the data collection process. In Chapter 4, I discuss various features and demonstrate how they contribute to reliable harassment detection. In Chapter 5, I present the algorithmic details, my experiments, and results using various classifiers. Chapter 6 concludes this document with suggestions for future work.

2. Related Work

In this Chapter, I discuss details of the past work related to cyberbullying and harassment detection in social media. These studies have used different social media sites for exploration of cyberbullying and harassment [8], [17], [18], [19], [24], [15], [14], [3], [1]. Table 2.1 provides a snapshot of the literature relevant to my work.

Authors & Year	**Dataset source**	**Features**	**Result**
Yin et al. (2009)	Kongregate, Slashdot, and Myspace	Content, sentiment and contextual (TF-IDF, N-grams, and foul language)	48.1 % F-score
Dinakar et al. (2011)	YouTube	General (TF-IDF, Parts of speech bigrams, list of profane words), and label specific (unigram and bigram)	66.7 % accuracy
Reynolds et al. (2011)	Formspring.me	SUM (to measure overall badness) and TOTAL (weighted average of bad words).	78.5% accuracy

Dadvar et al. (2012)	Myspace	Profane words count, second person pronouns, other personal pronouns, and the weight of words in each sentence.	23 % F-score
Chen et al. (2012)	Twitter	Unigrams, character n-gram, and Twitter user modeling features from LIWC (Linguistic Inquiry and Word Count) [7].	Not Provided
Huang et al. (2014)	Twitter	Textual features (number of exclamations, density of bad words, parts of speech bigrams), and social features (number of links, number of edges, number of nodes in relationship graph).	75 % ROC
Zhao et al. (2016)	Twitter	Bag of words features, latent semantic features, and bullying features based on word embedding.	78% F1 score
Despoina et al. (2017)	Twitter	user-based (number of tweets, lists subscribed, account age), text-based (hashtags count, uppercase	90.9% Precision, 91.3 Recall,

		letter, sentiment and curse words count), network-based (popularity, reciprocity, power difference)	81.7 ROC

Table 1. Overview of related work

Yin et al. [19] employed a supervised learning approach to classify harassment. They tested their approach on the data from Kongregate, Slashdot, and Myspace. Their approach combines local features, sentiment features, and contextual features such as TF-IDF (term frequency – invert document frequency), N-grams, and foul language to train a model for detecting harassing posts in discussion forums and chat rooms. They achieved a 48.1 percent F-score with 39.4 percent precision and 61.9 percent recall.

Dinakar et al. [18] obtain better results for detecting harassment by labeling YouTube comments according to the type of bullying it contains such as sexual, racial, or intelligence-related and then using the same binary classifier for individual labels. They used general features (such as TF-IDF, ortony lexicon [26] which contains a list of profane words, and parts of speech bigrams) and labeled specific features (such as topic-specific unigrams and bigrams). They achieve 80.20 percent, 68.30 percent, and 70.39 percent accuracies for the labels sexual, racial, and intelligence-related respectively.

Reynolds et al. [24] classified the document from Formspring.me as cyberbullying or not based on the occurrences of bad words. Initially, they assigned severity levels to each bad word in their dictionary. Later they divided the dataset into two different training sets: one containing the number of bad words, and another containing the density of bad words. Then they generated the features such as SUM (to measure overall badness, i.e., a weighted average of the bad words according to their severity) and TOTAL (normalized by simply dividing the number of words at each severity level by a total number of words in the post). They used a machine learning tool to train the different models such as the decision tree-based algorithm (J48), the rule-based algorithm (JRIP), the instance-based algorithm (IBK) and the support vector machine algorithm (SMO). They can correctly identify 78.5 percent of the data containing cyberbullying using the decision tree-based algorithm (J48) on both the datasets.

Dadvar et al. [25] showed that cyberbullying classification could be improved using gender information. This study showed that the vocabulary (like curse words) used by males and females differ. They trained two different classifiers for male and female corpus separately using the features (such as profane words, second person pronouns, other personal pronouns, and weight of words in each sentence) and then based on the proportion of each gender in the corpus, the result was calculated. They stated that the second person pronoun has a more important role in detecting online harassment, so they used the second personal as a single

feature and other pronouns as another feature. They also calculated the baseline performance metrics by running the classifier on entire corpus without gender separation. They were able to achieve 43 percentage gender-specific precision improving from a baseline precision of 31 percentage, 16 percentage in recall from 15 percentage, and 23 percentage in F-score from 20 percentage.

Chen et al. [17] implemented a demo system that can detect real-time cyber-bullying on Twitter. Their system can track tens of thousands of active children users on Twitter by using the features such as regular expressions (to identify emotional ASCII characters, URLs, and hashtags), character n-gram, unigrams, n-grams, and Twitter user modeling features from LIWC [7]. The main idea behind the system was to engage parents and teachers in monitoring their children's tweets to know whether they are being bullied or they are bullying others. A Gradient Tree Boosting model was used to differentiate between a normal tweet and a bullying tweet. However, the researchers did not reveal the performance of their system.

Huang et al. [8] show that considering the social relationships between the users on Twitter can improve the cyberbullying classification results. They build relationship graph for each tweet and extract social features from it (number of links, number of edges, and number of nodes in relationship graph). Textual features included a number of exclamations, the density of bad words, and Parts of

Speech bigrams. They achieved around 75 percent AUC of ROC and 76 percent true positive rate when classification used both textual and social features.

Zhao et al. [3] ran their experiments on a Twitter dataset where tweets contain at least one of the following words: bully, bullied, or bullying. They proposed a new method named Embedding-enhanced Bag-of-words (EBoW) that combines the bag of words features, latent semantic features, and bullying features based on word embedding. They used a linear support vector machine to classify the tweet as bullying or not. They also compared the performance of their method with the bag-of-words model, a semantics-enhanced bag of words model, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA) based on Gensim. They achieved 76.8 percent precision, 79.4 percent recall, and 78 percent F1 score.

Despoina et al. [1] ran their experiments on a Twitter dataset of tweets crawled by using 309 hashtags related to bullying and hateful speech. They divided the features into three categories: user-based features (number of tweets, age of account, the number of lists user subscribed to, and account verification status), text-based features (number of hashtags used, uppercase text, emoticons count, URLs count, and sentiment) and network-based features (popularity, reciprocity, power difference, hubs and authority, and influence). They made both 4-class (bully, aggressive, spammers and normal) and 3-class (bully, aggressive and normal) classifiers. While performing 3-class classification, they removed all the spam accounts from the dataset. They achieved 0.718 and 0.899 precision, 0.733

and 0.917 recall, 0.815 and 0.907 AUC of ROC while using 4-class classification and 3-class classification respectively.

2.1. Research Gap

The Work	Textual Features	User Features	Social Network Features	Features based on sequence of messages exchanged between users
Yin et al.	Yes	No	No	No
Dinakar et al.	Yes	No	No	No
Reynolds et al.	Yes	No	No	No
Dadvar et al.	Yes	Yes	No	No
Chen et al.	Yes	Yes	No	No
Huang et al.	Yes	No	Yes	No
Zhao et al.	Yes	No	No	No

Despoina et al.	Yes	Yes	Yes	No
Proposed Work	Yes	Yes	No	Yes

Table 2. A comparison of the proposed work with existing works on harassment or cyberbullying detection.

While these different studies highlight the role of the content of individual message or tweet, a network of users, and user profile information in the detection of harassment or cyberbullying in social media, they do not discuss and capture the context in which the users posted such messages or tweets on social media. The user mentioned in tweets are critically important for distinguishing between harassing tweets and non-harassing tweets that seems to use similar words but with different intents. Tweets are shown in Figure 5 that can be classified as harassing when considered individually or in isolation are simply a part of a series of tweets between two users exchanging tweets as friendly banter context. So, the present study focuses on bringing in contextual knowledge by using a sequence of messages (conversation) exchanged between the users to help reduce such false positives and improve the classifier's performance.

3. Harassment Detection using Conversations Framework

Before describing my approach to harassment detection using conversations framework, I describe the problem formally.

3.1. Problem Statement

Harassment is a form of unprovoked aggression that can create an intimidating or hostile environment for an individual or group of individuals. According to the Oxford dictionary[10], harassment is defined as "the act of annoying or worrying somebody by putting pressure on them or saying or doing unpleasant things to them." Formally, I define the problem as follows:

Definition. Harassment Detection on Twitter using conversations*.** Given an unlabeled tweet ***t from user ***U1*** that contains the twitter handle of user ***U2*** along with a set of past tweets **T** between ***U1*** and ***U2***, the harassment detection problem aims at automatically detecting if ***'t'*** is harassing or not.

My approach differs from the existing works by including past conversation tweets **T** between the users. Past conversation tweets **T** play a very crucial role in

[10] https://www.oxfordlearnersdictionaries.com/definition/english/harassment

my classification process because these tweets **T** can be used to judge the context and intent in which the actual tweet **t** was sent. I use the past sequence of tweets (conversation) between the users to understand the actual pragmatics of tweet, i.e., a friendly tweet that was exchanged between friends or defensive tweet sent in response to a harassing tweet or is itself a harassing tweet. So, my classifier for harassment detection using conversations can classify the friendly tweet or defensive tweet containing curse words correctly as a non-harassing tweet.

3.2. Thesis Statement

"Harassment detection on Twitter can be improved by harnessing contextual information from the sequence of messages (conversation) exchanged between users in addition to using tweet content and user profile information."

3.3. Framework

My system consists of three major components implementing the three main phases: corpus creation, feature extraction, and classification. During the corpus creation phase, I collect tweets from Twitter, then filter them and extract user profiles and conversations. During the feature extraction phase, I extract content features of the tweet from tweet filtering module, user profile features from user extraction module and conversation features from conversation extraction module. Finally, the classifier determines whether the tweet is harassing or not using these

features. I explain briefly what each module does in my framework and their importance.

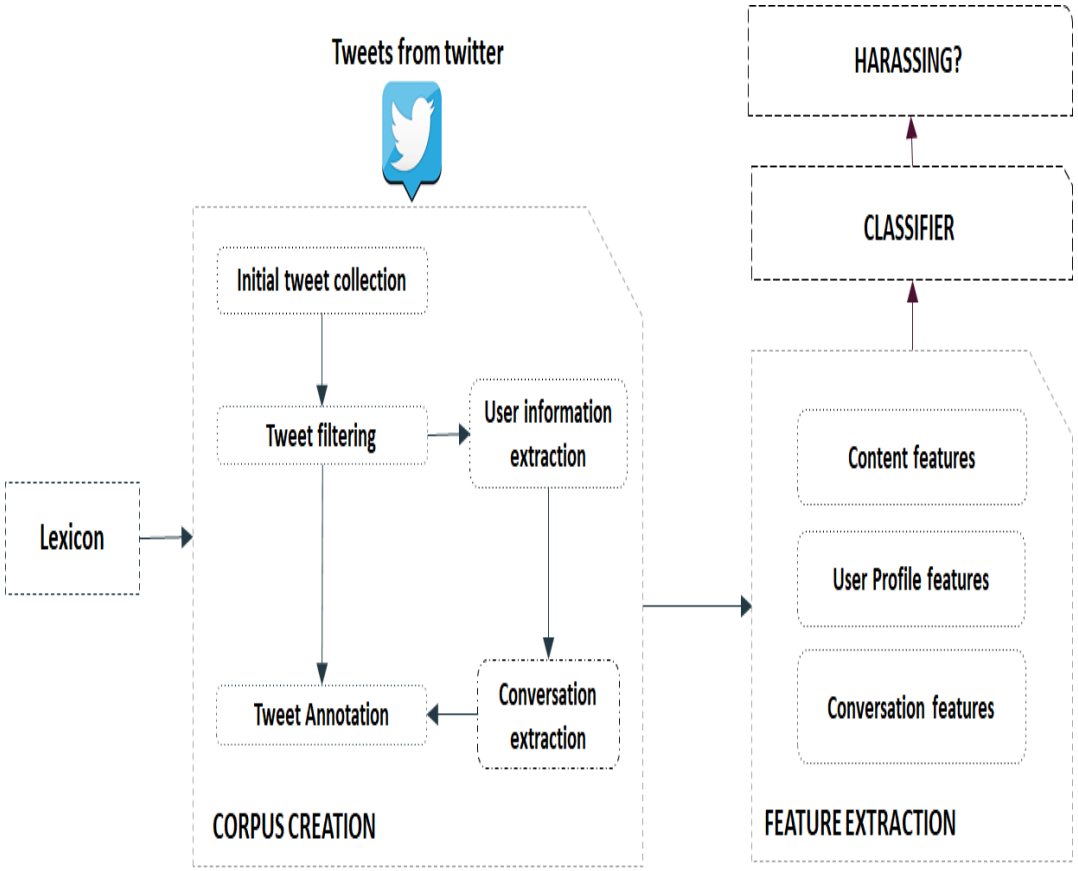


Figure 6. Harassment detection using conversations framework.

3.3.1. Corpus Creation

The corpus creation component provides the dataset for my study. This component includes five modules: initial tweets collection, tweet filtering, user information extraction, conversation extraction and tweet annotation.

The initial tweets collection module collects the tweets from twitter streaming API by using a lexicon containing a dictionary of curse words. So, this module is used to collect the tweets containing curse words from Twitter which has a higher likelihood of being harassing or raise false alarms as in previous studies.

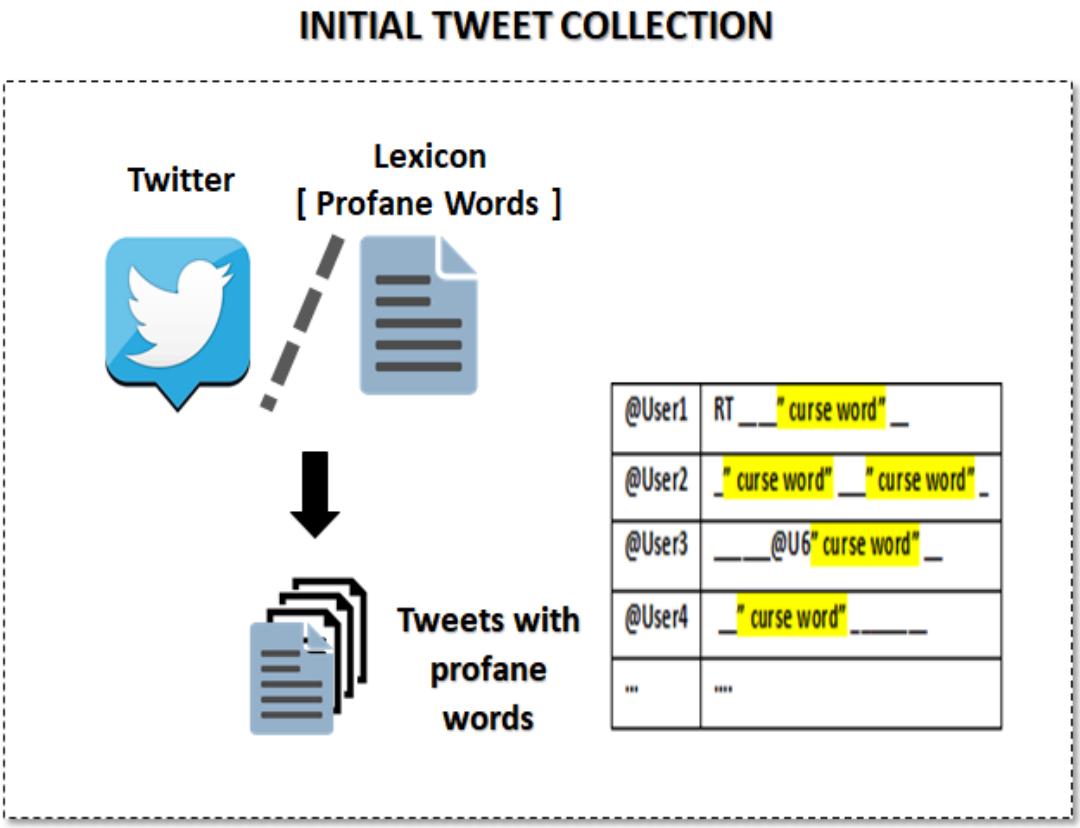


Figure 7. The initial tweet collection module.

Once the initial tweets are collected, I filter them to obtain tweets containing user mentions and directed towards a non-celebrity account or to a user account with a public profile. The tweet filtering module outputs tweets directed towards the user information extraction module and annotation module.

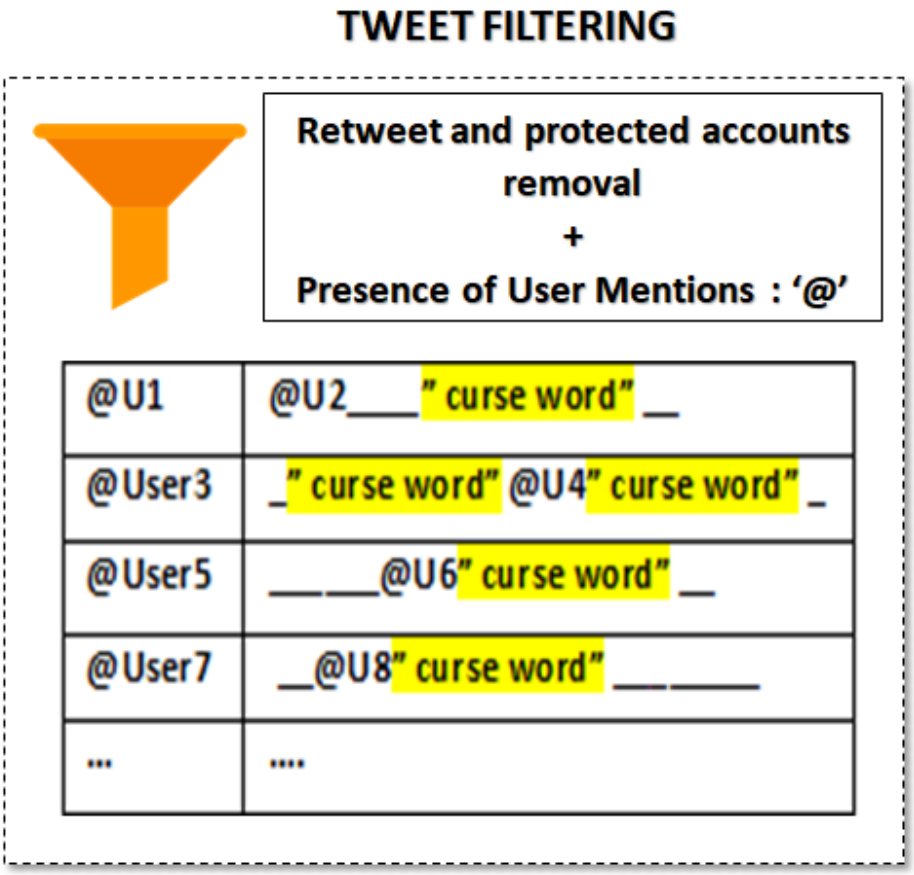


Figure 8. The tweet filtering module.

In the user information extraction module, I extracted the author twitter handle and mentioned user twitter handle from the tweet. This enables access to the entire user's profile information using the Twitter rest API and fed to the user profile features module of the feature extraction component. The author twitter handle and mention user twitter handle are sent to conversation extraction module.

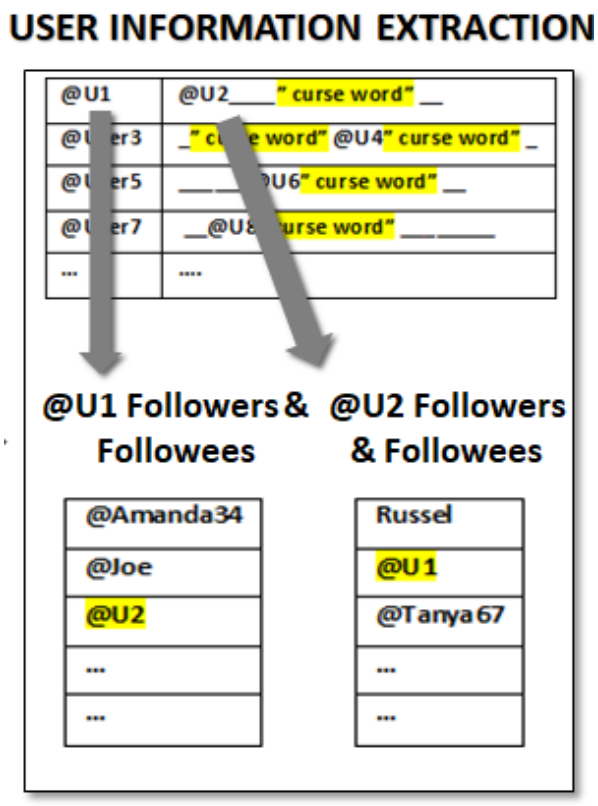


Figure 9. The user profile extraction module.

In the conversation extraction module, I crawled all the available tweets for the user profile and considered only the tweets that contain another user twitter handle (mentioned user if I crawled tweets using author twitter handle and vice versa) using the Twitter API. I also considered the date of the tweet posted to sort both the user's time-stamped tweet in ascending order.

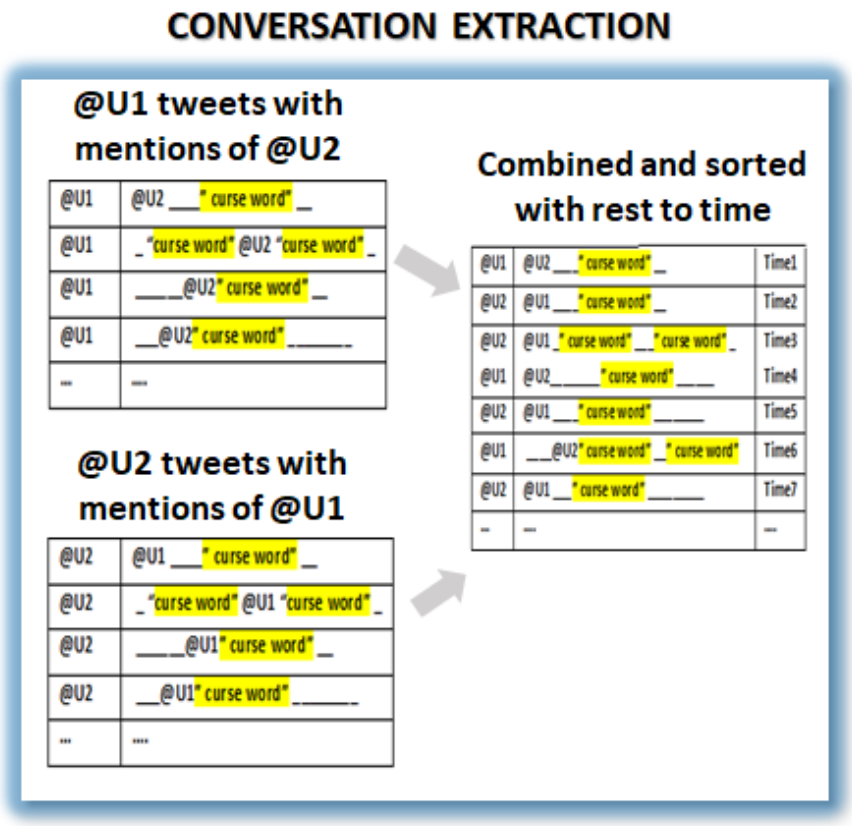


Figure 10. The conversation extraction module.

Once I extracted the conversation between the two users, I annotated the selected tweets, i.e., tweets from the tweet filtering module by using conversation related to that tweet. I will explain in detail the data annotation process in the data collection section.

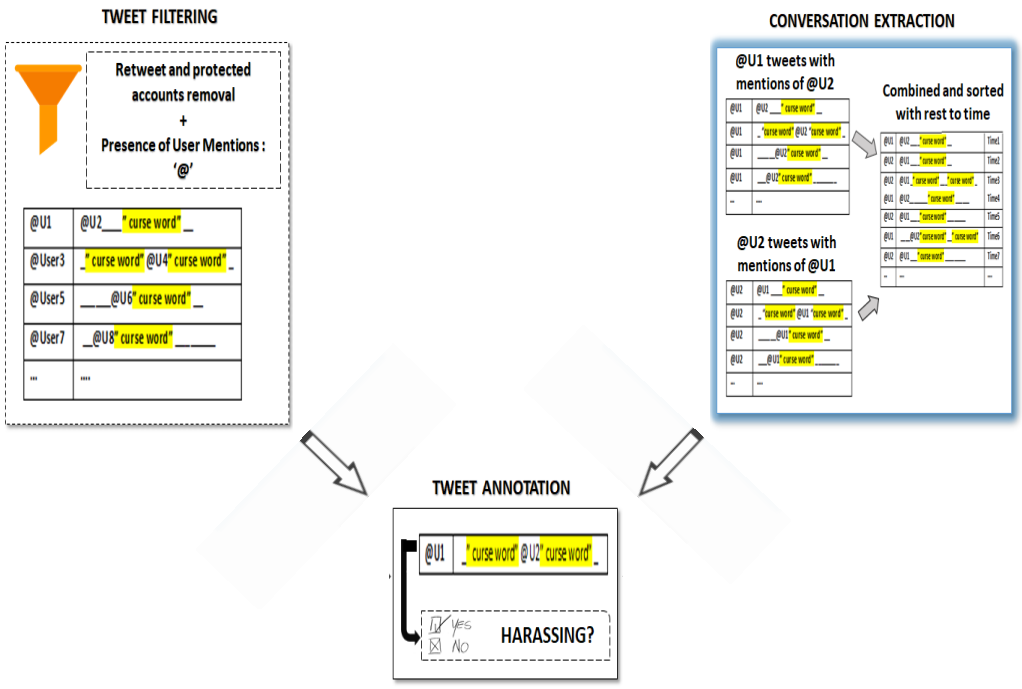


Figure 11. The annotation module.

3.3.2. Feature Extraction

The feature extraction component is used to extract different, meaningful, fine-grained and significant features from the corpus. This component is critical for developing an effective classifier. There are three modules in this component for extracting three categories of features: content features, user profile features, and conversation features.

The content features module generates the features based on the content of the tweet that is returned by the tweet filtering module in the corpus creation phase.

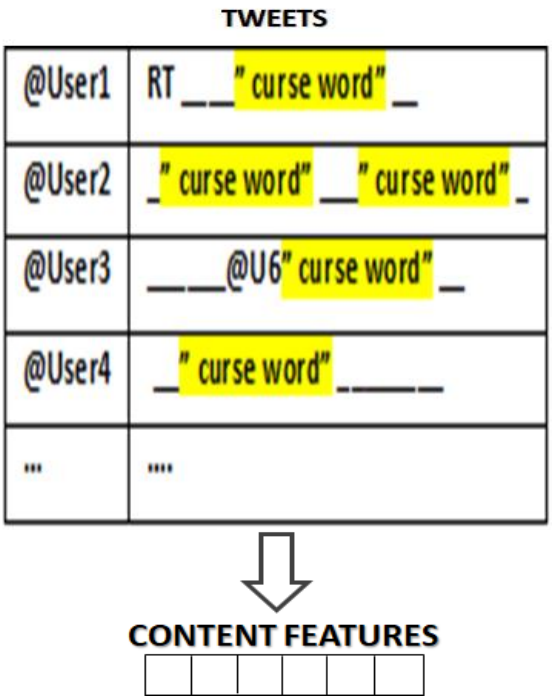


Figure 12. Extracting potential content features from tweets.

The user profile features module generates features using the user profile information from the user information extraction module in the corpus creation component.

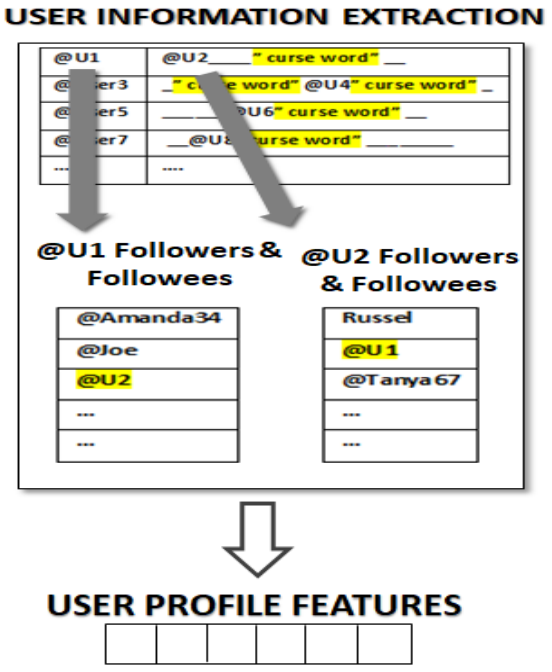


Figure 13. Identifying potential features from the user profile.

The conversation features module generates features based on the conversations from extracting conversations modules in the corpus creation component.

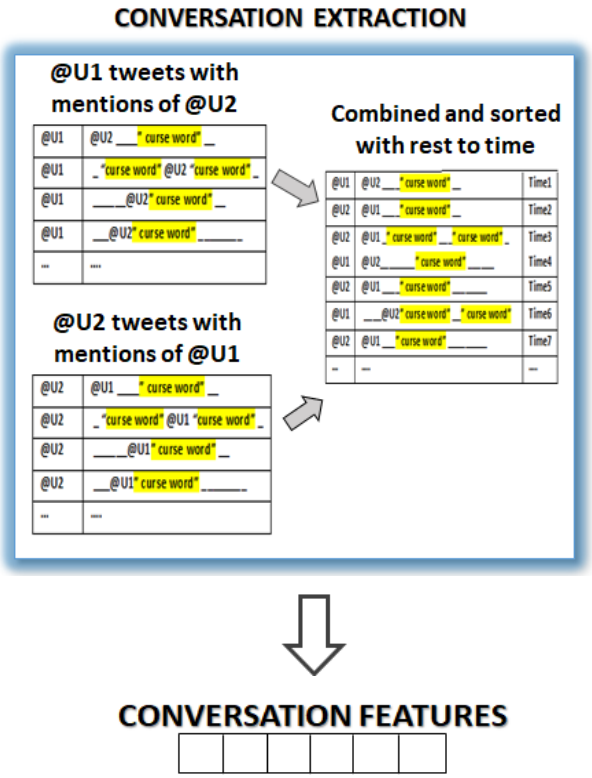


Figure 14. Extracting potential conversation features from a sequence of tweets between the users.

More details about the content features, user profile features, and conversation features are explained in Chapter 4 with their significance to the classification of a tweet as harassing or not.

3.3.3. Classification

Classification component is used to classify a tweet as harassing or not by using a comprehensive set of features generated in the feature extraction component. I use a supervised model to perform the classification task. More details about classifier are explained in Chapter 5 including its performance.

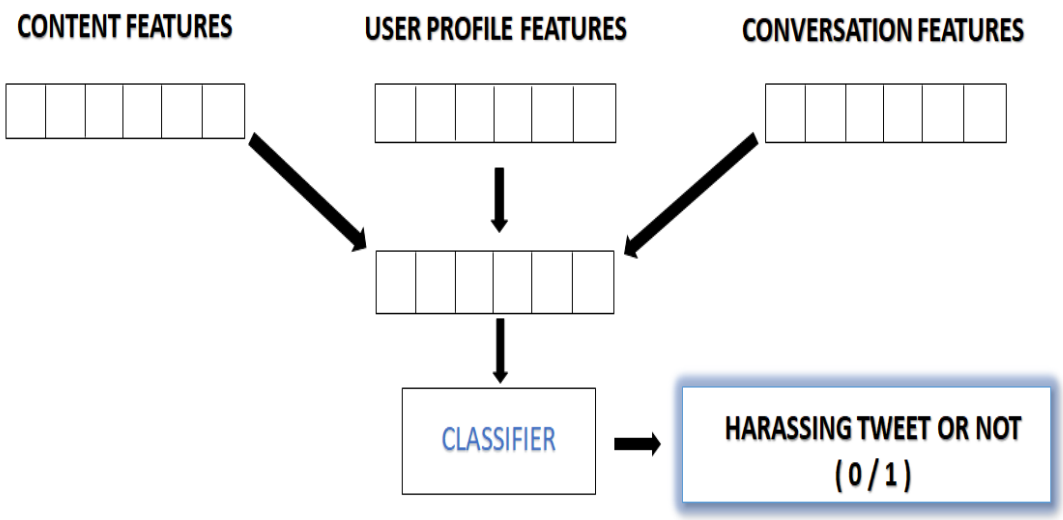


Figure 15. Generating the feature vector by combining the content features, user profile features and conversation features to predict the tweet as a harassing tweet or not.

3.4. Data Collection

I collected a corpus with offensive language because I hypothesize that harassing tweets normally contain offensive language. I tried to detect these using curse

words. However, not all tweets with curse words are harassing. Hence, I had to improve my approach to accurately classifying the tweets as harassing or not.

I used a lexicon developed by Wenbo et al. [11] that contains 788 words that include both curse words and their hieroglyphs (e.g., f*ck, f**k, @sshole). I used a lexicon-based approach for collecting tweets using the Twitter streaming API for one month. I retained tweets that contain "@username" to preserve conversation. From these data, I used the twitter handles of the users to extract the entire available conversation between them. In the data collection phase, I removed celebrity accounts and protected accounts because celebrities rarely respond to any abusive content directed at them, and it is not possible to extract tweets from protected accounts. I also removed the retweets and duplicate tweets from my data set to eliminate noise.

3.4.1. Annotation

One of the major hurdles I faced when annotating the data was that the annotators had to use the whole conversation between the users. Annotating an individual tweet is a significantly easier job. Some of the conversations are very long (based on the number of tweets in a conversation), and the annotators had to read the entire conversation to understand the relevant context and intent of the tweet. My dataset included 72 conversations each containing around 250 tweets and 47 conversations each containing more than 1000 tweets which made the annotation

task arduous. Figure 16 shows the frequency of conversations based on the conversation length, where conversation length is defined as the number of tweets within the conversation. I calculated the inter-rater reliability in terms of Fleiss Kappa [22] among my annotators. I obtained a kappa value of 0.65 which indicates substantial agreement among the annotators [23].

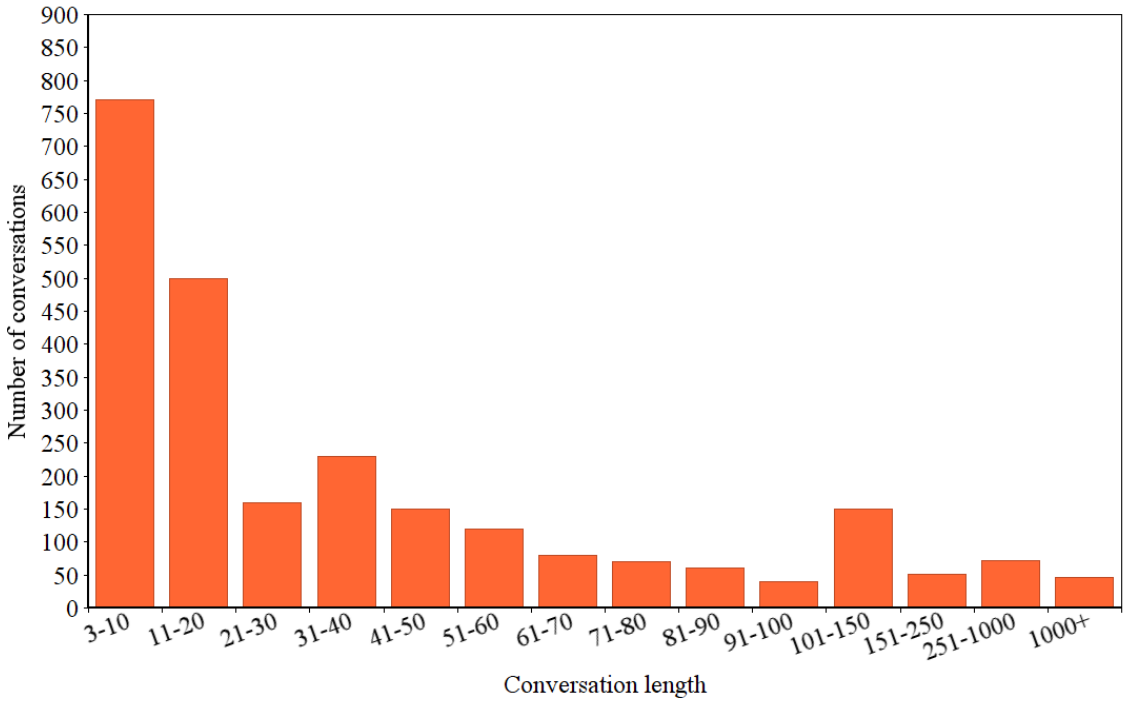


Figure 16. The frequency of conversations between the users by length.

4. Features

After preprocessing the data as explained in Chapter 3, I divided the features used in my model into three categories based on the scope of each feature as mentioned in my thesis statement.

i. Content features

ii. User profile features

iii. Conversation features

4.1. Content Features

Content features are those having the scope only over the tweet. These features help in analyzing tweet for harassment locally. The content features are as follows:

- **The density of curse words:** Based on the findings by Huang et al. [8], the density of curse words in a tweet correlates with it being harassing. So, I used curse word dictionary to count the number of curse words in a tweet and normalize it with tweet length.

$$\text{Density}(T_{cw}) = \frac{number\ of\ cursing\ words\ in\ a\ tweet}{tweet\ length}$$

where tweet length is the number of words in a tweet.

- **The density of capital letters:** Based on the findings of Dadvar et al. [9], I chose the density of capital words because they are more often used to express one's strong feelings in tweets which might end up being harassing.

$$\text{Density}(T_{cl}) = \frac{Number\ of\ capital\ letters\ in\ a\ tweet}{tweet\ length}$$

where tweet length is the number of letters in a tweet.

- **Parts of Speech:** A tweet can contain many non-standard lexical items and syntactic patterns. To identify these items and patterns related to harassment, I used the CMU POS tagger specifically designed for online conversational text [9]. I count the number of nominal + verbal or verbal + nominal, proper noun + verbal, URL or email address, subordinating conjunction, nominal + possessive, and predeterminers as some features to my classifier.
- **Emoticon count:** Emoticons usually indicate the emotion of a user. So, I count the number of emoticons present in a tweet. Harassing tweets very rarely contains emoticons in them whereas non-harassing tweets contain more number of emoticons as shown in Figure 17.

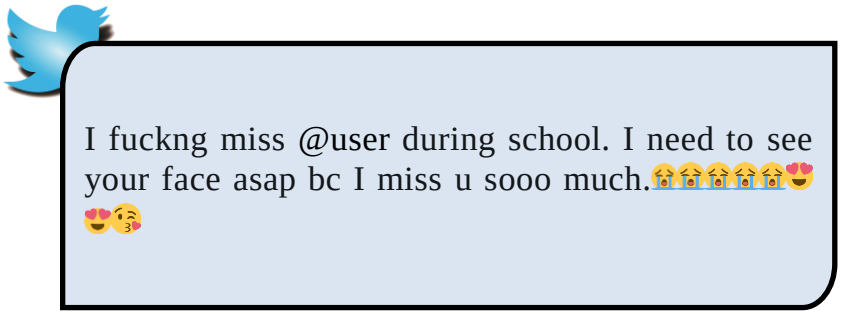


Figure 17. A non-harassing tweet containing more number of emoticons.

- **The frequency of exclamation and question marks:** Huang et al. [8] state that exclamation and question marks are related to emotional tweets and are often connected with harassment. So, I counted the number of exclamation and question marks present in a tweet.
- **Individual count of top twenty curse words:** Past study by Wang et al. [11] listed the top twenty words that are used on Twitter to curse one another. The individual counts of those twenty words are considered.
- **Tweet affect score:** According to the Oxford dictionary[11], "***affect somebody"*** is to *"make somebody have a strong feeling of sadness, pity, etc.*". To know the affect of a tweet on the mentioned user, I calculated the *affect score* by using the Warriner resource [12]. Warriner resource [12] is a file containing approximately14000

[11] https://www.oxfordlearnersdictionaries.com/definition/english/affect

English words with their valance, affect, friendliness, and dominance scores. These 14000 words are commonly used English words and include curse words. I calculated the overall affect score of a tweet by adding the affect score of each word present in the tweet.

$$\text{Affect Score}(t_{as}) = \sum_{i=1}^{j} A_i$$

where,

j is the number of words in a tweet **t**.

$\mathbf{A_i}$ is the affect score for the word '**i**' in tweet **t**.

4.2. User Profile Features

User profile features are those having the scope of the profile of the tweet's author and mentioned user referenced in the tweet. These features inform the relationship between the users. The user profile features are as follows:

- **Friend or not:** It is assumed that if the tweet's author and the mentioned user are following each other, they are friends. The exchange of curse words between a specific pair of friends may be usual, and tweets containing those words are not intended to be harassing. So, I include this as one of the features of my model to improve reliability by reducing false positives.

- **Follower count:** I assume that the person who has a harassing behavior may have only a few followers because people may want to disassociate themselves with such people. It can also represent the social popularity and likability of a user.

- **Following count:** I also assume that the person who has harassing behavior may follow fewer users. Following count determines the size of the social circle of a user.

4.3. Conversation Features

Conversation features have a scope over the past sequence of tweets exchanged between tweet's author and mentioned user present in the tweet. These features help us to learn the actual context and intent behind the original tweet. The conversation features are as follows:

- **Length of conversation:** This is one of the most indicative features to identify the relationship between two users. In my research, I found that the length of the conversation is high among the users that are friends.

- **One way or two-way conversation:** This important conversation feature represents whether only one user is directing messages to the mentioned person or whether both users are engaged in a dialogue.

In the former case, the probability is higher that the tweet is harassing given that the harasser might be sending tweets while the victim may be silently enduring it without responding.

- **The frequency of curse words:** I included the frequency of curse words, i.e., the total count of curse words present in an entire conversation between the users. Friends who naturally exchange curse words frequently will have more curse words in their conversation. So, the tweets that have an abnormal number of curse words can be non-harassing tweets based on other information.

- **Individual count of top twenty curse words:** As mentioned above, I identified the topmost twenty cursing words used on Twitter using Wenbo et al. [11] study. By using that I calculate the individual count of each cursing word used as a feature in my model. I used the same feature in content features where I count top twenty curse words in one tweet only, but here I counted those curse words in their conversation to know the usage of these words in the conversation between the users. These features in tweet and conversation are combined to improve the classification task.

- **Cursing percentage of the tweet with respect to their conversation:** I calculate the cursing percentage of a tweet by using the cursing words present in both the tweet and the conversation surrounding it. If the tweet contains more curse words and conversation also has more curse words, then cursing percentage decreases, which indicates that both the users have a habit of exchanging curse words regularly. Otherwise, the tweet contains more curse words than the normal tweet in a conversation, implying greater cursing percentage and that the tweet might be harassing. I observed that both tweets present in Figure 18 (a) and Figure 18 (b) contain an equal number of curse words, but they differ in the cursing percentage of the respective tweets because the number of curse words present in their respective conversations differed according to their usage of curse words in conversation.

$$\text{Cursing percentage} = \left(\frac{Number\ of\ curse\ words\ in\ a\ tweet}{Number\ of\ curse\ words\ in\ a\ conversation}\right) \times 100$$

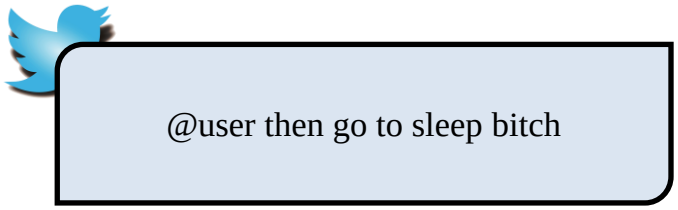


Figure 18(a). Non-harassing tweet

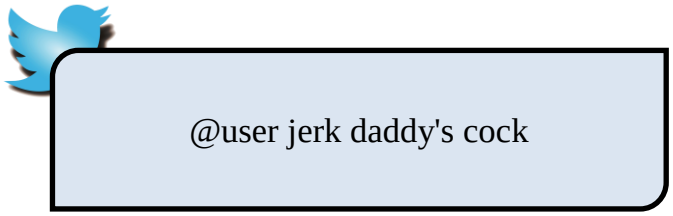


Figure 18(b). Harassing tweet

Figure 18. Tweets containing the same number of curse words but different cursing percentage relative to surrounding conversation.

- **Change of behavior during the conversation:** I divide the entire conversation into five equal parts and count the number of cursing words present in each part. This provides five additional features in my model. These features are meant to capture the sudden changes in the cursing behavior of the users while communicating online. It is meant to capture the temporal aspects of the conversation. Figure 19 represents the change of behavior during the conversation related to the harassing tweet. The number of curse words present in the conversation from the third part to the fourth part rises. Figure 20 represents the change of behavior during the conversation related to non-harassing tweet, using a scale with a much lower base rate. The

number of curse words present in each slice or part of the conversation is small and constant. Conversation among the friends rarely employs curse words.

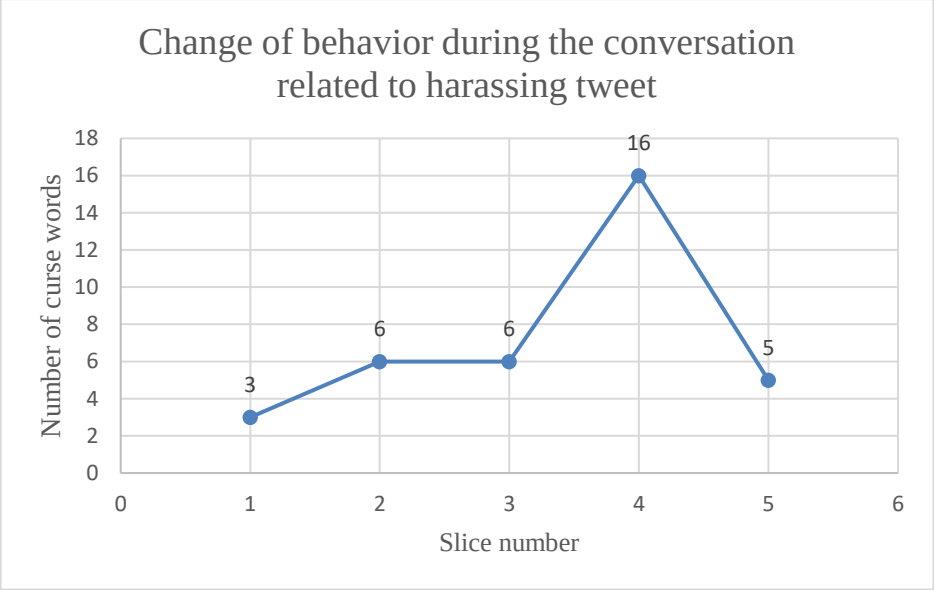


Figure 19. Change in curse word rate related to harassing tweet.

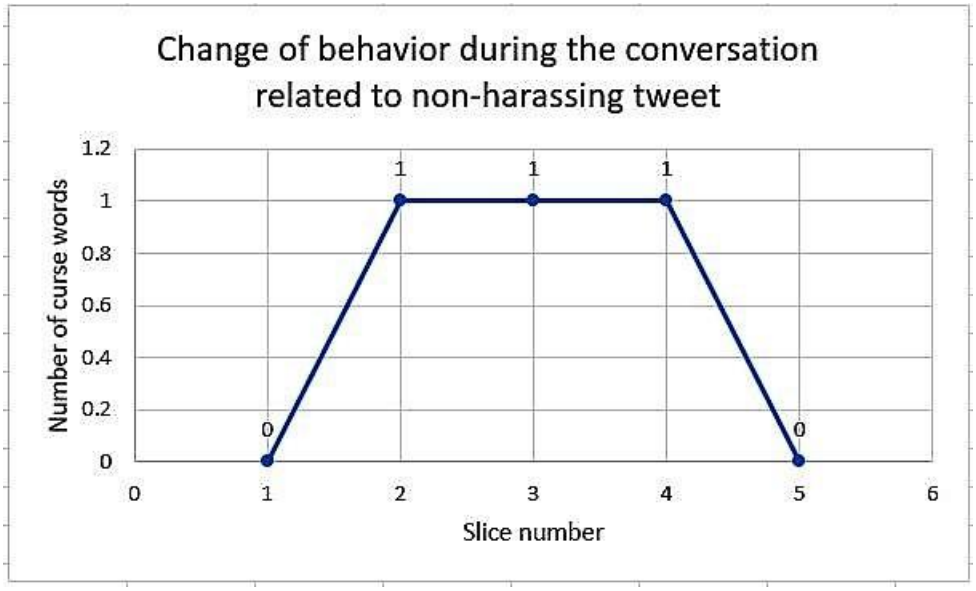


Figure 20. Constant in curse word rate related to non-harassing tweet.

5. Experiments and Evaluation

In this Chapter, I discuss how I handled dataset class imbalance problem, different machine learning algorithms used in the classification task, my evaluation metrics, the performance of my classifier, and finally, discuss challenges I faced while selecting features for the classifier.

By their very nature, harassment datasets have highly imbalanced classes because harassing tweets make up only a small proportion of all exchanged tweets. To deal with this class imbalance problem, I applied "SMOTE" (Synthetic Minority Oversampling TEchnique) approach [16] to my training dataset to create a balanced data set. SMOTE works by oversampling the minority class and under-sampling the majority class. I used Weka 3.69 to apply the SMOTE algorithm to my data and for the classification task. SMOTE will produce few duplicate entities to balance the classes in the dataset. Those duplicate entries can bias classification. So, I removed all the duplicate entries present in my dataset to improve the reliability of classification. Now, the dataset used for classification has unique entries that will increase the performance of the classifier when I run on unseen data it.

5.1. Machine Learning Classifiers

I evaluated my approach using several different classification algorithms such as meta-algorithm (Bagging), regression algorithm (Simple Logistic regression), and tree-based algorithm (Random Forest).

Breiman[12] first proposed the bagging algorithm I use, known as Bootstrap aggregating. It improves classification task by aggregating the performance of each classifier applied on a random redistribution of the training dataset[12] [13]. Bagging improves the stability and accuracy of the models used in regression and statistical classification by avoiding the overfitting problem and reducing variance present in the dataset[12].

Simple logistic regression uses a black box function (called SoftMax) to measure the relationship between a categorical dependent variable and independent variable[14]. It is useful only for a binary classification problem.

Random forest aggregates several learning methods including classification and regression. Random forest works by constructing multiple decision trees during learning from data and outputs the class, which is either the mode of the classes in classification problem or mean of the individual trees in regression[15].

[12] https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume11/opitz99a-html/node3.html
[13] https://en.wikipedia.org/wiki/Bootstrap_aggregating
[14] https://dataaspirant.com/2017/03/02/how-logistic-regression-model-works/
[15] https://en.wikipedia.org/wiki/Random_forest

5.2. Evaluation Metrics

I evaluated my work based on an F-Measure [6,4] and AUC of ROC [13]. I also considered the confusion matrix to check whether the classifier is biased towards any class. Since my solution addresses imbalanced datasets, accuracy is artificially high if the classifier is biased towards the majority class. For example, let us consider 10k non-harassing tweets and 100 harassing tweets in a dataset. If my classifier classifies 10090 as non-harassing tweets and only 10 as harassing tweets, then the accuracy of the classifier will be over 99.9 percent. In fact, classifying all the tweets as non-harassing still yields 99% accuracy! So, accuracy is a poor evaluation metric because trivially claiming all tweets to be non-harassing can still yield high accuracy.

To calculate these evaluation metrics, I need a confusion matrix, i.e., the matrix in which rows contain the actual class labels and columns contain the classifier predicted class labels.

$\frac{Predicted \rightarrow}{Actual}$	**Non-harassing**	**Harassing**
Non-harassing	True Negative	False Positive
Harassing	False Negative	True Positive

Table 3. Sample confusion matrix

F-Measure is defined as the harmonic mean of precision and recall [6, 4].

$$\text{F- Measure} = \frac{2 \times precision \times recall}{precision + recall}$$

Precision [6, 4] can be defined as the number of correctly classified harassing tweets to the total number of tweets predicted as harassment.

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

Recall [6, 4] can be defined as the number of correctly classified harassing tweets to the total number of harassing tweets.

$$\text{Recall} = \frac{tp}{tp + fn}$$

Receiver Operating Characteristic curve is true positive rate plotted against the false positive rate [13].

5.3. Performance Evaluation

I want to recall the statistics about my dataset. The size of the annotated data is 2500 tweets where 2000 tweets are used for training the model using 10-fold cross validation whereas 500 tweets are used as unseen data to evaluate the performance of the model.

First, I ran my experiments on training data using the 10-fold cross-validation. I obtained the results for each type of features in insolation and then using various

combinations. I also explored different classifiers to match it with the data characteristics. I tested my algorithm on the 500 unseen tweets (test data) and performed the analysis.

Table 4 and Table 5 showed the performance of different classification algorithms when I considered only the content features and only the conversation features respectively. Among all the algorithms, Random Forest performed the best, outperforming bagging algorithm and simple logistic regression. The classifier with only conversation features performs almost similar to the classifier using only content features.

$\frac{Evaluation\ metrics}{Classifier}$	**F-Measure**	**ROC**
Bagging	0.74	0.817
Random Forest	**0.805**	**0.882**
Simple Logistic Regression	0.674	0.736

Table 4. Classification results for the content model using different algorithms.

$\frac{Evaluation\ metrics}{Classifier}$	**F-Measure**	**ROC**
Bagging	0.75	0.824
Random Forest	**0.788**	**0.867**
Simple Logistic Regression	0.643	0.688

Table 5. Classification results for conversation model using different algorithms.

$\frac{Evaluation\ metrics}{Classifier}$	**F-Measure**	**ROC**
Bagging	0.802	0.877
Random Forest	**0.882**	**0.943**
Simple Logistic Regression	0.713	0.788

Table 6. Classification results for the composite model using different algorithms.

After analyzing the performance of content features and conversation features separately, I combined the content features, the conversation features, and the user profile features. The classification results for different algorithms such as bagging,

random forest, and simple logistic regression were reported in Table 6. The combination of these three sets of features helped us to reduce false positives and increase the performance of the classifiers.

This shows that the context gleaned from these features play an important role in understanding the semantics and intent of the tweet. A comparative analysis of all the classification algorithms shows that Random Forest performs the best (Table 6).

Now I would like to recall my thesis statement: *"Harassment detection on Twitter can be improved by harnessing contextual information from the sequence of messages (conversation) exchanged between users in addition to using tweet content and user profile information."* As I mentioned in the above statement performance of harassment detection classifier improved as shown in above results.

The following confusion matrix represents the classification results on unseen data.

$\frac{Predicted \rightarrow}{Actual}$	**Non-harassing**	**Harassing**
Non-harassing	221	32
Harassing	24	223

Table 7. Confusion Matrix for unseen (test) data

Now, I determine evaluation metrics on unseen data by using the formulas mentioned above.

Precision = 0.87

Recall = 0.90

F-Measure = 0.884

5.4. Comparative evaluation with respect to previous studies

Initially, I tried to get data from previous studies related to harassment on Twitter but could not. For a fair comparison, I have implemented some of the features used by previous studies that are essentially related to the content of the tweet.

Since Chen et al. [17] did not provide any quantitative information on the performance of their system, I was unable to evaluate the proposed classifier's performance against their counterparts.

The following Table 8 shows the results of Huang et al. [8] on their textual analysis. My content-based features and their textual features are similar except for the parts of speech tagger and additional content based features I use to improve the performance.

	Huang et al. [8]	My approach
ROC using Bagging	0.79	0.817
TPR using Bagging	0.715	0.74
ROC using Random Forest	0.831	0.882
TPR using Random Forest	0.748	0.805

Table 8. Comparison with Huang et al. [8] approach using only textual features and my approach using only content features.

5.5. Discussion

In this Section, I discuss the difficulties I faced while selecting the features of the classifier under:

a) N-grams
b) Sentiment

5.5.1. N-grams

Most of the studies on Twitter [20, 21, 5] use N-grams for classifying tweets. So, I also started with N-grams as one of the content features to differentiate harassing and non-harassing tweets. Once I built a classifier using N-grams as a feature and evaluated the performance, I got an F-measure less than 40 percent.

Later I analyzed the reason for the poor performance of the classifier. I separated the data into harassing data and non-harassing data based on the annotation. Then I ran my program to find top 5 unigrams and top 5 bigrams of both harassing data and non-harassing data. Shockingly, almost 80% of the unigrams and bigrams are same for both the classes even when I removed all the seed curse words from all tweets. So, the classifier is unable to predict the tweet as harassing or not, based solely on these features as shown in Figure 21. Table 9 and Table 10 shows the top 5 unigrams and bigrams of harassing data and non-harassing data before seed curse words removal and after seed curse words removed respectively.

Among the tweets that use curse words, there is not much divergence in the distribution or frequency of curse words that separate harassing and non-harassing tweets.

Top 5 unigrams of harassing data	Top 5 unigrams of non-harassing data	Top 5 bigrams of harassing data	Top 5 bigrams of non-harassing data
Fucking	Fucking	fuck you	hell yeah
Ass	Shit	hell yeah	fuck you
Bitch	Love	bitch ass	mine shit
Shit	Bitch	flour hoe	eat mine
Hoe	Good	mine shit	fuck it

Table 9. Analysis of N-grams on harassing and non-harassing data before seed curse words removal from tweets.

Top 5 unigrams of harassing data	Top 5 unigrams of non-harassing data	Top 5 bigrams of harassing data	Top 5 bigrams of non-harassing data
Me	You	Get up	Lol wtf
Get	Lol	Got some	Trust me
Up	Me	You are	Get up
You	Bro	They have	You are
Yeah	Up	Trust me	Yeah bro

Table 10. Analysis of N-grams on harassing and non-harassing data after seed curse words removed from tweets.

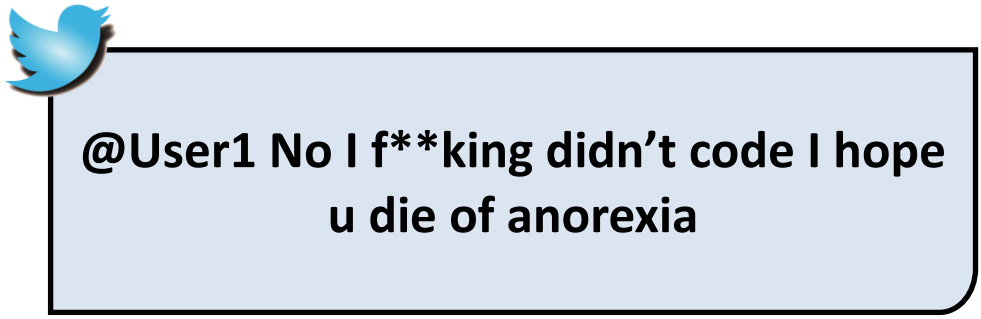


Figure 21 (a) Harassing Tweet

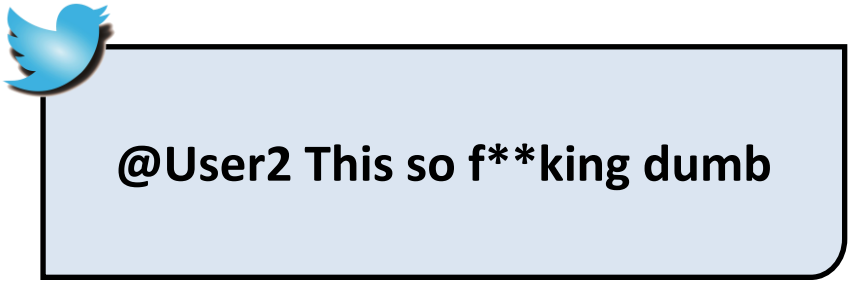


Figure 21 (b). Non-harassing Tweet

Figure 21. Both the harassing tweet and non-harassing containing the same n-gram.

5.5.2. Sentiment

Sentiment is another useful feature in harassment detection because tweet which meant to harass someone usually has negative sentiment. In general, the sentiment of the tweet is detected by using the words present in the tweet. Since my dataset has been crawled using cursing words, every tweet has at least one curse word, which tends to be negative sentiment. So, most of the non-harassing tweets also had negative sentiment.

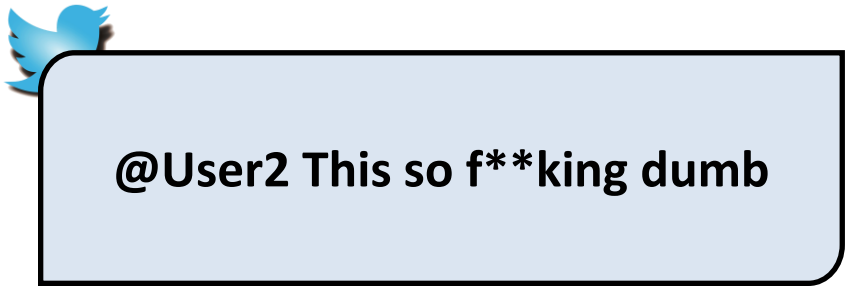


Figure 22 (a). Non-harassing tweet but negative sentiment

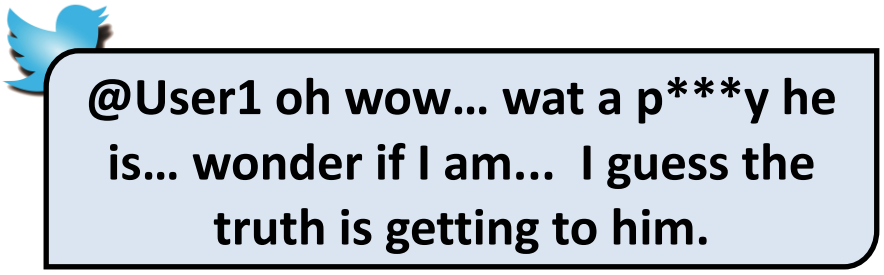


Figure 22 (b). Harassing tweet but neutral sentiment

Figure 22. Sentiment analysis results

So, N-gram and sentiment are not effective features to determinate harassing tweets from non-harassing tweets in the context of my datasets.

5.6. Error analysis

I observed two interesting false positives in my classifier.

1. Two people are talking about the third person without using his/her twitter handle. They are curse words about the third person without knowing her. This tweet and conversation contain features related to harassment, but the sender and the intended receiver are not harassing each other. So, while annotating, my annotators annotate this kind of tweets as a non-harassing tweet, but classifier identifies it as a harassing one. I observed this in Figure 23, where user2 is mentioning user1 and tweeting about another person named “Sam” in the tweet. It is one of the examples of false positives generated in my classifier.

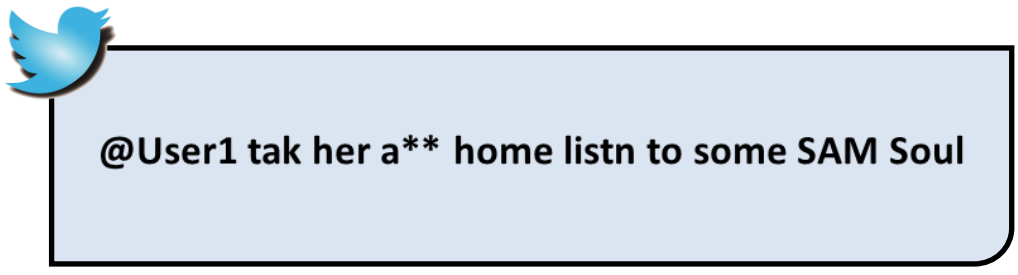


Figure 23. Example of false positive generated because of users tweeting about the third person without using his/her Twitter handle.

2. There are some people who are so fascinated by the online game and that they tweet about the game by mentioning another person who may be his/her partner in the game. I observed in Figure 24, that whereas two users are tweeting about the game, even though it looks like they are harassing each other.

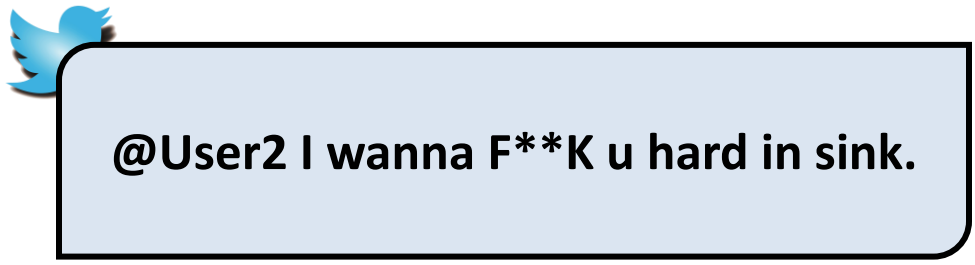


Figure 24. Example of false positive generated because of users tweeting about the game.

6. Conclusion and Future work

In this thesis, I designed a framework that can identify harassing tweets on Twitter. I collected the tweets using a curse words dictionary with 788 words and Twitter streaming API. After collecting the tweets, I extracted the conversations using the Twitter API, and Twitter handles of the tweet author and the user mentioned in the tweet. I used the SMOTE technique to create a balanced dataset for classification. Then I came up with a comprehensive set of discriminating features based on three categories: content features, user profile features, and conversation features.

I presented a novel approach that incorporates conversation-based features to improve detection and determination of tweet as harassing or not. From this study, I not only successfully improved the performance of the classifier but also explored meaningful features (such as one-way or two-way conversation, frequency of curse words in conversation, curse percentage of a tweet, friends or not, tweet affect score, density of curse words, and density of capital letters) and explained their significance for harassment detection.

I experimented using different machine learning classifiers and found that random forest with features mentioned in Chapter 4 works the best for my problem. The supervised machine learning classifier I used achieves an F-score of 88.2 and AUC for ROC of 94.3.

In future, I can use this work in detecting victims and harassers on Twitter by identifying the user in harassing tweets. The proposed framework can be used to identify the threatening tweets by slight modifications such as changing seed terms to reflect threatening context, use threatening words dictionary in place of curse word dictionary while generating features, learning the algorithm based on the annotated training data with respect to threatening context. This work can be further enhanced by adding features from social networks such as number of nodes, number of edges, number of links in a relationship graph for every tweet along with content features, user profile features, and conversation features.

Bibliography

1. Chatzakou, Despoina, et al. "Mean Birds: Detecting Aggression and Bullying on Twitter." *arXiv preprint arXiv:1702.06877* (2017).

2. Raisi, Elaheh, and Bert Huang. "Cyberbullying identification using participant-vocabulary consistency." *arXiv preprint arXiv:1606.08084* (2016).

3. Zhao, Rui, Anna Zhou, and Kezhi Mao. "Automatic detection of cyberbullying on social networks based on bullying features." *Proceedings of the 17th international conference on distributed computing and networking*. ACM, 2016.

4. R. Baeza-Yates, B. Ribeiro-Neto et al., Modern information retrieval. ACM press New York, 1999, vol. 463.

5. Muppalla, Roopteja, et al. "Discovering Explanatory Models to Identify Relevant Tweets on Zika."

6. C. D. Manning, P. Raghavan, and H. Sch¨utze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.

7. Pennebaker, James W., et al. *The development and psychometric properties of LIWC2015*. 2015.

8. Huang, Qianjia, Vivek Kumar Singh, and Pradeep Kumar Atrey. "Cyber bullying detection using social and textual analysis." *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*. ACM, 2014.

9. Owoputi, Olutobi, et al. "Improved part-of-speech tagging for online conversational text with word clusters." Association for Computational Linguistics, 2013.

10. Dadvar, Maral, et al. "Improving Cyberbullying Detection with User Context." *ECIR*. 2013.

11. Wang, Wenbo, et al. "Cursing in english on twitter." *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 2014.

12. Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. "Norms of valence, arousal, and dominance for 13,915 English lemmas." *Behavior research methods* 45.4 (2013): 1191-1207.

13. Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern recognition* 30.7 (1997): 1145-1159.

14. Nahar, Vinita, Xue Li, and Chaoyi Pang. "An effective approach for cyberbullying detection." *Communications in Information Science and Management Engineering* 3.5 (2013): 238.

15. Chen, Ying, et al. "Detecting offensive language in social media to protect adolescent online safety." *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012.

16. Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

17. Chen, Yunfei, et al. "4Is of social bully filtering: identity, inference, influence, and intervention." *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012.

18. Dinakar, Karthik, Roi Reichart, and Henry Lieberman. "Modeling the detection of Textual Cyberbullying." *The Social Mobile Web* 11.02 (2011).

19. Yin, Dawei, et al. "Detection of harassment on web 2.0." *Proceedings of the Content Analysis in the WEB* 2 (2009): 1-7.

20. Rajadesingan, Ashwin, Reza Zafarani, and Huan Liu. "Sarcasm detection on Twitter: A behavioral modeling approach." *Proceedings of the Eighth*

ACM International Conference on Web Search and Data Mining. ACM, 2015.

21. Bamman, David, and Noah A. Smith. "Contextualized Sarcasm Detection on Twitter." *ICWSM*. 2015.

22. McHugh, Mary L. "Interrater reliability: the kappa statistic." *Biochemia medica* 22.3 (2012): 276-282.

23. Landis, J. Richard, and Gary G. Koch. "The measurement of observer agreement for categorical data." *biometrics* (1977): 159-174.

24. Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using machine learning to detect cyberbullying." *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE, 2011.

25. Dadvar, Maral, et al. "Improved cyberbullying detection using gender information." (2012).

26. Ortony, Andrew, Gerald L. Clore, and Mark A. Foss. "The referential structure of the affective lexicon." *Cognitive science*11.3 (1987): 341-364.