INFLUENCE OF TIE STRENGTH ON HOSTILITY IN SOCIAL MEDIA

BY

BAHAR RADFAR

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science
in the Graduate College of the
Illinois Institute of Technology

Approved __________________________
Advisor

Chicago, Illinois
May 2019

ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to my advisor Prof. Aron Culotta for the continuous support of my research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research. I could not have imagined having a better advisor and mentor for my master's study.

It is my fortune to gratefully acknowledge the support of my close friend, Dr. Saeid Barati for his support and generous care throughout the research tenure. He was always beside me during the happy and hard moments to push and motivate me.

Last but not the least, I would like to thank my family, my parents and my brother for the encouragement, moral support, personal attention and care during this thesis and overall in life.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Online anti-social behavior, such as cyberbullying, harassment, and trolling, is a widespread problem that threatens free discussion and has negative physical and mental health consequences for victims and communities.

While prior work has proposed automated methods to identify toxic situation such as hostility, they only focused on individual words. While only a bag of keywords is applied to detect hostility, this is not enough as words might have different meaning based on the relationship between participants of the discussion.In this paper, we considered the friendship between sender and the target of a hostile conversation. First, we studied the characteristic of different types of relationship. Then, we set our goal to be more accurate hostility detection with reduced wrong red flags.

Thus, we aim to detect both the presence and intensity of hostile comments based on linguistic and social features from our well-defined relationships. To evaluate our approach, we introduce a corpus of over 12K annotated Twitter tweets from over +170,000 tweets. Next, we extracted useful features such as relationship type and length of the tweet to feed into our Long Short Term Memory(LSTM) and Logistic Regression(LR) classifier.By considering the relationship type in the classifier model we improved the hostility detection AUC by close to 5 % comparing to the baseline method. Also, the F-1 score increased by 4 % as well.

CHAPTER 1
INTRODUCTION

Over the past decade, the grown portion of human communication is occurring over social media. The main characteristic of social media is the constant connectivity which happens via cell phone or laptop anytime anywhere. Hence, regardless of the geographic location, time, or physical limitations, the users participate in various communication threads.

Social media outlets have become the largest point of insight into human communications. Although, the harassment and the aggressive and antisocial content inside, is a persistent problem. The amplification of social connectivity has led to a significant amount of hostility and cyberbulling.

The fact that the conversation could go on and on for an infinite time period, it provides a higher chance of heated conversation which could result in toxicity. In addition, by the essence of social media, that is easy access for users to join the public discussions, it is more probable that arguments go wrong further. However, none of these reasons as strong as the fact that cyberbullying does not include the emotional reactions face to face. For instance, a user makes an aggressive comment and logs off from the media while not caring about what happens later or if they are using a fake account (not associated with their real name). In the real world, it is possible that people cross their path again and dealing with the consequences of aggressive behavior in real life is more challenging and complicated. Hence, if we could detect and predict the toxic situation, we could prevent the later conflicts.

On social media, the feelings and emotions may not be passed through as easily as face to face conversations but still, some rules apply to both of these types of conversations. Words meanings and the conversation boundaries change depending on the type of relationship between the people who engage in a conversation.

A significant body of literature is done on cyberbullying in social media. Existing approaches include automated techniques, crowd-sourcing moderation, and user control. Some researchers focused on predicting whether a conversation is going wrong or not[9, 14, 8, 15]. Others have emphasized on the computational aspect of toxic content detection. Semantic variation of words, linguistic changes, and evaluation of users/communities are the goals. [7, 12, 4]. Furthermore, by extracting norms and hidden rules within the large community, more toxic content was discovered [2].

For this paper, we focused on hostile situations among all different classes of unacceptable content. Hostility comes from the Latin origin of "hostilis", which mean unfriendliness, opposition, and acts of warfare. A hostile conversation includes one or more words that represent harassing, bullying, threatening, and sexual language towards an individual or a group.

The prior research mostly focused on defining strict rules on how to detect hostility based on the text context itself. This is not sufficient for accurate hostility modeling as the word meaning changes based on the relationship between participants of an ongoing discussion. Close friends might use hostile known phrases for the appreciation of a specific situation or simply as a joke towards their close friend. This case motivates us to further analysis of the actual meaning of the common hostile words.

In this paper, we propose defining relationship categories for hostility detection analysis. In order to obtain the actual meaning of the words rather than their common use cases, we introduce four types of relationships that happen in two-way friendship social media such as Twitter and Instagram. These categories of relationships are:

1. Dual Friendship (two-way)

2. Sender follows the target (one-way)
3. Target follows the sender (one-way)
4. No Friendship (no connection)

By analyzing the behavior of the aforementioned groups, we constructed a set of characteristics which are unique and expresses a significant gap between the actual meaning of the words in each group. This case is more observable in the first category as friends tend to use hostile words more often in non-toxic situations.

We evaluate our approach method on a newly constructed dataset of 6.7K tweets from Twitter. We used a set of hostile words from HateBase and other online resources to collect the data.The tweets have at least a seed word, a sender, and a target. Then we asked annotators to annotate the tweet whether it's hostile or not . And if the tweet is hostile, then is it direct toward the target or not.Afterwords, we used the tweet content, length of the tweet, and the relationship type as the input feature for our classifiers.

We implemented a Long Short Term Memory and a Logistic Regression classifiers. We defined a base model that receives all the features except the relationship type. To compare the base and the proposed model, we use Area Under the Carve(AUC) in addition to numerical classification statistics.

As the result, by adding the relationship type definitions to classifier the AUC has improved by 4% and the F1-score has improved by 5% accordingly.

CHAPTER 2
RELATED WORK

Prior works have studied the users' behavior in social media in various computational methods. The first set of articles try to predict antisocial behavior in social media outlets or forums, while others make an attempt to find the relation between group norms and communication behavior. A variety of methods have been proposed for cyberbullying and hostility detection. These methods mostly approach the problem by treating it as a classification problem, where the comments been classified as antisocial or not.

While Zhang et al. and Liu et al. used the data to come up with forecasting methods which can predict if the conversation will go wrong [14, 9], others made attempts to predict the intensity of it in 5-15 next comments[9], or even further predicting whether a user will get banned in the future or not[8, 15]. These papers only detect antisocial and hostile behaviors in the existing comments rather than attempting to predict them[13, 1].

Researchers usually use different types of social media such as Twitter, Facebook, Instagram or forums such as Reddit or Wikipedia as their data source. The categorization happens in two ways in general. First, custom groups are defined and data is being categorized manually applying one or a couple of groups[13]. The second approach is using different unsupervised methods such as K-Special Centroid to identify the various types of conversations[14, 9]. Also, its possible to only use the plain text for the data categorization[1].

With data categorization, important features will be extracted where they are essential in the classification stage. While a group of studies used linguistic features such as Unigram[13], Word2vec, etc[9], others have used more specific features such as the number of comments in each thread [13], deciding based on the time order of comments, if it's the final comment or not[9], the number of people who participated in the conversation, their gender, age, and number of followers or following[1], or even the effects of the authors' mood[8, 3].On the contrary, some have employed the features related to conversation content, for instance, vulgarities[1] or the level of politeness[14]. Nevertheless, these methods are different to one or another for feature extraction, and they applied a combination of features rather than a single feature for more efficient feature engineering.

Comments are not the only content available to be considered as the data source. Another article focused on peoples actions and discussed the fact that although some people may not always write a comment, it is possible they interact via liking the post itself or the existing comments (if liking the comment is available in that specific social media). This can result in an elevated discussion with more and more comments in the future where it might even start or stop a hostile conversation[14].

The second set of computational studies emphasized finding the different semantic variation of words, linguistic change and evaluation of users and communities, and in general, find the words meaning based on their contextual users[7, 12, 4].

Some researchers looked for nouns and generated graphs to keep track of words that appear together in similar context. The graph is a co-occurrence graph, where the nodes represent the terms and the edge between them indicates their co-occurrence in a context. As the number of edges between two nodes increases, their weight and importance will increase as well. Therefore, they pruned the graph and only kept the ones that were higher than the threshold. They used the number of shared neighbors as their similarity function and have tried to detect different communities using clustering[7]. Also, the linguistic features can be applied to differentiate semantic variation of a word [12].

Others have tried to extract norms and hidden rules inside a big community by defining micro, meso, macro as the different levels of the norm. After collecting and preprocessing, k-mean clustering has been used on the prediction matrix. As a result, they ended up with the cluster of subreddits which share norms among themselves at three different levels (macro, meso, and micro). They have used topic modeling and open coding to extract these norms[2].

Another set of articles have worked on the users' level adaptability during their life cycle, which started with users writing their first comment and ended by leaving the website. They have realized that at the beginning the users are more flexible and they mostly try to learn the norm of the community. They also realized that when the users stop learning they mostly leave the website as well[4].

Investigating the prior work on the computational analysis of social media leads us to realize that hostile words can not always be considered the same way. Not only the words can have different meanings based on the conversation's topic, but also they can be defined differently depending on the type of the relationship between the sender and the receivers of the conversation.

The hostility detection can not be used as a cookie cutter with a predefined list of words. We should look at each conversation separately and consider the type of relationship between the participants and redefine each word's meaning based on that.

Our approach takes each tweet, extract the participants, their relationship and then extract the meaning of the tweet. This will help us to find which words can be used both as hostile and nonhostile and also what may cause that. For instance, some words, which previously labeled as the top 10 used hostile words on social media, can also be used as an act of surprise or even only for exaggeration purposes. However, one of the biggest challenge here is on how to predict the hostility in the future if we are not aware of relationship between the participants.

CHAPTER 3
DATA ANALYSIS

3.1 Motivation

In this section, we describe our procedure that collects hostile discussions from Twitter and annotates the tweets accordingly.

Our goal is to detect hostility more accurately, specially in cases where there is a relationship between the participants. We tried to identify the exact connection between the sender/target of the posts when a possible hostile situation might going on.

In general, social media is divided into two groups according to their type of users relationship:

1. One-way friendship: Social media, such as Facebook, have a one-way following mechanism. On these social media, friendship can be designed as a non-directed graph with peoples as nodes and relationships as the edges. A graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge and there can exist at most one edge between each pair of nodes.
2. Two-way friendship: Social media, such as Twitter and Instagram, have a two-way following mechanism. On these social media, friendship can be outlined as a directed graph. Each user can be presented as a node and relationships can be presented as directed edges between pairs of nodes. If there be an edge between node A and node B, which points to B, then we can say A is friend with B, but not the other way around.

Knowing the type of relationship between the sender and the target of the message is a key part of this research. Therefore, we selected Twitter from the second group to be used in our evaluation. The fact that Twitter has a better API and more text-based environment made it a better option than Instagram.

3.2 Data Collection

For data collection, we created the dataset using the Twitter API as the social media of interest. Data is collected from newly posted from January until the end of March 2019. We originally recorded +200,000 tweets which later re-sampled 12,000 tweets to keep the data balanced and be able to label the tweets in the later stages of this research.

3.2.1 Raw Data. To detect hostility in tweets, the content should be parsed and analyzed to find inappropriate words or phrases. We looked for tweets with words that are known for their high level of hostility and then checked whether the relationship type has changed their meaning (in this case, one or both of them unfollow the other user) or they have lowered the hostility in their heated-up discussion.

The first step was to gather a list of hostility representative words which score high on the hostility spectrum. We used a combination of different methods to generate the seed words.

1. Our Primary Source of hostile words was Hatebase[0], an online lexicon of words that are used primarily in Hate Speech. One advantage of using Hatebase is that for every word in its database, it also has information about the number of sightings, which allows ranking every word in its vocabulary according to this value. Since not all hostile words are related to hate speech, we used other online sources to collect more words that were used in a hostile context.

2. We took an intersection of all the hostile words list we found online[6, 10] using

different lists such as Google list of bad words and top swear words[11], and combined it with the top 20 most used Hate Speech words from Hatebase [5].

As a result, we end up with the 65 words such as n-word, f-word, and c-word which then been used as the hostility seed words (The complete list of hostile words can be found in the appendix)

Using the TwitterApi data stream collection and the filter features, we collected over 200,000 tweets between January and March 2019. This dataset only contains the tweets which at least contains one of the seed words, been posted in reply to a tweet, and the target person is not the same as the sender.

In addition to the tweet content, we also needed to collect the relationship between the original poster and the target exactly after the tweet has been posted because their type of relationship might change after a hostile argument. Therefore, after receiving each tweet, we collect their relationship using the ’friendships/show’ request.

The next step was to classify each tweet based on its type of relationship. Twitter is a two-way friendship type of social media, thus, if we assume ”A” write a comment that includes a hostile word in reply to ”B”’s previous comment, then at the time the tweet been posted, they can be classified as one of the following four groups :

1. Dual friendship: ”A” follows ”B” and ”B” follows ”A” as well.
2. No Friendship: Neither ”A” nor ”B” follows the other person.
3. One-way Friendship
   (a) Sender follow Target: Only ”A” follows ”B”.

(b) Target follow Sender: Only ”B” follows ”A”.

Another point of view is to see if these comments can affect their participants' type of friendship in any special way and also if there is any difference between how other people will react to different groups of these tweets. Therefore, we have waited for two weeks and we have updated our dataset for each tweet, using their unique ids.

We also collected each person last 200 tweets, for both senders and targets, and store them in a separate file. These data are used as a conversation outlines for our in prediction purposes.
All collected data have then been preprocessed and checked with items below:

- The tweets have been replaced with their full-length version of them.
- Tweets have been replaced with their lowercase version.
- All stop words and links have been removed.
- The emojis have been replaced with their text version.

3.3 Data Annotation

After collecting the tweet, we sampled 3,000 tweets from each type of relationship and merge them together. This results in a huge dataset with 12,000 sampled tweets. The whole idea was to create a large and balanced dataset which then can be employed in our classification models. We stored the information on our samples in the following five main columns:

- id: Contains the unique id of each tweet as the unique key.
- text: The actual full text of each tweet, fed to the parser to examine the possible involvement of hostile words.

- target: The tweets we are using are all non-starter tweets. That means they are posted in reply to another tweet, and the target refers to the writer of the previous tweet which this tweet is replying to.
- sender: The writer of the current tweet.
- link: The link to the actual tweet.

Although the purpose of this research is to search on hostility detection with considering friendship type, we do not want the results to be biased. Therefore, we have removed the relationship information before presenting to the annotators.

These sampled tweets have then been divided into smaller files, with some overlapping between them and been given to 63 annotators and asked them to check if not only they are hostile or not, but also if their hostility is directed toward the target.

> "@*** Lmao my old hurt ass fucked it up wife.. tried to beat him to the punch of fucking me over because Convinced myself he was or would fuck me over when he wasnt doing either one Im glad to say today I HAVE GROWN ."

We have also provided them a file of 20 labeled sample data and a manual which not only defines each column in our dataset but also give annotators some hints to help them to understand the whole process better. For instance, if a message has the following characteristics, then it has a higher chance of being hostile:

- If the message is mostly in Capital Letters (CAPS)
- If the message has excessive letters or exclamations points. It tends to show more emotion towards targeting someone.

> “@*** @*** @*** DGHahahhahhahahhahhahhhhahahhhahhahahahahahhahahah Fuck you.”
>
> “@*** @*** YOU LIED YOUR ASS OFF ON THIS. YOU DID TRY TO GET HILLARY ARRESTED OVER EMAILS.. YOU JUST FAILED...LIAR.”
>
> “BALL DONT LIE BITCH!!!”

More examples of hostile tweets can be found in the Appendix.

After the initial analysis of the labeled data, we realized that the data points are too sparse for the classifier. The reason is that people usually have a different understanding of the same discussion. While one conversation might seem hostile to an annotator, the exact same conversation can be friendly in another annotator point of view. Also, we noticed that annotators needed additional information on the relationship status of the sender/target to have a better insight into the conversation. Thus we decided to re-label data with only three annotators and we made sure that they agree on some examples of hostile tweets.

To this end, annotators first started with annotating the first 500 tweets altogether to get a better understanding of the hostility definition. Then, they annotate over 2000 tweets individually. We also asked them to not label the ambiguous ones that they are not sure about, so the second annotator can go through and label the data. If they disagree on the final label, then we asked the third person to label the data as the tiebreaker. In the end, we ended up with 6.7K of cleaned and relabeled tweets which then can be used in our next steps.

3.4 Data Processing

We now describe our procedure for processing the labels after going through a cleaning session. At a high level, we looked at the tweets from different angles such as length of tweet content and type of the relationship to extract the features.

In tweeter, the discussion happens between two users as we identify them as the sender and the target. The sender uses the hostile words against the target. Hence, we decided to categorize the tweets based on the relationship of sender/target. The four main categories include:

- No Friendship: neither sender nor target is following each other.
- Dual Friendship: Both the sender and the target are following each other.
- Target follow Sender: The sender does not follow the target.
- Sender follow Target: Target is not a follower of the sender.

These categories cover all the possible relationship scenarios on Twitter. After introducing the relationship categories, we attempted to characterize each category in terms of various metrics like their follower base, the total number of likes and retweets, and so on.

For the characterization step, we first need to extract additional information regarding the tweet content and the participants. In some cases, the tweet which has hostile words might be deleted after a while. Looking at the number of likes and retweets helps us to recognize hostile situation better. For instance, more tweet interactions (likes and retweets) increases the chance of being shown in the explore section of the Twitter. The explore section includes highly rated, trends of the day, or viral tweets. Therefore, if a tweet is shown in the explore section, there is a higher probability that other users will see that tweet and consequently the discussion could elevate more quickly.

The additional updated information for the tweets per relationship category includes :

- The average length of a tweet: The longer tweets might indicate a hostile situation is going on as the sender is rambling with hostile words.
- The ratio of deleted tweets: In a hostile situation, users are more likely to delete their tweets if the discussion elevates.
- The average number of followers: When the target has a high number of followers (assume the target is a celebrity), the hostile situation happens more often as users might be salty and offensive to the celebrity while if the tweet was coming from a normal person (with a lower number of followers), the hostile situation would probably not happen.
- The number of likes/retweets: More tweet interaction might result in being more seen by other users, and causes a hostile situation.

To collect the additional information, we used the tweet and users' unique ids. Although, we waited two weeks after our initial data collection to update our information to give the users sufficient time to react and possibly make changes. Table 3.1 shows the update result. Table 3.3, 3.2, and 3.4 shows more information regarding the tweets' targets and senders.

Table 3.1. Dataset Global Information

Group	Size	Avg-length	Avg-likes	Avg-retweet
Dual Friendship	79909	14.73	1.45	0.12
No Friendship	59465	22.45	2.55	0.23
Source Follow Target	35786	17.59	1.63	0.12
Target Follow Source	2786	19.56	2.90	0.16

Table 3.2. Targets' Information

Group	Avg. # Followers	$> 10^3$	$> 10^4$	$> 10^5$	$> 10^6$
Dual Friendship	37222.96	30846	5488	1045	138
No Friendship	961484.50	32027	19654	10613	4188
Source Follow Target	2168448.27	23170	8454	13904	6209
Target Follow Source	6262.01	562	87	23	3

Table 3.3. Senders' Information

Group	Avg. # Followers	$> 10^3$	$> 10^4$	$> 10^5$	$> 10^6$
Dual Friendship	3324.47	25156	1810	95	4
No Friendship	1512.69	10879	994	76	10
Source Follow Target	1194.38	4576	310	16	1
Target Follow Source	13355.50	1467	427	63	5

Table 3.4. Number of Unique Participants in Each Category

Group	# sender/receiver	# unique senders	# unique receivers
Dual Friendship	79909	74504	73396
No Friendship	59465	55931	44436
Source Follow Target	35786	34093	21358
Target Follow Source	21358	2714	2749
Total	177946	162545	134238

CHAPTER 4
METHODS AND FRAMEWORK

4.1 Feature Engineering

To interpret the structure of our model we need to optimally select features from the tweets and evaluate how much they will help for a more accurate hostility detection. We considered the features below:

- The length of the tweets:

  Based on the real-life experience, we note that actual hostile messages should have a higher number of words on average. Figure 4.1 is illustrates the average length of the tweets for each relationship category. We suggested that it is more possible to say more words and ramble around if the user is angry or he/she wants to prove his/her point of view. Subsequently, both of these attitudes mostly result in hostile conversations.

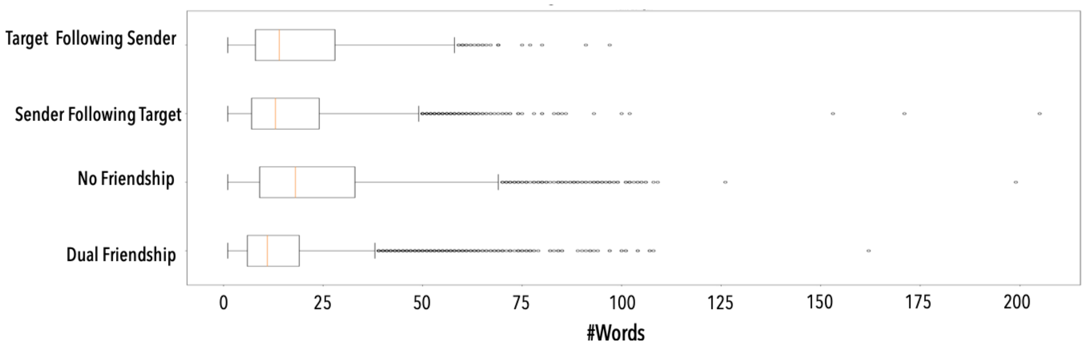


Figure 4.1. Box-plot of tweets' length in each category (Log Scale)

- The relationship category:

  Many hostile words can be used as an act of surprise or exaggeration, especially when it comes from a closer relationship. On the other hand, using some of these words such as b-word and n-word in some circumstances acts as a positive word and show the closeness between the sender and the target.

After topic modeling on different categories of the relationships, we realized that some hostile words are more likely to be used in both hostile and non-hostile context. These words have been assorted with both highly positive groups of words such as "LOL" and "love" and negative ones like "motherf****". Differently, words such as "racist" they have only been used in hate-based or hostile conversations.

While analyzing the features above, we encountered a couple more interesting features which do not applied in the final evaluation.

- The ratio of deleted tweets, after two week period:

  It is highly possible that the tweets in a hostile discussion are deleted because they were somehow hostile, antisocial, or hate-based. Therefore, we can address the fact that probably the tweets in "no friendship" group (with a higher rate of deleted tweets), are more likely to be hostile as well. Table 4.1 shows the ratio of deleted tweet regarding each relationship category.

- The average number of followers of the target:

  The correlation may not imply causation. Although the "dual friendship" category has the lowest ratio of hostility in our labeled data, it has one of the lowest targets' average follower numbers. This has a simple explanation. When we talk to a person which we are in the small circle of their friends, the words may lose their actual meaning and especially their negativity power.

  Figure 4.2 and 4.3 show the average number of follower for target and sender respectively for all relationship categories.

As you can observe, when the target is following the sender, usually sender has a larger number of followers in average which means that the sender is probably a celebrity. Imagine the situation where a celebrity replies back to a user with profane words as he/she might be angry about something or re-acting to a prior conversation. That may explain why they have the lowest size in our data. However, in the "Sender follow Target" the target is mostly the famous one. For instance, the account of US president. Every time that he tweets something, there are many hostile and aggressive replies back to him. Obviously, President is not following the senders.

Table 4.1. Global Group's Information

Friendship Category	Size	Deleted Tweets	
		Quantity	Ratio
Dual Friendship	79909	7656	9.58%
No Friendship	59465	8917	14.99%
Source Follows Target	35786	4089	11.43%
Target Follows Source	2786	357	12.81%

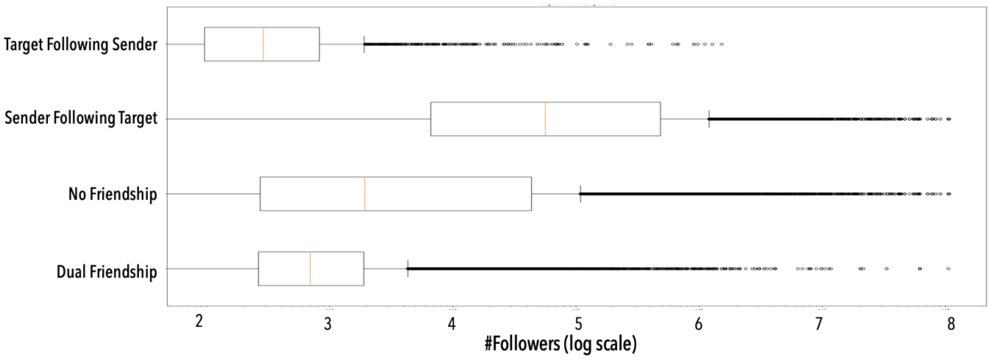


Figure 4.2. Box-plot of number of targets' followers in each category (Log Scale)

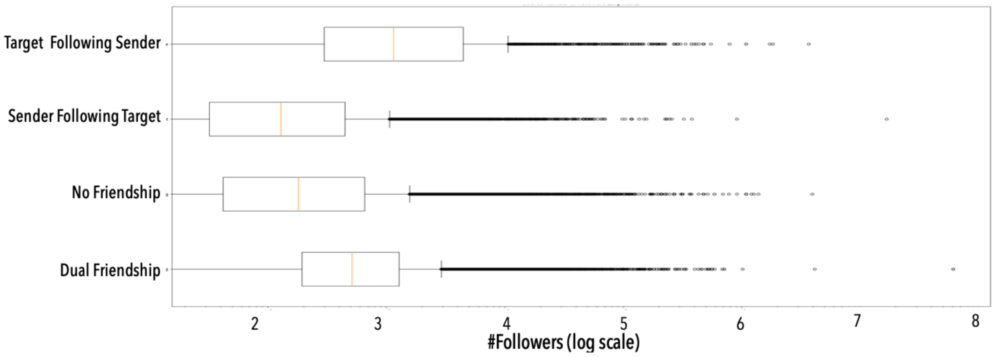


Figure 4.3. Box-plot of number of senders' followers in each category (Log Scale)

4.2 Experimental Methods

The features extracted from data processing were used to construct a model for detecting hostility. We tested several machine learning techniques to select the best classifier. We employed Logistic Regression (LR), Long Short Term Memory (LSTM), and Bi-directional (BD-LSTM). These techniques are discussed as follows.

4.2.1 Logistic Regression. Logistic regression is used when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression describes data and explains the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

4.2.2 Long Short Term Memory(LSTM). We implemented a Neural Language Model using a specific type of Recursive Neural Networks(RNN) known as Long Short Term Memory (LSTM).

A neural network is the closest replication of humans neurons. The neural network is made of multi-hidden layers, at least one layer, which the output of each layer is the next layer's input. As the data pass through each layer the network discovers more information about the data.

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture. Unlike standard feedforward neural networks, LSTM has feedback connections.It does not only process single data points (such as images), but also entire sequences of data (such as speech or video).

4.2.3 Bidirectional LSTM. Bidirectional Recurrent Neural Networks connects two hidden layers of opposite directions to the same output. With this form of RNN, the output layer can get information from past and future states simultaneously.

Our proposal is adding relationship to the classifier for better hostility detection. we have a BD-LSTM and a LR classifier. Both are using features bellow:

- Tweet Context: Exact text of tweet with stopwords removed and emojis replaced
- Length of Tweet: Size of the tweet in words
- Relationship Type: The relationship between the sender and the target immediately after the tweet has been posted.

Overall the LSTM models are more sophisticated that LR models and they are more capable of learning the text meaning. However, when it comes to a smaller amount of data, then the lightweight methods such as LR works the same. Therefore, we decided to consider the LR model in our evaluation step as well.

CHAPTER 5
EVALUATION

5.1 Comparison Metrics

Our data set contains an unbalanced distribution of classes; therefore, selecting the evaluation metric is considerably important. We chose Area Under the Curve (AUC) as our main comparison metric.

AUC provides an aggregate measure of performance across all possible classification thresholds. Since it shows high robustness for evaluation classifier, we can interpret AUC as the probability that the model ranks a random positive instance more highly than a random negative one. AUC has three main benefits:

- AUC provides robustness where the imbalanced data classes exist.
- AUC is scale-invariant. Instead of ranking predictions based on their absolute values, it measures how well they are ranked.
- AUC is classification-threshold-invariant. It measures the quality of the model's predictions regardless of what classification threshold is chosen.

The key advantage of AUC is that it is more robust than accuracy precision, recall, and f1-measure in class imbalanced situations. Given a 79% imbalance, the accuracy of the default classifier that consistently issues ”positive” will be 79%, whereas a considerably interesting classifier that actually deals with the issue is likely to obtain a worse score.

The ROC curve (receiver operating characteristic curve) denotes the rate of the true positive versus false positives at different threshold settings. The area provides a signal of the discriminatory rate of the classifier at various operating points.

Therefore, AUC gives a robust classifier performance measurement under imbalance class distribution data. A high AUC indicates an improved classification for both class instances regardless of the imbalance. In addition to ROC and AUC, we report the F1, precision, and recall as reference measures.

5.2 Hostility Distribution

Table 5.1 provides detailed statistics of the tweets coming from different groups. Tweets are cleaned and labeled from the annotation step. They include at least a profane word and contain a distinct sender and target.

Table 5.1. Hostility ratio

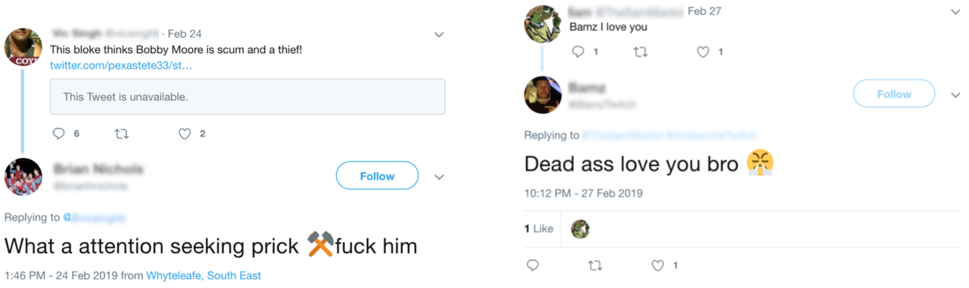
Relationship category	# Tweets	Non-Hostile Ratio	Hostile Ratio
Dual Friendship	2,176	70.54%	29.46%
No Friendship	1,516	25.07%	74.93%
Sender Fallows Target	1,610	38.70%	61.30%
Target Fallows Sender	1,469	41.59%	58.41%

As we expected, while the tweets in the dual-friendship group have toxic content, but they are not considered as hostile toward each other. More than 70% of the tweets are considered as normal, the non-hostile conversation. On the contrary, when there is no connection between the sender and the target, the hostility probability arises. When users do not know/follow each other, they behave more aggressively towards the other person during a heated conversation. This is mostly because they will not handle the real-life consequences of harassing someone that they do not know. Almost 75% of tweets with seed words are indicating a hostile discussion. We did not see a meaningful difference between the other two categories in terms of hostile/non-hostile ratio. The reason is the "Target follow Sender" category is includes rare groups of tweets with a wide range of users which make it much harder to find a pattern for.

Table 5.2. Directed vs non-directed hostility

Relationship Category	# Hostile Tweets	Directed Ratio	Non-Directed Ratio
Dual Friendship	641	42.12%	57.88%
No Friendship	1,136	63.82%	36.18%
Sender Fallow Target	987	60.69%	39.31%
Target Follow Sender	858	59.09%	40.91%

Table 5.2 expands the hostile column of the previous table to represent the hostility direction of the tweets. Sometimes a hostile conversation is happening but the harassment or negativity is not towards the participants. The "Dual friendship" group exposes more non-direct hostility comparing to the rest. Most often friends (aka users that are following each other) are complaining about a third party or they are using the profane words to show affection or support in a healthy manner. Figure 5.1 below displays an example for each case.



[redacted] · Feb 24
This bloke thinks Bobby Moore is scum and a thief!
twitter.com/pexastete33/st...

This Tweet is unavailable.

[redacted]

Replying to @[redacted]

What a attention seeking prick ⚒fuck him

1:46 PM - 24 Feb 2019 from Whyteleafe, South East

[redacted] Feb 27
Bamz I love you

[redacted]

Replying to [redacted]

Dead ass love you bro 😤

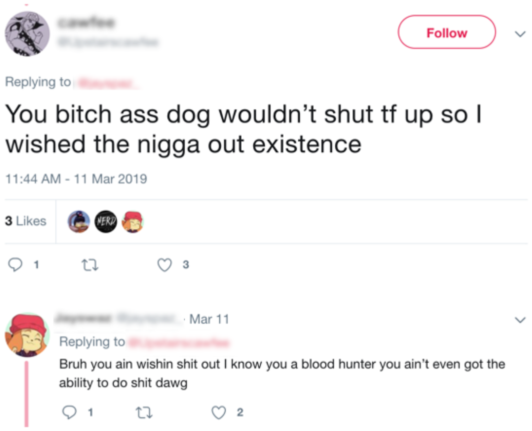
10:12 PM - 27 Feb 2019

(a) Attacking a third party

(b) Showing affection

Figure 5.1. Example of using profane words in a non-hostile manner in dual friendship category

In the "No Friendship" group, hostility is more directed. Less than 40 % of tweets are non-hostile. Users in this group blatantly harass each other and either want to cool their temper or anger the target. Figure 5.2 shows an example of direct hostility between two users who do not follow each other.



[redacted]

Replying to [redacted]

You bitch ass dog wouldn't shut tf up so I wished the nigga out existence

11:44 AM - 11 Mar 2019

3 Likes

1 3

[redacted] · Mar 11

Replying to [redacted]

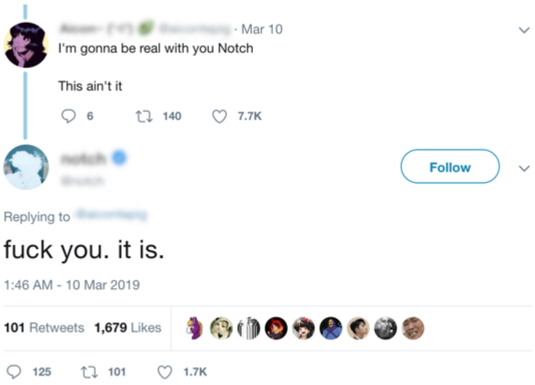
Bruh you ain wishin shit out I know you a blood hunter you ain't even got the ability to do shit dawg

1 2

Figure 5.2. Example of using profane words in a hostile manner in No Friendship category

When there is a one-way friendship between users, users feel more comfortable in starting or entering a directed hostile conversation. More than half of the tweets are direct hostile in both groups.

Below is one example where a celebrity (user with more than 1 M followers) is sending a direct hostile tweet towards another user.



[redacted] · Mar 10

I'm gonna be real with you Notch

This ain't it

6 140 7.7K

[redacted]

Replying to [redacted]

fuck you. it is.

1:46 AM - 10 Mar 2019

101 Retweets 1,679 Likes

125 101 1.7K

Figure 5.3. Example of a celebrity using profane words in manner in "Target follow Sender" category

5.3 Base Model

To show the benefits of applying the relationship categories, we compared our proposed method against a base model for each classifier.

- Classifier with LSTM: We implemented a neural network model using an existing sequence learning bidirectional LSTM with 2 layers of size 256 cells. We randomly split the dataset into training (70%), validation (15%), and test (15%) and trained the model on the tweet's text and its length. We train the model for 20 epochs, using Adam estimation for parameter update and drop out with regularization.

  The same training procedure with the same dataset has been applied for the base model. The only difference is that our base model does not receive the relationship information.

- Classifier with LR: We implemented a three different Logistic regression model:

  - Logistic Regression using only the tweet content
  - Logistic Regression on the tweets' content for each relationship category individually.
  - Logistic Regression using the length and relationship category features

  We randomly split the dataset into training/validation (85%) and test (15%). We then implemented 5-fold cross validation on our train/validation dataset.

5.4 Hostility Detection

5.4.1 Using LSTM Model. Figure 5.4 shows the accuracy of our proposed model versus the base model. Our model achieves high accuracy only after 3 epochs while the base model maximum accuracy never reaches 0.74.

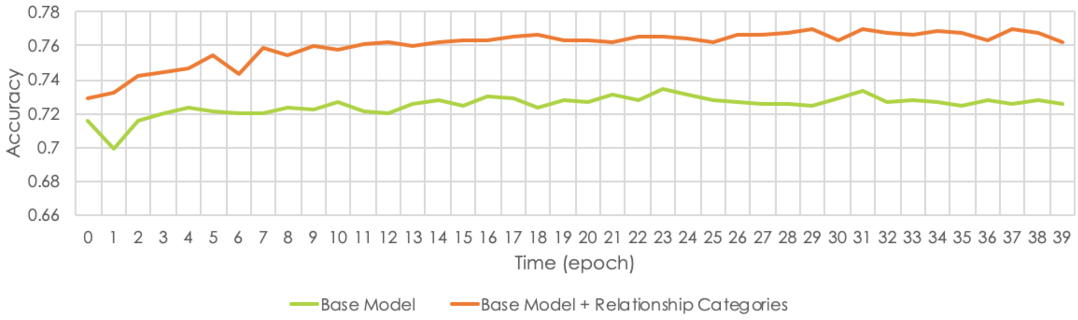


Figure 5.4. Validation accuracy for the LSTM models

Although accuracy drops multiple times, but it still grows eventually as the number of epochs increases. The figure above shows the addition of relationship groups helps to demonstrate the hostility problem more accurately.

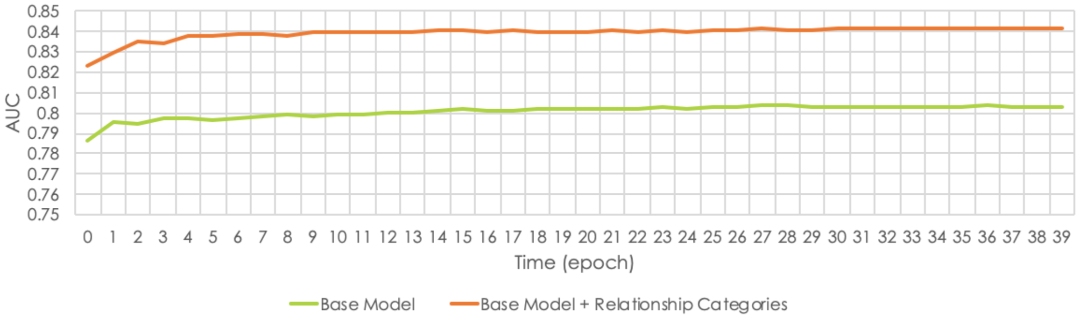


Figure 5.5. Validation AUC for the LSTM models

Figure 5.5 displays the AUC for both our model and the base model. A higher AUC represents more improved classification accuracy. As expected, it is easier to discover hostility when acknowledging the relationship between the sender and the target of the tweet. Table 5.3 reports the statistical results for the models. Defining the relationship groups clearly achieves higher F1, precision, and recall respectively.

The figures and the table above suggest that there are sufficient reasons to consider the relationship groups as a part of the classifier. Hostile situations not only

Table 5.3. LSTM Models Statistical Results

Relationship category	Precision	Recall	F1_score
Base Model	0.7189	0.7121	0.7103
Our Model	0.7588	0.7532	0.7522

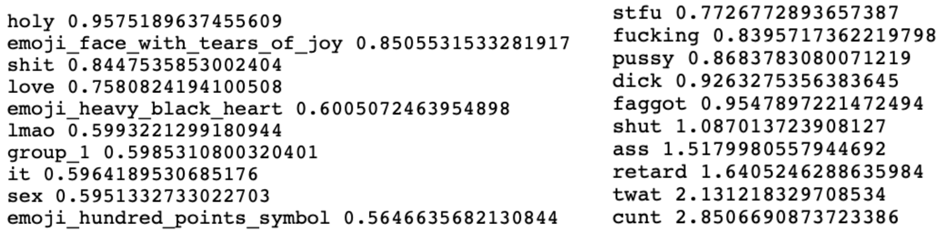
rely on the exact words but also their actual meaning and how they are used. Adding extra information about the relationship between the sender and the target resulted in more accurate detection.

5.4.2 Using LR Model. We implemented a Logistic regression model using our TF-IDF method for tweet content, in combination with tweet length and relationship category features.

Figure 5.6 and 5.7 indicates the top 10 coefficient words for our proposed model and baseline model respectively.

Moreover, figure 5.8, 5.9, 5.11, and 5.10 represent the LR base model for individual relationship categories.

As you can see, all the methods were able to detect the positive/negative words and separate them clearly. However, the LR with the relationship category model was also able to distinguish between the hostile words that are only used as negative phrases and the one that can also consider as a positive phrase. Overall LR+Relationship model improves the f1-measure by 4%. As a conclusion, table 5.4 represents the numerical statistics regarding different base models and the proposed LR model. The last row indicates how much improvement we can gain by considering the relationship categories in the LR model.



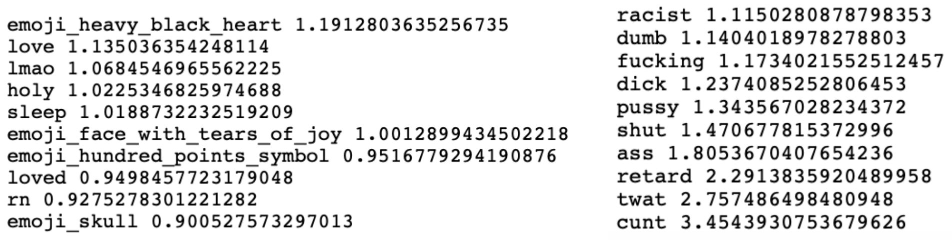
```
holy 0.9575189637455609
emoji_face_with_tears_of_joy 0.8505531533281917
shit 0.8447535853002404
love 0.7580824194100508
emoji_heavy_black_heart 0.6005072463954898
lmao 0.5993221299180944
group_1 0.5985310800320401
it 0.5964189530685176
sex 0.5951332733022703
emoji_hundred_points_symbol 0.5646635682130844
```

```
stfu 0.7726772893657387
fucking 0.8395717362219798
pussy 0.8683783080071219
dick 0.9263275356383645
faggot 0.9547897221472494
shut 1.087013723908127
ass 1.5179980557944692
retard 1.6405246288635984
twat 2.131218329708534
cunt 2.8506690873723386
```

(a) Non-Hostile

(b) Hostile

Figure 5.6. Top 10 coefficient words using LR on tweet's content + length + relationship category



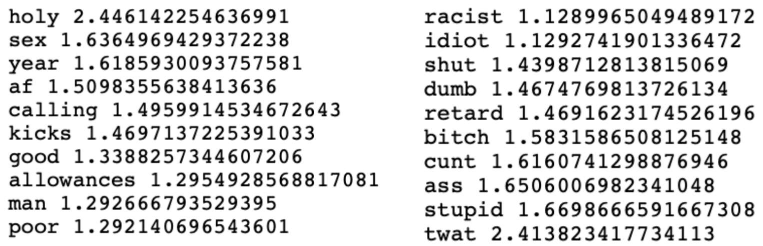
```
emoji_heavy_black_heart 1.1912803635256735
love 1.135036354248114
lmao 1.0684546965562225
holy 1.0225346825974688
sleep 1.0188732232519209
emoji_face_with_tears_of_joy 1.0012899434502218
emoji_hundred_points_symbol 0.9516779294190876
loved 0.9498457723179048
rn 0.9275278301221282
emoji_skull 0.900527573297013
```

```
racist 1.1150280878798353
dumb 1.1404018978278803
fucking 1.1734021552512457
dick 1.2374085252806453
pussy 1.343567028234372
shut 1.470677815372996
ass 1.8053670407654236
retard 2.2913835920489958
twat 2.757486498480948
cunt 3.4543930753679626
```

(a) Non-Hostile

(b) Hostile

Figure 5.7. Top 10 coefficient words using LR on tweet's content



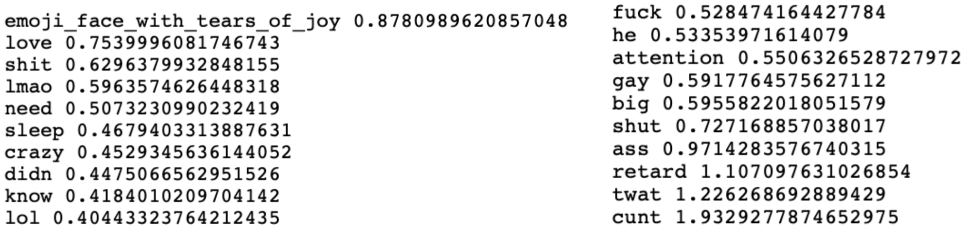
```
holy 2.446142254636991
sex 1.6364969429372238
year 1.6185930093757581
af 1.5098355638413636
calling 1.4959914534672643
kicks 1.4697137225391033
good 1.3388257344607206
allowances 1.2954928568817081
man 1.292666793529395
poor 1.292140696543601
```

```
racist 1.1289965049489172
idiot 1.1292741901336472
shut 1.4398712813815069
dumb 1.4674769813726134
retard 1.4691623174526196
bitch 1.5831586508125148
cunt 1.6160741298876946
ass 1.6506006982341048
stupid 1.6698666591667308
twat 2.413823417734113
```

(a) Non-Hostile

(b) Hostile

Figure 5.8. Top 10 coefficient words using LR with tweet's content on No Friendship category



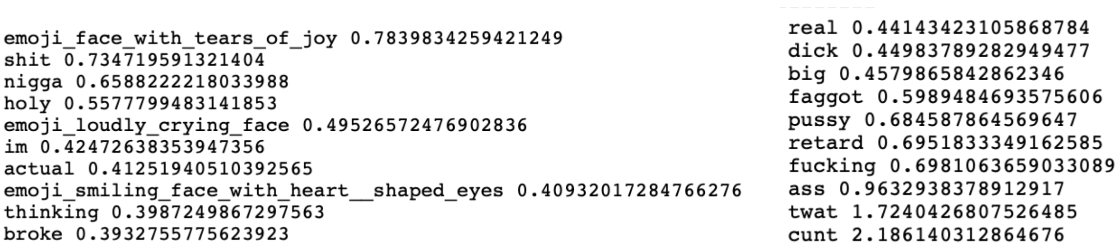
```
emoji_face_with_tears_of_joy 0.8780989620857048
love 0.7539996081746743
shit 0.6296379932848155
lmao 0.5963574626448318
need 0.5073230990232419
sleep 0.4679403313887631
crazy 0.4529345636144052
didn 0.4475066562951526
know 0.4184010209704142
lol 0.40443323764212435
```

(a) Non-Hostile

```
fuck 0.528474164427784
he 0.53353971614079
attention 0.5506326528727972
gay 0.5917764575627112
big 0.5955822018051579
shut 0.727168857038017
ass 0.9714283576740315
retard 1.107097631026854
twat 1.226268692889429
cunt 1.9329277874652975
```

(b) Hostile

Figure 5.9. Top 10 coefficient words using LR with tweet's content on Dual Friendship category



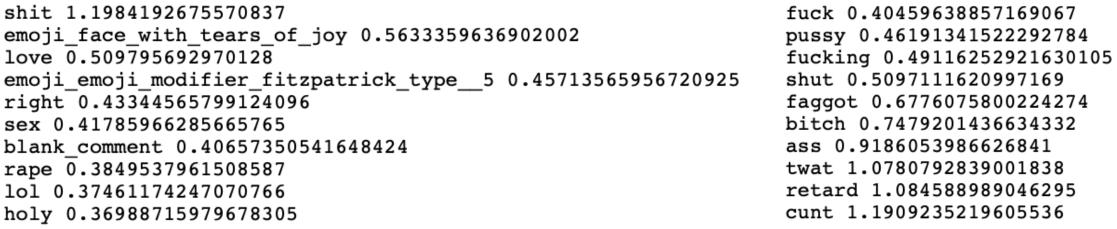
```
emoji_face_with_tears_of_joy 0.7839834259421249
shit 0.734719591321404
nigga 0.6588222218033988
holy 0.5577799483141853
emoji_loudly_crying_face 0.49526572476902836
im 0.42472638353947356
actual 0.41251940510392565
emoji_smiling_face_with_heart__shaped_eyes 0.40932017284766276
thinking 0.3987249867297563
broke 0.3932755775623923
```

(a) Non-Hostile

```
real 0.44143423105868784
dick 0.44983789282949477
big 0.4579865842862346
faggot 0.5989484693575606
pussy 0.684587864569647
retard 0.6951833349162585
fucking 0.6981063659033089
ass 0.9632938378912917
twat 1.7240426807526485
cunt 2.186140312864676
```

(b) Hostile

Figure 5.10. Top 10 coefficient words using LR with tweet's content on Sender follow Target category



```
shit 1.1984192675570837
emoji_face_with_tears_of_joy 0.5633359636902002
love 0.509795692970128
emoji_emoji_modifier_fitzpatrick_type__5 0.45713565956720925
right 0.43344565799124096
sex 0.41785966285665765
blank_comment 0.40657350541648424
rape 0.3849537961508587
lol 0.37461174247070766
holy 0.36988715979678305
```

(a) Non-Hostile

```
fuck 0.40459638857169067
pussy 0.46191341522292784
fucking 0.49116252921630105
shut 0.5097111620997169
faggot 0.6776075800224274
bitch 0.7479201436634332
ass 0.918605398626841
twat 1.0780792839001838
retard 1.084588989046295
cunt 1.1909235219605536
```

(b) Hostile

Figure 5.11. Top 10 coefficient words using LR with tweet's content on Target follow Sender category

Table 5.4. Statistical Results of Logistic Regression Models

Relationship category	Precision	Recall	F1_score
All categories + (text) feature	0.71	0.71	0.71
All categories + (length + relationship) features	0.64	0.64	0.64
Dual Friendship category + (text) feature	0.72	0.72	0.72
No Friendship category + (text) feature	0.64	0.64	0.64
Sender Follow Target category + (text) feature	0.73	0.73	0.73
Target Follow Sender category + (text) feature	0.69	0.68	0.68
All categories + (text + length + relationship) features	0.75	0.75	0.75

CHAPTER 6
CONCLUSION

Toxic content in social media is increasing every day and more users are using aggressive, threatening, or bullying language in the heated conversations. We focused on the hostility in social media and proposed a method to detect hostile conversations on Twitter more accurately. While prior research addressed the problem by only analyzing the exact words, we tried to find the actual meaning of the word using the relationship between the participating users. We find that

- Friends (users who follow each other) use profane words mostly a non-hostile behavior to show support or affection. Even if there is hostility within this group, it is frequently towards a third party.
- Users who do not follow each other tend to engage more in hostile conversations and their aggressiveness is usually directed toward the target.
- If the target is not following the sender, and there is a hostile conversation happening, usually sender has a huge follower base and might be a celebrity replying to a previous insult.
- When there is no friendship between users in a hostile discussion, the average length of tweets increases as they show more rage in the ongoing conversation to be able to prove their point of view.

We used an RNN with bi-directional LSTM and a LR classification model with features including the tweet content, length of the tweet, and the relationship between the sender and the target. Our BD-LSTM model improved the AUC of hostility detection by 4% while providing the F1 of 5% as well. Similarly, the F1-score got increased by 4% using a classifier with LR. We showed that employing the relationship categories as a feature for the classification model results in higher detection accuracy.

There are several approaches available for future work. First, evaluating our proposed model on a larger dataset and also teach annotators more deeply. When using machine learning techniques, the size of data for train and test plays a vital role in the overall performance. Larger dataset provides better and more accurate results. For our research, however the neural network is capable of detecting non regular patterns as long as it receives a huge amount of data, but it requires significantly large dataset. Our main challenge was that in our first round of annotation, we chose to work with 63 different annotators which means 63 different definitions of hostility. Therefore, the initial result was not acceptable as we wanted to be due to sparsity of labels. Although, in the second round, we decreased the size of annotators to only three and therefore result was much more uniform.

Second, inferring more user attributes (e.g. age, gender, ethnicity, geo-location, etc) may provide additional insight for detection. For instance, the N-word is not considered profane in some countries. Third, applying more complex linguistic techniques helps with better sampling (less manual annotation, towards more automation). Using a profane word in a tweet expressing happiness or excitement should not be considered as a possible candidate for a hostile tweet. Finally, use the last 200 tweets of both sender and the target of the tweets to create a conversation chain and use those additional information's as insight to both detect and also predict hostility in twitter.

APPENDIX A

LIST OF SEED WORDS

The following seed words are the profane words that we used to collect tweets from Twitter. They are gathered from HateBase and other online resources.

[’twat’, ’nigger’, ’pussy’, ’coon’, ’cunt’, ’nigga’, ’fudgepacker’, ’bitch’, ’slut’, ’anal’, ’anus’, ’ass’, ’bastard’, ’bitch’, ’boob’, ’cock’, ’cum’, ’cunt’, ’dick’, ’dildo’, ’dyke’, ’fag’, ’faggot’, ’fuck’, ’fuk’, ’handjob’, ’homo’, ’jizz’, ’kike’, ’kunt’, ’muff’, ’nigger’, ’penis’, ’piss’, ’poop’, ’pussy’, ’queer’, ’rape’, ’semen’, ’sex’, ’shit’, ’slut’, ’titties’, ’twat’, ’vagina’, ’vulva’, ’wank’,’cunt’, ’raghead’, ’cripple’, ’retard’, ’knacker’, ’gypo’, ’yellow bone’, ’niggur’, ’shyster’, ’sand nigger’, ’Uncle Tom’, ’muzzie’, ’faggot’, ’peckerwood’, ’greaseball’, ’twat’, ’Jihadi’, ’porch monkey’]

APPENDIX B

HOSTILITY DETECTION MANUAL

The word hostile comes from the Latin word hostis, an enemy. Hostility is seen as a form of emotionally charged aggressive behavior. Some synonyms are inimical, antagonistic, unfavorable, and unfriendly. Therefore, a hostile comment is defined as a message which contains harassing, threatening, or offensive/harsh language directed toward a specific individual or group.

If a message has the following characteristics, then it has a higher chance of being hostile.

- If the message is mostly in Capital Letters (CAPS)
- If the message has excessive exclamations or letters (It tends to show more emotion towards targeting someone)

  Ex: "@***@****@***_ DGHahahhahhahahhahhahhhhahahhhahhahahahahahhahahah Fuck you."

  "@*** WHO THE FUCK WON SOMEONE ANSWER"

On the contrary, the message should not be considered as hostile if:

- the message targets the original poster (Self-degradation)

  Ex: "FUCK ME im a terrible person this is from @***"

- a hostile word is used as a placeholder to refer to some object

  Ex: "I dont have to pay rent so I can buy dumb shit"

- a hostile word is used as an element of surprise or exaggeration

  Ex: "@*** How the fuck is this possible? (am drunk but this shit is impossible)"

- some hostile words just occur by itself with no reference to targeting anything

  Ex: “@*** FUCK”

B.1 How to use the table?

There are seven columns in the table, which five of them already has data inside of them and the rest you need to use to label the data.

- id: Contain the unique id of each tweet which can then be used in our next steps.
- text: The actual full text of each tweet, which you have to use to decide if the tweet is hostile or not.
- target: The tweets we are using are all non-starter tweets. That means they are posted in reply to another tweet, and the target refers to the writer of the previous tweet which this tweet is replying to.
- sender: The writer of the current tweet.
- link: The link to the actual tweet.

B.2 How to label the data?

As you can see there are two empty columns which we need you to help us with:

- Hostile: For this column, you only need to consider the text column and decide if the text is hostile or not. (1 = hostile, 0=not-hostile)
- If hostile is it directed?: For this column, you need to check to use the sender, target, and the text column and decide if the comment is hostile and the sender

is referring to the target at the same time. (1= hostile and direct, 0= not hostile or not direct)

If the text is not hostile in the first place, then use 0 for both columns

For instance, the following tweet is hostile but not a direct hostile. Therefore, the hostile column should be 1 and the If hostile is it directed? column should be 0.

Ex: "@*** Lmao my old hurt ass fucked it up wife.. tried to beat him to the punch of fucking me over because Convinced myself he was or would fuck me over when he wasnt doing either one Im glad to say today I HAVE GROWN "

APPENDIX C

SAMPLE ANNOTATED TWEETS

C.1 Not Hostile

- “@*** @*** drunk i really birthed you before al these niggas no cap! First nigga to make yo nose bleed cause you my lil bro”
- “@*** @*** nigga they just won like 14 of their last 16 you serious?”
- “@*** Every guy pays for pussy rich or poor”

C.2 Hostile and Direct

- “@*** Nigga trash as hell”
- “@*** You nasty nigga.”
- “@*** Fuck off and die”
- “@*** what the fuck did you just say to me, you little prick”
- “@*** @HatfieldAnne Shit, Ive got those...is that bad?”

C.3 Hostile and non-Direct

- “@*** Wellll I tried thanks to our wager but within 30 seconds of turning it on some cunt made comment about the border wall. Shortly after that a rich millionaire cunt commenting about how bad women and people of color have it. Screen reader support enabled. ”
- “@*** Because she is a total twat”
- “@*** Hang the cunt!!!! Crossed eyed bitch”
- “@*** @*** this bitch look like she stink”

BIBLIOGRAPHY

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications. *Comput. Hum. Behav.*, 63(C):433–443, October 2016.

[2] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):32:1–32:25, November 2018.

[3] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1217–1230, New York, NY, USA, 2017. ACM.

[4] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 307–318, New York, NY, USA, 2013. ACM.

[5] HateBase contributors. Hatebase. `https://hatebase.org/`, 2019. [Online; accessed Feb-2019].

[6] Jacob Emerick. List of dirty naughty obscene and otherwise bad words. `https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words`, 2018. [Online; accessed Feb-2019].

[7] David Jurgens. Word sense induction by community detection. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, TextGraphs-6, pages 24–28, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] Srijan Kumar, Justin Cheng, and Jure Leskovec. Antisocial behavior on the web: Characterization and detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 947–950, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[9] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. *CoRR*, abs/1804.06759, 2018.

[10] Luis von Ahn. List of bad words. `https://www.cs.cmu.edu/~biglou/resources/bad-words.txt`. [Online; accessed Feb-2019].

[11] Robert James Gabriel. Google profanity words. `"https://github.com/RobertJGabriel/Google-profanity-words`, 2017. [Online; accessed Feb-2019].

[12] Marco Del Tredici and Raquel Fernández. Semantic variation in online communities of practice. *CoRR*, abs/1806.05847, 2018.

[13] Amy X. Zhang, Bryan Culbertson, and Praveen Paritosh. *Characterizing Online Discussion Using Coarse Discourse Sequences.*

[14] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. Conversations gone awry: Detecting early signs of conversational failure. *CoRR*, abs/1805.05345, 2018.

[15] Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J. Taylor. Characterizing online public discussions through patterns of participant interactions. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):198:1–198:27, November 2018.