The Impact of Toxic Replies on Twitter Conversations

by

Nazanin Salehabadi

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science at The University of Texas at Arlington

December 2019

Arlington, Texas

Supervising Committee:

Shirin Nilizadeh, Supervising Professor

Dajiang Zhu

Changkai Li

ABSTRACT

The Impact of Toxic Replies on Twitter Conversations

Nazanin Salehabadi, M.S.

The University of Texas at Arlington, 2019

Supervising Professor: Shirin Nilizadeh

Social media has become an empowering agent for individual voices and freedom of expression. Yet, it can also serve as a breeding ground for hate speech. According to a Pew Research Center study, 41% of Americans have been personally subjected to harassing behavior online, 66% have witnessed these behaviors directed at others, and 18% have been subjected to particularly severe forms of harassment online, such as physical threats, harassment over a sustained period, sexual harassment, or stalking. Recently, many research studies have tried to understand online hate speech and its implications, focusing on detecting and characterizing hate speech. One limitation of these works is that they analyze a collection of individual messages without considering the larger conversational context. Our project has two objectives: First, we characterize the impact of hate speech on Twitter conversations, in terms of conversation length and sentiment, as well as user engagement; Second, we demonstrate the feasibility of automatically generating hate replies to some tweets, using retrieval models. For the first objective, we: (1)

extracted toxic tweets and their corresponding conversations; (2) defined a toxicity trend score for conversations; and (3) studied the impact of toxic replies on twitter conversations using statistical methods. For the second objective, we: (1) created a knowledge database for toxic tweets and replies; (2) implemented a retrieval model that uses Doc2vec embedding, which identifies N top tweet-reply matches for a specific tweet; (3) proposed a ranking algorithm based on Word2vec that identifies the best hate reply for the tweet; (4) evaluated our approach by implementing some alternative approaches and running several studies on Amazon Mechanical Turk.

ACKNOWLEDGEMENTS

I would like to thank Dr. Shirin Nilizadeh, my supervising professor, for her guidance and support in this research project. Dr. Nilizadeh provided tremendous guidance and direction in my thesis work, from getting started with reading relevant papers, to the final stages of evaluation and writing my thesis. I also would like to thank the rest of the committee members, everyone who participated in the experiments for the thesis.

My heartfelt appreciation goes to my parents, Mohammad Salehabadi and Fereshteh Vosoughi, who have always supported me in every step of my life. Their constant devotion and encouragement were among the most significant factors that helped me complete this challenging journey. I would like to thank my sister and brother, Nooshin and Omidreza, for being always supportive and helpful.

Last but not least, I would like to express my love and deepest appreciation to my husband, Mohamadreza, for her endless support, encouragement, understanding and patience.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

API – Application Programming Interface

HS – Hate Speech

HS- – Toxic Comment After Hate Speech

HS+ – Non-Toxic Comment After Hate Speech

NLP – Natural Language Processing

IR – Retrieval Model

CHAPTER 1

INTRODUCTION

There is no explicit definition for hate speech, but the majority holds the opinion that it is speech that objects individuals or groups in a way that is harmful to them. Hate speech is a broad term and is based on different attributes such as race, ethnicity, gender, and sexual orientation, or threats of violence against others. Hate speech towards individuals and minority or majority groups can cause violence or social disorder. In the United States hate speech is legally protected under the free speech stipulations, but in many countries, such as United Kingdom, Canada and France, hate speech is illegal and offenders face large fines and imprisonment (Davidson et al., 2017).

Hate speech directed to individuals is categorized as Directed and toward groups is categorized as Generalized. Directed hate speech is very personal but in contrast Generalized hate speech specially in religious and ethnic area is more general. Both Directed and Generalized hate speech are informal speech, but informality and anger can be found more in Directed hate speech (ElSherief et al., 2018).

Hate speech in social media also can be used to express hatred, humiliation and insult against a targeted individual or group (Davidson et al., 2017). Social media is a collection of online communications channels, interaction, content sharing and collaboration. Examples of social media companies include Facebook, Twitter and Youtube (Ring,2013). Social media improves the communication, access to knowledge, spreading information, and social interaction. Social media is a dominant part of daily lives, easily facilitating and improving

communication and exchange of points of view (ElSherief et al., 2018). Hate speech content on social media can stay available for public for a long time and cause more damages to victims and empowers the offenders. Social medias consider their own rules to protect people from hate speech attack; however, websites like Twitter and Facebook are accused of not preventing hate speech, but they claim they have founded policies to stop the use of their platforms for attacks on people (Davidson et al., 2017). Environments like Twitter's conversations around specific topics may facilitate the quick and wide spreading of hateful messages.

The destructive effects of hate speech content on websites and their substantial influence and harms on victims and community has been widely recognized. Hate speech spreads hatred, violence, discrimination and conflict against a person or a group. Hate speech poses serious dangers for the cohesion of a democratic society, the protection of human rights and the rule of law. There are dangerous links between hate speech and violence. Hate speech can cause social disorders and hate crime and the danger of hate speech is more than hurt feelings. Pew Research Center revealed a study that shows 60% of Internet users are victims of offensive name calling, 25% witnessed someone physically threatened, and 24% witnessed someone being harassed for a long period of time. As a result, hate speech has become an important focus for research (ElSherief et al., 2018).

With the rapid growth of online social networks, people have become increasingly in exposure of experiencing abusive language and hate speech. When an individual decides to engage in an online discussion, they are exposing themselves to the risk of being harassed by hate speech commenters. Instances of offensive comments are quite common and have

negatively impacts the dynamics of the online community, the user experiences, and direction of discussions. Reading hate speech primes your brain for hateful actions. Current research shows that humans understand language by activating sensory, motor and emotional systems in the brain. According to this new simulation theory, just reading words on a screen activates areas of the brain in ways similar to the activity generated by literally being in the situation the language describes. Most online hate on conversations looks a lot like fear, and they are used to spread the fear. Fear, more than hate, feeds online bigotry and real-world violence (The Conversation, 2019).

Others have studied the impact of online toxicity in conversations on individuals from different perspectives. First, the harm done to its targets, from online spaces being experienced as hostile. The harm may not be visible, but definitely a large amount of cyberbullying is motivated by hate. For example, lesbian, gay, bi-sexual and transgender (LGBT) youth are almost twice as likely to report having been bullied online as those who are straight, young women in compare to young men are twice more likely to have been sexually harassed online. Young people who experience online hate are more in danger of experiencing anxiety and depression. Second, the risk that those who face it may be radicalized by it, becoming more sympathetic and possibly even active; Radicalization refers to the process by which people come to believe that violence against others and even oneself is justified in defense of their own group. Third, the effect that it has on the values and culture of the online spaces. The purpose is to make an online environment more hostile. It includes posting snarky jokes about

an unfolding news story, tragedy, or controversy; retweeting hoaxes and other misleading narratives ironically, to condemn them, make fun of the people involved, or otherwise assert superiority over those who take the narratives seriously (MediaSmarts, 2019). Three above mentioned items only study the toxic impact on individuals who are targeted or not targeted by a toxic comment in an online conversation.

In this project, we study Twitter conversations with toxic replies to understand how a toxic reply impacts the whole conversation and also to identify factors that might increase or decrease the toxicity of the whole conversation. For that, we leverage the literature from social science and phycology.

(Zhang et al., 2016) studied conversation dynamics in a debate dataset and founds that losers are more interested in imposing their ideas to gain control of discussion, however, winners are more active than losers in contesting their opponents' opinion. We are also interested in studying the flow of conversations and examine the activity level of toxic and non-toxic commenters. We define the toxicity flow of a conversation as toxicity trend, and group conversations into three classes of positive, negative and neutral. Then, we study the activity level of toxic commenters in these groups. Active toxic commenters contribute to the negativity of a conversation, while active non-toxic commenters contribute to the positivity of a conversation. We define a conversation neutral if toxic and non-toxic commenters compete and have the same level of activity. We also examine the impact of toxic replies on

the length of conversations, to see if negative Twitter conversations are shorter in compare to positive conversations.

(Zhang et al., 2018), studied linguistic cues that predict a conversation's future health and found that an initially healthy conversation can be affected by personal attacks. We also examine how the location (index) of a toxic reply affect the negativity and length of conversations. For example, if a toxic reply occurs closer to the head or tail of a conversation, how it affects user engagement in the conversations.

(Zhang et al., 2016), also showed that the audience feedback on debates is related to debate outcome. (Kwak et al., 2018) also studied toxic comment in online games, and found that players of the online games are surprisingly not active in generating toxic, and most of the toxics are generated by third parties the contribution of third parties. We also examine if only a limited number of users are posting toxic replies. We also study if some individuals stop their engagements in the conversation when observing toxic replies, and if their engagement can affect the total positivity or negativity trend of the conversation. We also investigate if toxicity can encourage additional toxicity.

Finally, we study the impact of having an immediate non-toxic reply after the first toxic reply on the conversation. The goal is to understand if having a positive or constructive comment after some toxic comment can help saving the conversation and encourages user

engagements. Therefore, we examine if having an immediate non-toxic or toxic after first toxic affects the trend of rest of conversation to be positive or negative.

In summary, to study the impact of toxic replies on Twitter conversations, we examine the following research questions:

1. How many of conversations are positive? negative? neutral?
2. Compare the length of positive vs. negative vs. neutral conversations.
3. Initial toxic comment occurs in which index of conversation (with considering index 0 as main tweet)?
4. Number of unique users per conversation and number of unique users participating in toxic comment per conversation.
5. Number of unique users participating in conversations after initial toxic.
6. Conversations started with a toxic main will be negative.
7. The position of hate comment affects the length of conversation.
8. What percent of conversations get non-toxic reply after initial toxic?
9. Having a positive reply on a toxic comment changes the discussion to be positive.

Findings. We found that toxic comments affect conversations negatively, where they can reduce user engagement in conversations and decrease the length of conversations. We also found most of conversations have neutral toxicity trend. We also found that having more user engagement can help positivity of conversations. We also found that most of the time, the first toxic comment is also the first reply to the main tweet, which might show that the author

of first tweet as well as his/her followers decide to not response to the toxic commend and the conversations dies. We also found that a limited number of unique users post toxic comments. After the first toxic reply, users either stop participating in the discussion, or if they do, they contribute to the positivity. Most of the time conversations started with a toxic tweet (i.e., the main tweet is toxic) will be negative. We also found that conversations that include a toxic reply immediately after the first toxic replies are more negative, whereas conversations that have an immediate positive comment after the first toxic comment are more positive.

Based on our findings, it is clear that even having one toxic reply on a tweet can negatively impact the conversation. These findings motivated us to examine the feasibility of automatically generating toxic replies for tweets. We proved that offenders can accelerate their destructive behavior by creating hate speech.

Natural Language Processing (NLP) Techniques can be used in generating new hate speech for an utterance. There are two types of models to create new sentence in the same way as human does, Information Retrieval models (IR) and Generative models. Information retrieval models are based on obtaining relevant reply for a given utterance. Retrieval models query a targeted utterance in a repository and return a relevant reply that best matches the given utterance. Generative models generate new replies, they create a new sentence word by word. Generative models are typically based on sequence to sequence (seq2seq) models. Generative models receive a sequence of words as input and then generates a sequence of words as output. Seq2seq models are based on two recurrent neural network (RNN) models,

one as encoder and the other as decoder. Encoder receives input as vectors (also known as Embedding). Decoder decodes the vectors to a sentence (Song et al., 2018).

Natural Language Processing (NLP) Techniques are used in Question-Answering, Chatbots and dialog systems. Generative models are state of the art for generating text, and generated text based on them are not natural and have lower readability score. Most of previous works on Question-Answering used retrieval models (Lee at al., 2018). We used Retrieval models to create a new toxic reply for an utterance.

Adversaries can abuse these tools to create toxic comments with the goal of negatively affecting conversations and social media. Social media and other forms of communication are being exploited as platforms for bigotry. Cases of genocide and crimes against humanity could be the next frontier of social media jurisprudence, drawing on precedents set in Nuremberg and Rwanda. The Nuremberg trials in post-Nazi Germany convicted the publisher of the newspaper Der Sturmer; the 1948 Genocide Convention subsequently included "direct and public incitement to commit genocide" as a crime. During the UN International Criminal Tribunal for Rwanda, two media executives were convicted on those grounds. As prosecutors look ahead to potential genocide and war crimes tribunals for cases such as Myanmar, social media users with mass followings could be found similarly criminally liable (Council on Foreign Relations, 2019).

For the two mentioned objectives, we considered Twitter as the source for extracting hate speech conversations as it's very popular channel and hate speech is so common in this channel. We extracted different classes of hate speeches from Twitter and considered all

tweets that fall under any one of the hate speech categories as toxic speech (Koratana et al., 2019). We also used Amazon Mechanical Turk to validate our result in second objective. Amazon Mechanical Turk is a web service that provides public with facility of completing tasks which cannot be done systematically and requires manual effort. We asked workers to score our created toxic reply based on relevancy to the given utterance.

The contribution of this research to the body of knowledge is (1) developing a framework for collecting conversations in twitter. (2) Analyzing conversations that include toxic comments for better understanding of the factors that impact the flow of discussions. (3) Developed an information retrieval model for generating toxic replies.

CHAPTER 2

RELATED WORK

Recently many related works have studied hate speech focusing on hate speech detection, classification and characterization. There has been no attempt on studying conversations that include hate speech, and no other work has proposed to generate hate toxic. A few studies have tried to understand the impact of hate speech on hate targets and flow of conversations. For example, (Kwak et al., 2018) shows that online games make players particularly vulnerable to the exhibition of, and negative effects from, cyberbullying and toxic behavior. (Zhang et al., 2018) also studies conversation dynamics in Oxford-style debate dataset, which is from the public debate series "Intelligence Squared Debates" and shows how the outcome of debate depends on aspects of conversational flow including number of discussion points, talking points and discussion feedback. (Zhang et al., 2018) studies linguistic cues that predict a conversation's future health. (Maity et al., 2018) studies the factors associated with incivility in Twitter, which caused users to leave Twitter.

Hate speech detection and classification. Most current works in the domain of hate speech have focused on detection algorithms, including binary classification, and multi-class classification that identify the type of toxic speech. For example, (Koratana.et.al 2019) Classifies comments to seven groups of clear, toxic, obscene, insult, identity hate, severe toxic, and threat. (Zhang et al. 2016) separated hate speech detection methods to two methods of classical methods: (1) feature engineering including Logistic Regression, Bayesian models,

SVM, and random forest (Davidson et al., 2017), and (2) deep learning methods (Koratana et al., 2019).

Hate speech detection and classification also have been handled by a variety of features, including lexical properties, such as n-gram features (Nobata et al., 2016), character n-gram features (Mehdad and Tetreault 2016), Character n-gram, demographic and geographic features (Waseem et al., 2016), average word embeddings, and paragraph embeddings (Nobata et al., 2016; Djuric et al., 2015), and linguistic, psychological, and affective features inferred using an open vocabulary approach (ElSherief et al., 2018). (ElSherief et al., 2018) deeps on online hate speech features and categorizes hate speech to Directed and Generalized groups. Directed hate speech is a speech which targets individuals. Generalized hate speech is a speech which targets a group and its members. Directed hate speech is very personal in contrast to Generalized hate speech which is general especially in religious and ethnic areas. Both Directed and Generalized categories are informal speech. Other work has used sentiment analysis (Dinakar et al., 2012; Sood, Churchill, and Antin 2012; Gitari et al., 2015).

Hate speech Characterization. Over the past few years, several approaches have been proposed to measure abusive behavior on social media platforms like Facebook, Twitter and etc. (Chen et al., 2012) used both textual and structural features to predict a user's intrinsic desire in producing toxic content in YouTube, while (Kayes et al., 2015) found that users tend to flag toxic content posted on Yahoo Answers in an exorbitance correct way. Also, some users considerably deviate from community norms, posting a large amount of toxic content. Through careful feature extraction, they also showed that it is possible to use machine learning

approaches to predict which users will be suspended. (Chatzako et al., 2017) studied the properties of bullies and aggressors and employ supervised machine learning to classify Twitter users, also studied the users of tweets with the #Gamergate hashtag. (ElSherief et al., 2018) Compared the characteristics of hate instigators and hate targets from multiple perspectives and showed both hate instigator and target users are more likely to get attention on Twitter, i.e., they obtain more followers, are retweeted and listed more.

Information Retrieval models for text generation. We propose an information retrieval model to generate relevant and meaningful toxic replies for some utterances. Retrieval models are based on obtaining relevant reply for a given utterance, they query a targeted utterance in a repository and return a relevant reply that best matches the given utterance. This process is done by a ranking function which ranks the candidates in a query. There are different methods to design the ranking function. The purpose of ranking function is to rank relevant sentences on top of non-relevant ones (Novgorodov et al., 2019). Question and answering systems are mostly based on traditional information retrieval models, which use TF-IDF rankings. (Chen et al., 2017; Wang et al., 2017b) proposed comprehensive knowledge dataset machine which is so efficient in answering to a given question. For this purpose, they used an open-domain system where a given question have to be searched from a huge knowledge database like Wikipedia, although these traditional retrieval models are so efficient, the answer candidates retrieved and ranked at the top by such systems often are not best answers to questions (Lee at al., 2018). In another work, (Zhang et al., 2018) created a dialogue system, where human and machine communicate as partners in a two by two conversation, leveraging both

generative (LSTM) and retrieval models (i.e., TF-IDF and Star Space separately with considering human profiles). They use a crowdsource database (utterances between crowd workers who were randomly paired) and show that retrieval models outperform generative models. (Mendes et al., 2011) proposed an approach to "answer selection" in question-answering system using semantic relations. (Mendes et al., 2011) explored semantic relations detected between the candidate answers to a given question, using the corpus of factoid questions, the factoid corpus dataset was gathered through available data from the Text Retrieval Conference (TREC), Corpus included both questions and answers, however, this approach proved to be more effective in factoid questions (Acosta et al., n.d.). (Medved et al., 2018) used two different approaches to find an answer for a given question: sentence embedding (Doc2Vec) (Mikolov et al., 2014) and word embedding (Word2vec) (Mikolov et al., 2013). Word2Vec and Doc2Vec are explained more in chapter five. By using Word2Vec, words with similar meanings will have similar vectors and as the result similar positions in vector space, Doc2Vec modifies the word2vec algorithm to unsupervised learning of continuous representations for larger blocks of text, such as sentences, paragraphs or entire documents, and as the result each sentence is represented by one vector in vector space. Despite the efficiency of the traditional retrieval systems, the candidate answer retrieved and ranked at the top by such systems often do not contain answers to questions. (Lee et al., 2018) re-ranked and filtered returned documents from TF-IDF by paragraph ranker using bidirectional long short-term memory. BILSTM encodes returned documents and given question by using two RNNS, and then they calculate the probability that a returned paragraph

contains the answer to the given question by calculating similarity. In this project, we propose an informational retrieval model that uses two algorithms for ranking the retrieved documents, i.e., Doc2Vec as first ranker and Word2Vec as second ranker. Most of the time, retrieval algorithms top ranked returned result is not the exact answer to the question. We tried to re-rank returned answers and come up with the most suitable answer. Our main focus is on re-ranking retrieved sentences to correctly answering the questions (Lee at al., 2018). In Chapter eight, we show that this technique is effective in providing more relevant replies to some specific tweets.

CHAPTER 3

COLLECTING TWITTER CONVERSATIONS

In this chapter, we describe the datasets used for our analysis, study and experimentation in first objective "Study the impact of toxic replies on Twitter conversations". We also describe the features and characteristics which are available in our datasets. The dataset is gathered from Twitter, and we used Twitter Standard API to extract the data. We considered Twitter as the source for extracting hate speech conversations as it's very popular channel and hate speech is so common in this channel.

Twitter API provides users with facility of accessing features of Twitter. Twitter API allows users to have read and write access in Twitter, such as posting a tweet, reading user's profile information, finding tweets based on special content, or finding tweets based on tweet-id, which is an ID that identifies tweets. Standard Twitter API returns a batch of relevant Tweets which cover and match a specific predefined query. Standard Twitter API is not an exhaustive source for extracting tweets from twitter and just a sample of tweets is available and allowed to be extracted. Return rate limit is 180 tweets per 15 minutes for any authentication keys and tokens. Returned result of Standard Twitter API is in json format. Before connection to Standard Twitter API, authentication is required which is handled by accessing keys and tokens through Twitter developer account and creating Twitter app. We used "twitter" and "tweepy" libraries in python to interact with Twitter API and extract tweets. By creating Twitter app and authentication tokens, we established access to the Standard Twitter API.

The first purpose of this study is to investigate the conversations which include toxic comments. As the inherent of this study is conversational base, we need to gather conversations with toxic replies. There has been no attempt on studying conversations that include toxic comments.

Obtaining Toxic Replies. The paper (ElSherief et al., 2018) considers different classes for hate speech including archaic, class, disability, ethnicity, gender, nationality, religion and sex-orient. The paper introduces top 10 words or contents which represent any classes of hate speech in a best way. At the first step, we used these top 10 words for extracting toxic content. Top ten words for 8 classes of hate speeches can be found in Table 3 .1. All 8 hate speech classes are defined as "Toxic" class in Table 3.1. We considered all 8 classes in our study.

Toxic	Top 10 Contents indicating each class
Archaic	'Anti','wigger','hillbilly','bitch','white','chinaman','verbally','prostitute','vegetables'
Disability	'retards','legit','Only','yo','phone','#Retard','sniping','retarded','Asshole','upbringing'
Class	'Catholics','hollering','#racist','Cracker','#Virginia','Rube','#redneck','ALABAMA', 'batshit', 'DRINKS'
Sexorient	'meh','#faggot','#faggots','queers','hipster','NFL','pansy','Cuck','CHILDREN','FOH', 'wrists'
Gender	'dyke','dykes','chick','cunts','hoes','bitches','#CUNT','judgemental','aitercation', 'Scouse', 'traitorous'
Nationality	'Anti','wigger','bitch','white','chinaman','Zionazi','Zionazis','#BoycottIsrael', 'prostitute','#BDS'
Religion	'Algebra','Israelis','extermination','Jihadi','lunatics','catapults','Muzzie','Zionazi', '#BoycottIsrael', 'rationalize'
Ethnicity	'Anglo','spics','breeds','hollering','actin','coons','Redskins','Rhodes','#wifebeater', 'plantation'

Table 3.1. Shows top 10 words for 8 classes of hate speech (ElSherief et al., 2018)

We used Twitter Streaming API to extract tweets containing any of the ten words mentioned per class and we only extracted tweets in English language. We extracted data in

different rounds. We removed redundant repeated tweets from the gathered toxic data. We also used Hate Sonar, hate speech detection library for Python, (Davidson, Thomas, et al., 2017) to make sure if any gathered tweet is really toxic. We separated tweets which were considered as normal by Hate Sonar in a separated file. We randomly chose 100 of them and checked if they are really normal and not toxic. We found most of them normal because of two reasons. First, some tweets didn't contain special toxic content mentioned previously. Second, the specific toxic content was accompanied by other words which changed the meaning of the sentence to a normal sentence. For example, tweets containing "chick-fill-a" were gathered also as toxic comment because they contained content "chick" as toxic content. We let off separated normal tweets as normal and did not combine them with toxic tweets.

After gathering of data for any round we cleaned the tweet status. API streaming returns result in json format. To gather a conversation based on a toxic tweet for any gathered tweet we needed to first, extract the main tweet which the current toxic tweet is the reply to that, and second, extract replies to the main tweet. A conversation includes a "main" tweet and all its main replies. In this way, we could gather a whole conversation that at least contains one toxic tweet reply.

Obtaining main tweets for toxic replies. To obtain the main tweets, which current toxic comment is the reply to them, we handled the following process. We first used two features in Tweet metadata, namely, "in_reply_to_status_id" and "in_reply_to_screen_name", which refer to the tweet-id and username of main tweet which the current tweet is the reply to that.

There is no direct way to extract replies to a particular “main” tweet. We propose an algorithm for obtaining the replies. This algorithm leverages an algorithm called “Reply algorithm” to find the replies to the main tweet. Reply algorithm uses the “GetSearch()” method defined in “twitter” library in python to extract replies. We defined language item in “GetSearch()” method to be English and only gathered English replies. After fetching “in_reply_to_status_id” and “in_reply_to_screen_name”, from all gathered tweets, we used Twitter search API to do the following query [q="to:$ in_reply_to_screen_name ", sinceId = $ in_reply_to_status_id]. This query gathers all tweets which targeted specific user with username “in_reply_to_screen_name” and they are created after the tweet with “in_reply_to_status_id” tweet-id. Then we checked all the gathered results from previous step if their “in_reply_to_status_id_str” feature in their metadata corresponds to main tweet’s tweet-id. We considered any matched tweet a reply to the toxic tweet. This way, we completed the procedure toward creating conversations for a toxic tweet. Whole process regard conversation gathering is displayed in Figure 3.1. We ignored replies to replies as we needed net conversations. Sub conversations (replies to replies) could act as a noise in our data for studying conversations behavior and features.

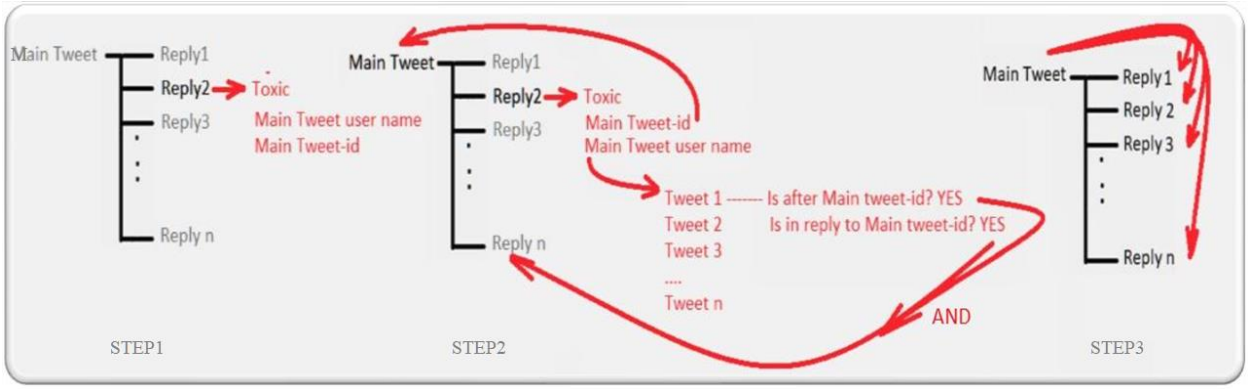

Figure 3.1 Shows the process for gathering a conversation

We considered minimum size of dataset for each class on 10,000 conversations per class. Different classes had different number of gathered data per round. Some classes like Archaic, and Nationality could have more than 10,000 conversations in 24 hours, Disability could have more than 10,000 conversations in just 10 hours, but for some other classes like Religion and Ethnicity we could not gather minimum number of data which we considered after several long rounds, so removed them from our research.

The result of all gathered data is in Table 3.2. Per toxicity class, the duration which was taken to gather the data is mentioned in "Duration" column. Total number of gathered conversations is specified in "Total" column.

Class	Duration	Total#
Archaic	24hr	114,589
Gender	2days	26,908
Class	6days	10,327
Nationality	24hr	130,302
Sex-orient	3days	10,378
Disability	10hr	20,657
Religion	11days and 10hr	1,345
Ethnicity	10days and 12hr	4,588

Table 3.2. Shows The total number of conversations per toxicity class

Gathered conversations had lengths between 2 to 17. Conversations with all lengths are considered in this study (2-17). Number of length 2 conversations in gathered dataset was 164,191 and number of conversations with length between 3 to 17 was 45,189. Number of conversations with different length can be found in Figure 3.2.

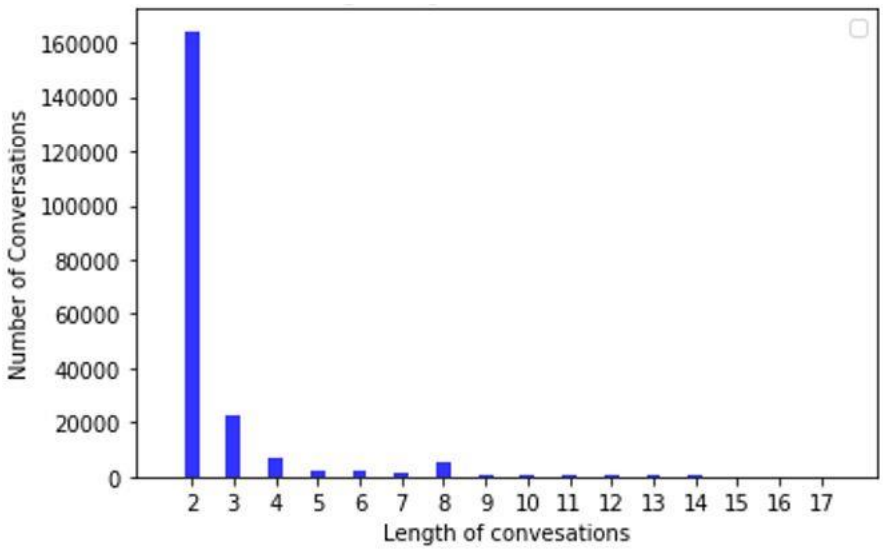

Figure 3.2. Shows Number of conversations with different length in gathered dataset

After replies extraction by replies algorithm, we used crawler algorithm to extract their text content. Crawler algorithm uses “tweepy” library and “get_status()” method to extract tweets metadata by tweet-ids gathered from replies algorithm. We removed conversations which we couldn’t extract main tweet status through get_status() method. It could happen because the main tweet was deleted, or the account was private. Descriptive statistics on all gathered dataset for all of 8 classes for main tweets and replies can be found in below Table 3.3. The table shows that different kind of users are considered in our study, some of them have 0 followers, friends and number of tweets and some of them have so many of them. Based

on the age in the table some the users are new and some of the are old, and verified feature is 3.014%.

Feature	Type	Count	Min	Max	mean	Median
Followers	Count	291,834 (100%)	0	108,440,077	56,291,605	409
friends	Count	291,834 (100%)	0	1,559,300	1,355	411
#tweets	Count	291,834 (100%)	0	4,860,247	24,642	7,545
age	Count	291,834 (100%)	0	13	5	5
verified	Boolean	8,798 (3.014%)				

Table 3.3. Shows Descriptive statistics on all gathered dataset

Data Cleaning. Since gathered replies from "Reply algorithm" are not in order based on time and date. We ordered tweets based on exact time and date. Another problem that we faced in this step was that Twitter Standard API search has recently posed a 7-day limitation. It means that it returns tweets generated in recent 7 days from current day. It's possible to assign a parameter named "until" to return tweets generated before the given date ("until"), but still has 7-day limitation, means that returns tweets generated in recent 7 days before the given date (until). This means if any main tweet was older than 7 days of the time replies algorithm was running, we would miss a part of the conversation. Therefore, we removed conversations with main tweets which the difference between main tweet's creation time and last reply's creation time is more than 7 days. As the result, we removed all the conversations that are not completed and are partially collected. The difference between main tweet and last reply can be found in Figure 3.3. We also separated this figure to two other figures, Figure 3.4 and Figure

3.5 to show the result clearer. Figure 3.4 shows conversations if the difference between main tweet and last reply is more than 7 days, so we removed these conversations. Figure 3.5 shows conversations if the difference between main tweet and last reply is less than 7 days, so we kept all of these conversations.

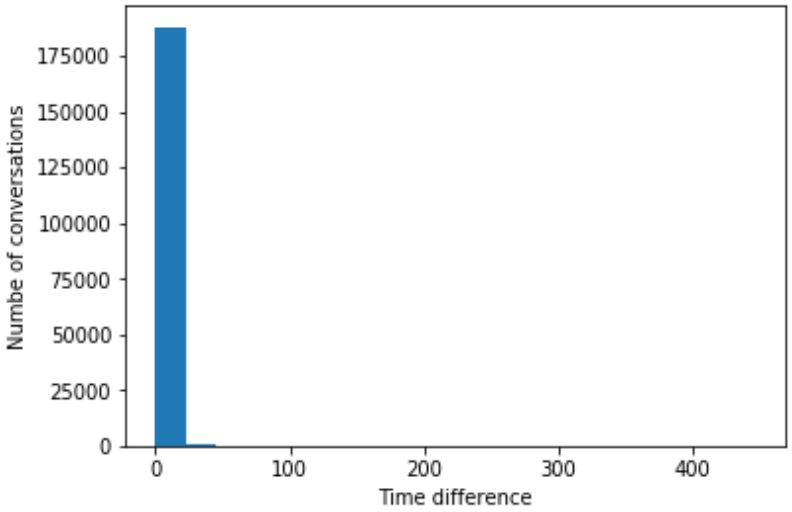

Figure 3.3. Shows the difference between main tweet and last reply

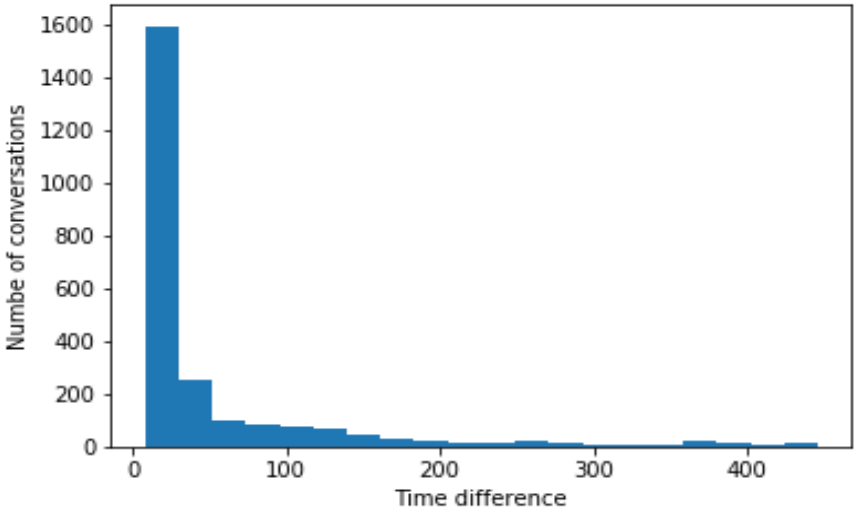

Figure 3.4. Shows conversations with time difference more than 7 days

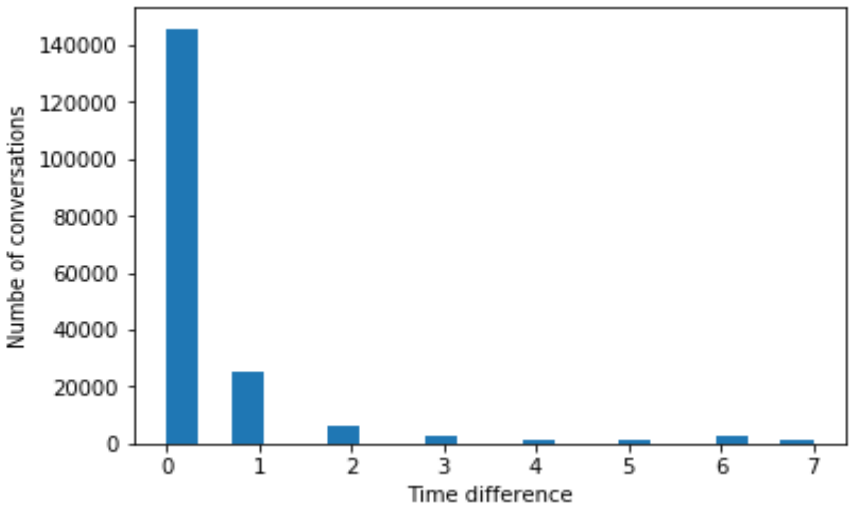

Figure 3.5. Shows conversations with time difference less than 7 days

Toxicity Detection in the conversations. After ordering tweets in conversations based on date and time, we checked all tweets in a conversation by Google Perspective API to find toxic comments. Google perspective API uses machine learning models to score the toxicity of a comment (ProgrammableWeb, 2019). Comments which in google Perspective API has probability of being toxic with more than 5% is considered as toxic and comments which in google Perspective API has probability of being toxic with less than 5% is considered as non-toxic. We labeled all non-toxic tweets as 1 and all toxic comments as 0.

Google Perspective API uses machine learning models to score the perceived impact a comment might have on a conversation. Google perspective API refers to different models of study, we applied following models in our analysis Toxicity, Severe-Toxicity, Obscene, Likely-To-Reject, Inflammatory. Each of the models supports detection of a type of toxic comment and the descriptions of all them can be found in Table 3.4. For all gathered conversations in dataset, we studied 5 models of Google Perspective API.

Models	Description
Severe- Toxicity	A very hateful, aggressive, disrespectful comment
Toxicity	Rude, disrespectful, or unreasonable comment
Likely- To - Reject	likelihood for the comment to be rejected according to the New York Times's moderation.
Obscene	Obscene or vulgar language
Inflammatory	Intending to provoke or inflame

Table 3.4. Shows Google Perspective models

We removed conversations that google perspective API could not recognize its main tweet toxicity score. We also removed conversations if their length has become below 2 due to the reply reduction in both get-status step and google perspective detection. Finally, we removed conversations if toxic was not detected in any of tweets in a conversation as just having the "hate speech" keywords do not guarantee toxicity. Final dataset after cleaning includes 96,253 conversations.

CHAPTER 4

TOXIC CONVERSATION STUDY

This chapter studies conversations including toxic comments. The dataset which was used for this study was a repository of conversations with different length (from length 2 to 17), including main tweet and all replies to the main tweet including at least one toxic tweet. All results from 5 models of Google Perspective API had similar results, so we only displayed the result of Severe Toxicity in this study.

We define conversation toxicity trend or the collective toxicity of a conversation as positive (non-toxic), negative (toxic) or neutral. A conversation is labeled as positive if the majority of comments in the conversation are non-toxic. A conversation is labeled as negative if the majority of comments in the conversation are toxic, and a conversation is labeled as neutral if the same amount of comments in the conversation are toxic and non-toxic. For example, trend of a conversation in Table 4.1 is non-toxic as total number of non-toxic tweets is more than total number of toxic tweets, and also ratio of total number of positive tweets over length of the conversation is more than 0.5.

Conversation tweets	**Probability**	**result**
Me clicking on the “RIP Tommy” trending section just to see the Tommy Robinson isn’t dead.	0.058	Non-toxic
I got proper fucking exited as fuck	0.79	Toxic
Trust me, Cam RIP won't even trend.	0.040	Non-toxic
Sick Cunts a fella is dead and you joke What Wankers #RipTommy	0.93	Toxic
lol London will burn, why would you wish that?	0.27	Non-toxic
You do realise Tommy Breslin was a successful football manager and his family and friends will see this?	0.016	Non-Toxic

Table 4.1 shows how collective toxicity of a conversation is calculated

We proposed some research questions and tried to answer them to find the impact of toxic replies on Twitter conversations.

1. How many of conversations are positive? negative? neutral?
2. Compare the length of positive vs. negative vs. neutral conversations.
3. Initial toxic comment occurs in which index of conversation (with considering index 0 as main tweet)?
4. Number of unique users per conversation and number of unique users participating in toxic comment per conversation.
5. Number of unique users participating in conversations after initial toxic
6. Conversations started with a toxic main will be negative
7. The position of hate comment affects the length of conversation.
8. What percent of conversations get non-toxic reply after initial toxic?
9. Having a positive reply on a toxic comment changes the discussion to be positive.

CHAPTER 4.1

RESULTS IN TOXIC CONVERSATION STUDY

In this chapter we examine all research questions that were mentioned in the previous chapter, and provide the results obtained from the whole dataset on Severe-Toxicity.

Study item 1: How many of conversations are positive? Negative? Neutral?

We found 10,879 (%11.3%) of conversations are positive, 20,332 (31.2%) are negative and 65,042 (67.5%) are neutral. Positive conversations are conversations which non-toxic commenters are more active in them. Negative conversations are conversations which toxic commenters are more active in them. Neutral conversations are conversations which both toxic and non-toxic commenters are active at the same level. Figure 4.1 also shows total number of conversations in each collective toxicity trend. Based on these result number of neutral conversations is more than negative conversations and number of negative conversations is more than positive conversations. The result shows that in most of conversations toxic and non-toxic commenters are active at the same level. With a high difference after neutral, toxic-commenters are more active than non-toxic commenters. Result shows a higher number of negative conversations in compare to positive conversations, so it means toxic comment most of the time has destructive effects on conversation and encouraged other users to engage in toxicity, or as an alternative interpretation, other users could not change the trend of the conversation by non-toxic comments. Less number of conversations had positive trend, and it means toxic comment did not have so much destructive effects on some conversations and did not encourage

other users in toxicity engagement, or an alternative interpretation is that in some conversations a user or some users changed the trend of the conversation by non-toxic comments. Figure 4.1 shows most of negative conversations get ratio of 0 which means that all tweets in these conversations are toxic, and it means toxic comment encouraged other users in toxicity engagement, In compare to negative conversations, positive conversations do not get exactly ratio of 1, and their ratio is displayed in a range between 0.5 and 1 because they include at least one toxic comment, but it shows toxic comment did not have so much destructive effect on conversation and could not encourage other users in toxicity engagement, or there are other users who tried to change the conversation trend by non-toxic comments.

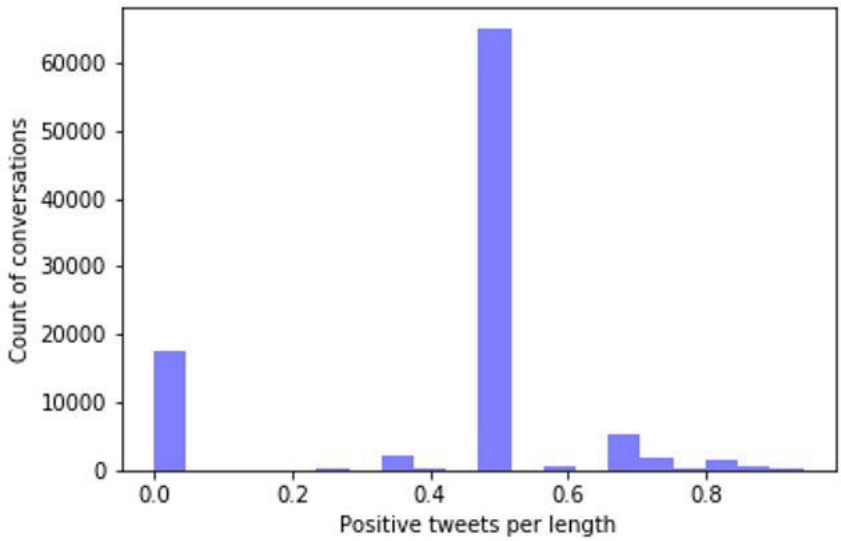

Figure 4.1 Shows total number of conversations in each collective toxicity trend

Study item 2: Compare the length of positive vs. negative vs. neutral conversations.

The length of conversations for negative, positive and neutral conversations are displayed below in two different figures, one in percentile and the other in count. Figure 4.2 shows

length of conversations for negative, positive and neutral conversations for total number of conversations in count and Figure 4.3 shows length of conversations for negative, positive and neutral conversations for total number of conversations in percentile.

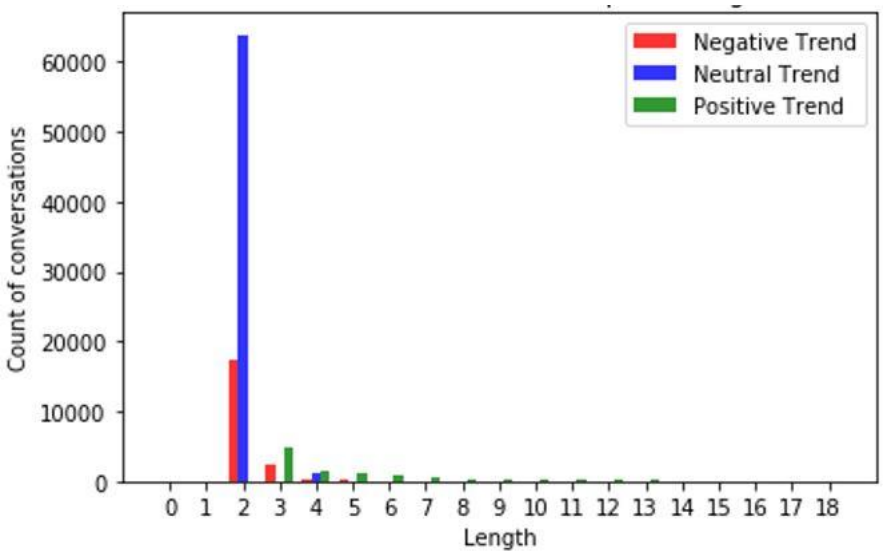

Figure 4.2 shows length of conversations for total number of conversations in count

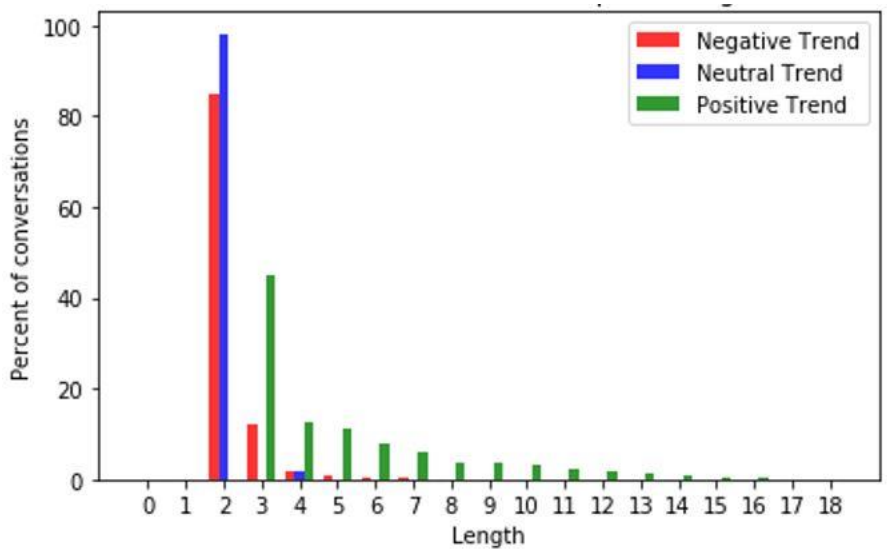

Figure 4.3 shows length of conversations for total number of conversations in percentile

Above figures show that the number of negative conversations is more than positive conversations, but length of positive conversations is longer. Perhaps having more user

engagement helps the positivity of conversations, or maybe users stop in engaging in conversations which most of their tweets are toxic. Our finding also approves that of Kwak et al., 2018 study, which shows players of the online games are not surprisingly active in generating toxic. Our result also shows just limited number of users are interested in toxic engagement.

- Study item 3: The first toxic reply occurs in which index of conversation, where main tweet has index 0?

Figure 4.4 shows percentage of conversation which their first toxic comment occurs in an index. Also, Table 4.2 shows this percentage in more details per index.

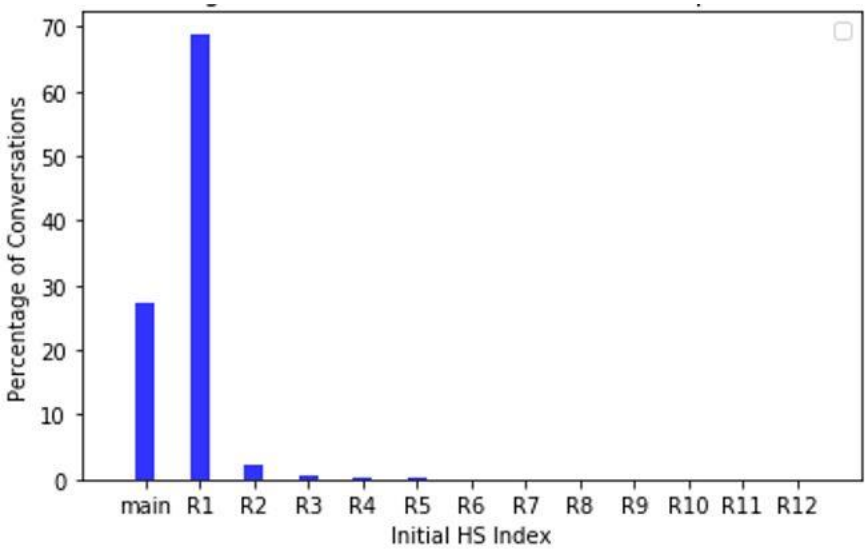

Figure 4.4 Shows percentage of conversation with initial toxic in special index

Index	Percent
0	27.27
1	68.86
2	2.39
3	0.68
4	0.34
5	0.17
6	0.09
7	0.06
8	0.04
9	0.02
10	0.014
11	0.005
12	0.005
13	0.001

Table 4.2 Shows percentage of conversation with initial toxic in special index

Based on the results, most of initial toxic comments occur in first reply to the main tweet. So, perhaps it shows the probability of being attacked by toxic is highest in the first reply to the main tweet, even if main tweet is toxic or non-toxic. After first reply, main tweet with index 0 has highest percentage of being toxic. So as a complementary to Zhang et al,.2018 which shows whether an initially healthy conversation may be affected by personal attacks, we show in which index of conversation the probability of being attacked is more than other indexes.

Study item 4: Number of unique users per conversation and number of unique users participating in toxic comment per conversation.

Number of unique users per conversation is displayed in Figure 4.5 and number of unique users participating in toxic comment per conversation is displayed in Figure 4.6. So, as an approval to Kwak et al., 2018, study, which shows players of online games are not surprisingly

active in generating toxic, we also see that just a limited number of unique users participate in conversations that include toxic replies. Also, we showed the number of unique users posting toxic comments in more details by showing number of conversations in log in Figure 4.7.

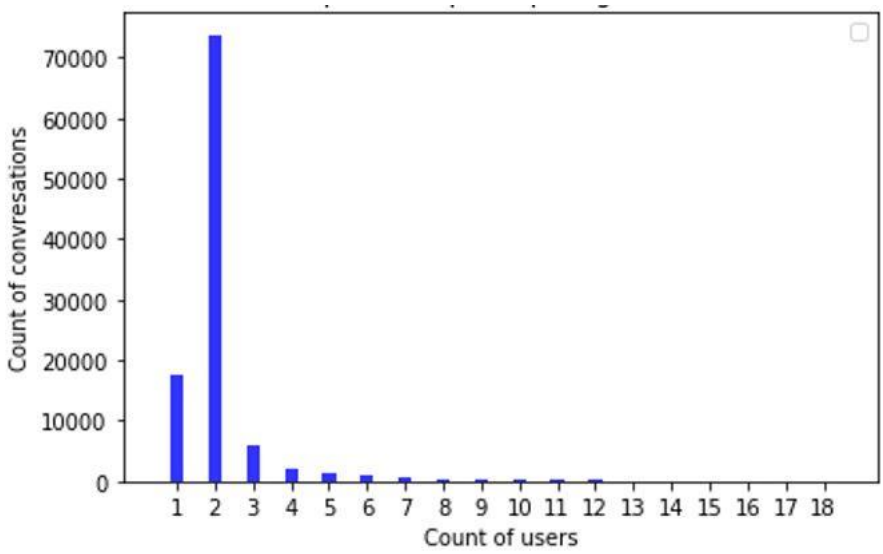

Figure 4.5. Shows number of unique users per conversation

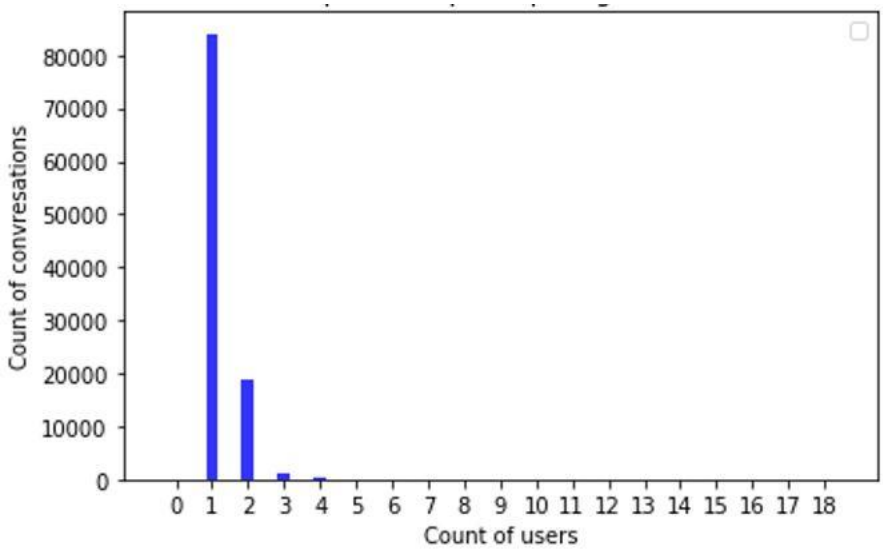

Figure 4.6. Shows number of unique users participating in toxic comment

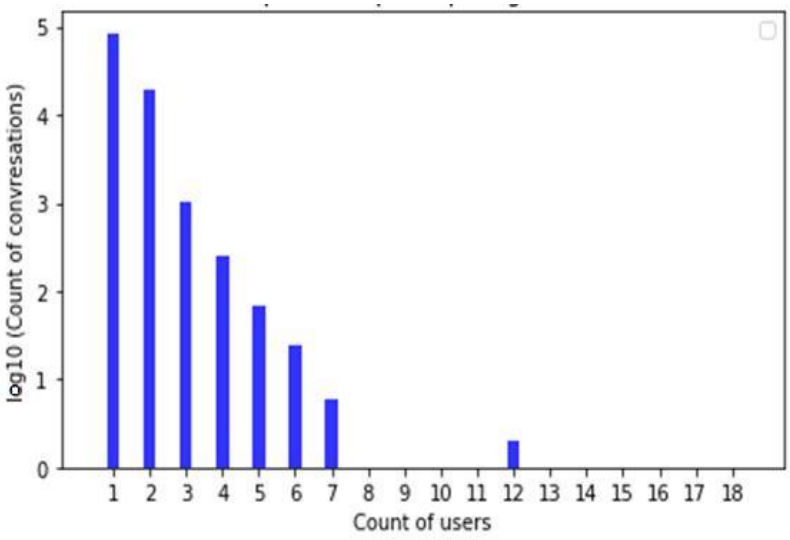

Figure 4.7. Shows number of unique users participating in toxic comment with total number of conversations displayed in log

Study item 5: Number of unique users participating in conversations after the first toxic reply.

Conversations with length 2 are removed from this study. Figure 4.8 shows number of unique users participating in negative or positive conversation after initial toxic. Results show after the first toxic reply, users either stop participating in the discussion, or if they do, they contribute to the positivity. So, as the complementary to Zhang et al,.2016 study, which shows audience feedback on debates relates to debate outcome, perhaps user engagement or feedback can affect the flow of debate or trend of a conversation.

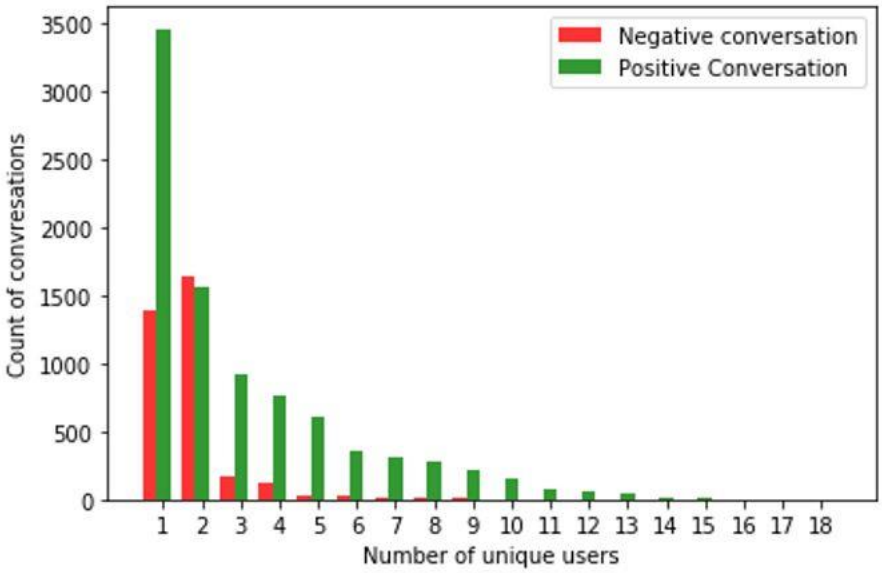

Figure 4.8. Shows number of unique users in negative and positive conversations

Study item 6: Conversations started with a toxic main will be negative.

We studied this study item by considering the correlation between two variables of "total trend of the conversation" and "main tweet (toxic or non-toxic)". Pearson Correlation between these two variables was 0.740 with p-value of 1.4012985e-45. The correlation showed most of negative conversations start with a toxic main, however most of positive conversations start with a positive main. We also found from total number of 10,879 positive conversations, 1,660 of them start with toxic main, and 9,219 of them start with non-toxic main, and from total number of 20,332 negative conversations, 19,529 of them start with toxic main, and 803 of them start with non-toxic main. This result also approves Zhang et al., 2016 which shows toxic engagement can be surprisingly increase by toxic demand, our results also show toxic comment can encourage other toxics in a conversation.

Study item 7: The position of toxic comment affects the length of conversation.

We studied this study item by considering the correlation between two variables of "initial toxic index" and "Length of the conversation". Spearman Correlation between these two variables was 0.127 with p-value of 1.4012985e-45. As correlation was weak, it is also a complementary to previous result, which regardless to the initial index of toxic comment, after the first toxic comment, users either stop participating in the discussion, or if they do, they contribute to the positivity.

Study item 8: What percent of conversations get non-toxic reply after initial toxic?

As an example, for this study, imagine a conversation shown below, where the first toxic reply is reply 2. We want to examine in what percent of conversations reply 3 is also a toxic reply and in what percent of conversations reply 3 is a non-toxic reply.

Example: "Main tweet, reply1, reply2 (initial toxic), reply3 (toxic or non-toxic), reply4, reply5.

In Figure 4.9, "HS+" shows conversations that get non-toxic reply after the first toxic reply (i.e., when reply 3 is non-toxic), "HS-" shows conversations that get toxic reply after the first toxic reply (i.e., when reply 3 is toxic). Based on the results, portion of conversations that get non-toxic reply after initial toxic is less than portion of conversations which get toxic reply after initial toxic. This can show that toxicity is contagious, and one toxic encourages next toxic.

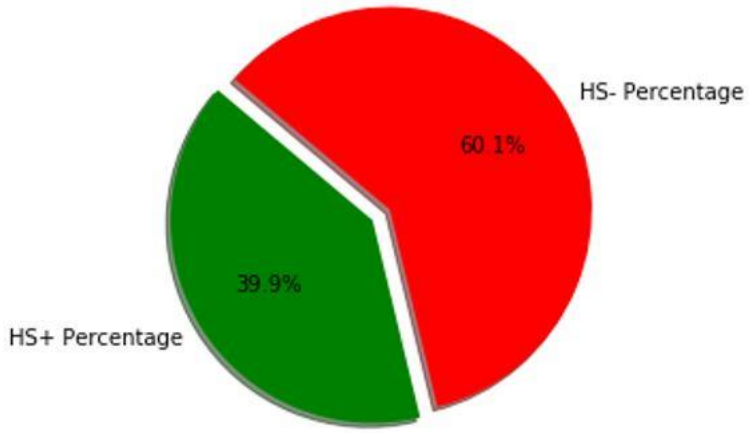

Figure 4.9. Shows what percent of conversations get toxic or non-toxic reply immediately after initial toxic

Study item 9: Having a positive reply after a toxic reply can affect the flow of conversation positively.

As an extension to previous study item, Figure 4.10 shows conversation toxicity trend after the first toxic reply, where "HS+" represents conversations which get non-toxic reply after initial toxic, and "HS-" represents conversations which get toxic reply after initial toxic. The results show that conversations which get toxic reply after initial toxic are more negative and conversations which get non-toxic reply after initial toxic are more positive. This finding also refers to the contagious toxicity which occurrence of a toxic after initial toxic may encourage other toxicity and makes the conversation trend negative. It also shows that if users post a non-toxic reply after the first toxic, it can change the conversation trend positively.

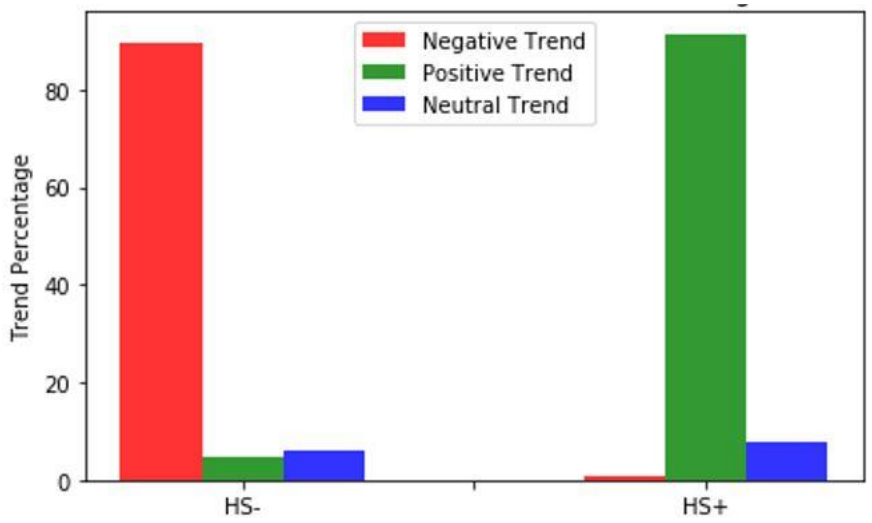

Figure 4.10. Shows conversations toxicity trend after initial toxic with considering getting toxic or non-toxic reply immediately after initial toxic.

We also examined this study item by analyzing the correlation between two variables of "Conversation trend after initial toxic" and "Non-toxic or toxic reply immediately after initial toxic". Spearman correlation was 0.856 with p-value of 1.4012985e-45. The correlation also approves previous results which indicate conversations get toxic reply after initial toxic are more negative and conversations get non-toxic reply after initial toxic are more positive.

Based on these results, in conversations which at least include one toxic comment, we found number of negative conversations are more than positive conversations, but length of positive conversations is longer. Perhaps having more user engagement helps the positivity of conversations. We also found limited number of unique users participating in hate comments. Based on the results, most of initial toxic comments occur in first reply to the main tweet. After the first toxic reply, users either stop participating in the discussion, or if they do, they contribute to the positivity. Most of the time conversations started with a toxic main will be

negative and conversations started with a non-toxic main will be positive. Most of the conversations get toxic reply immediately after initial toxic and it can show toxicity is contagious and as the result makes conversation trend more negative. However, if conversations get non-toxic reply immediately after initial toxic, they help to make conversation are more positive.

CHAPTER 5

THE USE of INFORMATION RETRIEVAL MODEL For BAD

The other objective of this research is to create a new toxic speech for an utterance. For this purpose, we tried to use Natural Language Processing (NLP) to generate a relevant toxic reply for any given tweet. As explained in Chapter 2, NLP models can be divided into Generative models and Retrieval models. We used Retrieval models in this study. Retrieval models are based on obtaining relevant reply for a given utterance. Retrieval models query a targeted utterance in a repository and return a relevant reply that best matches the given utterance (Song et al., 2018). Figure 5.1 shows an Information Retrieval model, which includes the following components: Knowledge Database, Retrieval algorithm and Ranking algorithm.

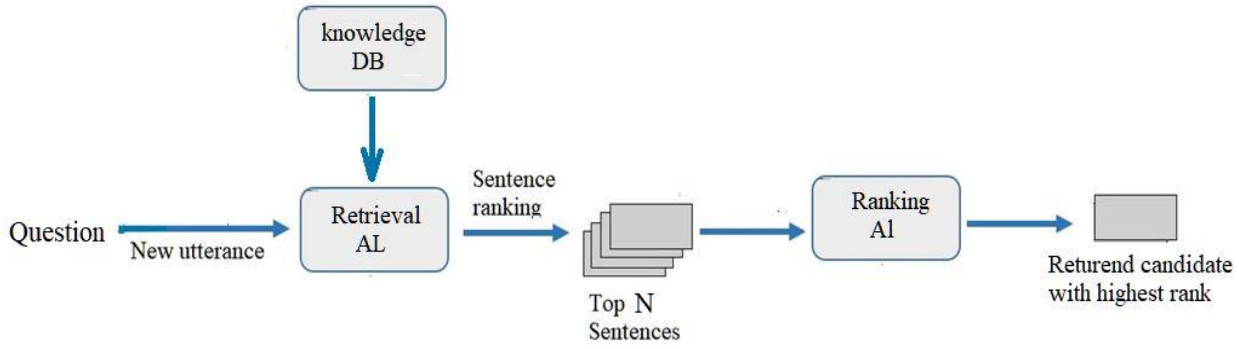

Figure 5.1. Shows Retrieval model

There are different models for text retrieval. The problem in text retrieval is to design a ranking function which ranks the candidates in a query. There are different methods to design the ranking function. The purpose of ranking function is to rank relevant sentences on top of non-relevant ones. The way the relevancy is measured defines a ranking method. Based on different definitions of relevancy, there are different kinds of ranking modules.

The Retrieval algorithm phase is based on calculating similarities between a given utterance and candidates in knowledge database. The Ranking algorithm phase is based on calculating similarities between the given utterance and candidates retrieved from previous step in Retrieval algorithm. In similarity calculation, Cosine Similarity as a common method is used in different sentence representations methods (Novgorodov et al., 2019). Selected sentence representation method should provide fixed-length vector representations to compare and find the difference. Two sentence representation algorithms are used in this project, Paragraph vectors (Doc2Vec) as Retrieval algorithm and Word vectors (Word2Vec) as Ranking algorithm. We used Genism library for implementation.

In Word2Vec, word vectors are inferred from predicting next word in the sentence, and all word vectors are initialized randomly. As an indirect result of this predictions, words can capture semantics. There are two types of Word2Vec, Skip-gram (sg) and Continuous Bag of Words (CBOW) (QV Le at al., 2014). In Skip-gram, input is a word vector and output is a context word. In CBOW, input is a sum vector of multiple words and output is a context word (JH Lau at al., 2016) (Figure 5.2) (Mikolov et al., 2014). After training words with similar meanings will have similar vectors and as the result similar positions in vector space (Mikolov et al., 2014).

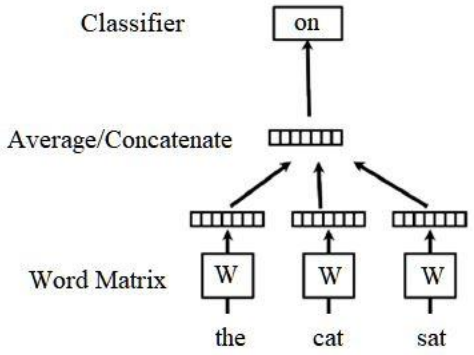

Figure 5.2 Shows Word2Vec CBOW method (Mikolov et al., 2014)

Paragraph vectors (Doc2Vec) is a sentence representation method and is an unsupervised algorithm for representing variable length of texts. Any piece of text in this algorithm is represented by a dense vector (fixed-length vector). This approach is inspired by word vectors. Paragraph vectors are the extension to the word vectors. Paragraph vectors in Doc2Vec algorithm are trained to predict words in the paragraph by gradient descent and back propagation (Rumelhart et al., 1986), training continues till convergence. Paragraph vectors in training are unique per text piece, but word vectors are shared among all texts. Paragraph vectors were initially introduced for sentiment analysis and text classification (Mikolov et al., 2014). There are two types of Doc2Vec, Distributed Memory Model of Paragraph Vectors (PV-DM) and Distributed Bag of Words Paragraph Vector (PV-DBOW).

DMPW works same as CBOW. The input is the concatenation of paragraph vector and several word vectors and output is a context word. It means that this time paragraph vectors also contribute in next word prediction (Figure 5.3) (Mikolov et al., 2014).

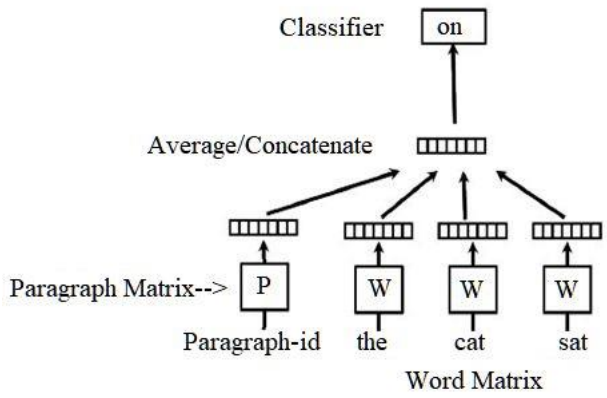

Figure 5.3. Shows Doc2Vec DMPV method (Mikolov et al., 2014).

DBOW works same as Skip-gram with the difference that the input is a paragraph vector and output is a context word (Mikolov et al., 2014).

Most of the time, retrieval algorithms top ranked returned result is not the exact answer to the question. We tried to re-rank returned answers and come up with the most suitable answer. Our main focus in this study is on re-ranking retrieved sentences to find the most relevant toxic reply for a tweet. Most of previous studies on question and answering systems and text retrieval algorithms used TF-IDF or specialized vector spaces, or simple string matching for Retrieval algorithm. We used paragraph vectors (Doc2Vec) (Mikolov et al., 2014). Most of previous studies didn't apply second ranking for their retrieval model. Recently specialized machine learning approaches are used as second ranker approach. We used word vectors (Word2Vec) (Mikolov et al., 2013).

We combined two methods of Doc2Vec as first ranker (Retrieval algorithm) and Word2Vec as second ranker (Ranking algorithm) to get a better result. We tested the use of only Doc2Vec or Word2Vec methods separately and we found the results not acceptable. We

propose to work on top 10 returned results of Doc2Vec and try to find the best toxic reply among those ten by using Word2Vec method.

CHAPTER 6

CREATING KNOWLEDGE DATABASE

In this chapter, we explain how we created knowledge database for second objective which was “examine the feasibility of automatically creating toxic replies for a tweet” through Natural Language Process models. The aim of this work is to create relevant and appropriate comment which can act like a toxic comment and makes the online conversation flow more negative. In this regard, we needed to gather datasets including toxic comments. As the inherent of this study is an interactive model, and we must create one toxic reply in response to an online utterance, we needed to gather conversations including toxic replies. We decided to gather length two conversations including one main tweet and one toxic reply to the main tweet.

The same as our previous dataset, we tried to extract new tweets containing special content from Twitter. The first step in this regard was to recognize the contents which can refer to a toxic speech tweet. We used the top ten words that can help detecting gender-based toxic content (ElSherief et al., 2018), including: 'dyke', 'dykes', 'chick', 'cunts', 'hoes', 'bitches', '#CUNT', 'judgemental', 'aitercation', 'Scouse', and 'traitorous'. We used Twitter Streaming API to extract tweets containing any of the ten words and we only extracted tweets in English language.

We extracted data in different rounds. In the first round, we ran script for around 6 hours, and as the result 115,906 tweets as metadata (646 MB json file) were gathered. After

gathering of data for first round we cleaned the tweet status. For any gathered tweet metadata, we extracted "id_str", and “in_reply_to_status_id”. First feature refers to the current tweet’s tweet-id and the second feature refers to the main tweet’s tweet-id. Main tweet is a tweet which current toxic tweet is the reply to that. We removed redundant repeated toxic tweets from the gathered toxic speech data. We used Hate Sonar library to make sure if any gathered tweet is really toxic speech. We removed tweets which were considered as normal by Hate Sonar in a separated file. We randomly chose 100 of them and checked if they are really normal and not toxic speech. We found most of them normal because of two reasons. First, some tweets did not contain special toxic content mentioned previously. Second, the specific toxic content was accompanied by other words which changed the meaning of the sentence to a normal sentence. For example, tweets containing “chick-fill-a” were gathered also as toxic speech because they contained content “chick” as toxic content. We let off separated normal tweets as normal and did not combine them with toxic speech tweets. From 115,906 gathered tweets, 18,791 of them were recognized as normal (non-toxic) and 97,139 of them as toxic.

Till this step, for the first-round data gathering we collected toxic’s tweet-id and main tweet’s tweet-id (main tweet is a tweet which current toxic tweet is reply to that). The result was a length 2 conversation with a toxic or normal main tweet and a toxic reply. As we just could gather a small number of conversations in first round, we decided to have longer script run and more rounds. We used Crawler algorithm to extract whole tweet-ids in each length 2 conversation.

The result for total rounds can be found in Table 6.1. All dataset is gathered in June 2019 and total number of 2,373,306 tweets as toxic is gathered by Twitter Streaming. Among these number of tweets, 1,646,720 (69.38%) of them is considered as toxic by hate sonar and 598,253 of them is considered as normal (removed from dataset). From 1,646,720 recognized toxic tweets 110,655 of them referred to a main tweet in their status. So, till this step we gathered 110,655 length 2 conversations.

Date	Total	Toxic	Normal	Main
June 2019	2,373,306	1,646,720	598,253	110,655

Table 6.1. Shows Total Gathered datasets for second objective

We Removed conversations that we could not extract main tweet for their toxic reply in get-status() method. We removed conversations which their reply was not considered as toxic by Severe-Toxicity (Main tweet can be toxic or not). Our final cleaned dataset included 63,478 conversations with length 2 including main tweet and its corresponding toxic reply. Descriptive Analysis on the dataset for main tweets and toxic replies can be found in below Table 6.2. The table shows different kind of users are considered in our study, some of them have 0 followers, friends and number of tweets and some of them have so many of them. Based on the age in the table some the users are new and some of the are old, and verified feature is 3.14%.

Feature	Type	Count	Min	Max	mean	Median
Followers	Count	169,546 (100%)	0	108,716,624	73,676	569
friends	Count	169,546 (100%)	0	1,566,827	1,255	432
#tweets	Count	169,546 (100%)	0	4,853,285	30,908	11,011
age	Count	169,546 (100%)	0	13	4	4
verified	Boolean	5,323 (3.14%)				

Table 6.2. Shows Descriptive Analysis on all gathered dataset

We cleaned the dataset (removed usernames, hashtags, punctuations, tweet-ids from tweet), and whole dataset was in lower case. We split our dataset to 90% train and 10% test. To create our knowledge database, we inspired from Key-value profile memory network of Saizheng Zhang (2018). We applied this approach to our dataset, we considered main tweets in our dataset as keys and their corresponding toxic replies as values (Zhang et al,.2018), (Figure 6.1).

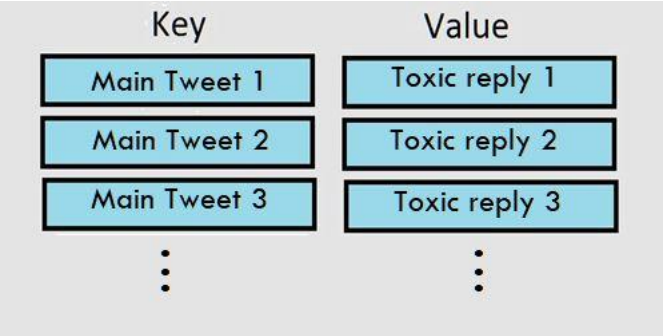

Figure 6.1. Shows how we create knowledge database

CHAPTER 7

ALGORITHM

In this chapter we show how our algorithm works. We trained our Word2Vec model based on our train dataset. We modeled all keys in train set through Doc2Vec. In first ranking per any new utterance, Doc2Vec uses Cosine Similarity to return top 10 sentences from dataset (keys) which match the given utterance more than others. We listed top ten returned utterances (keys) and their corresponding replies (values) as candidates to reply to the given utterance.

In second ranking we used average word vectors by using Word2Vec for top 10 keys and corresponding values and the new utterance to gain a new vector for new utterance and each key and value. Then we computed two distances. First distance between key and new utterance, second between value and new utterance. Distance is calculated by cosine similarity. We averaged the gained distances from key-utterance and value-utterance. We repeated these steps for all 10 keys and values and at the end returned the value which has shortest average distance as appropriate answer to the given utterance. Total implemented algorithm can be found Figure 7.1:

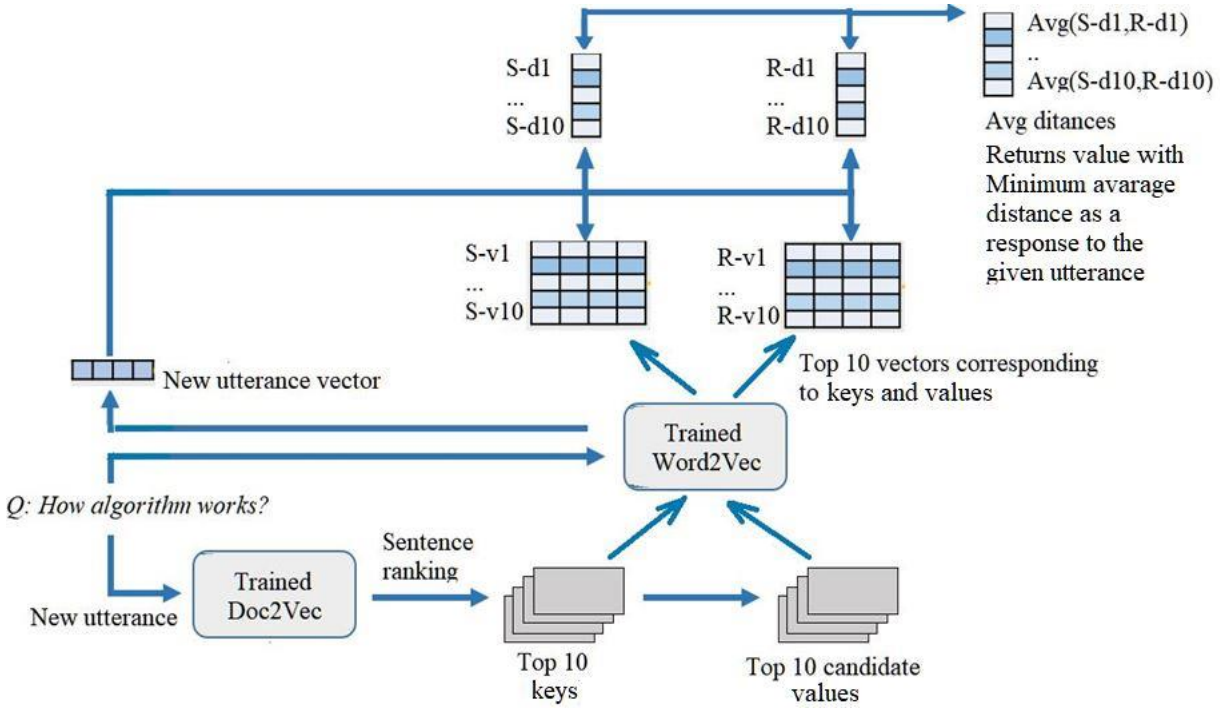

Figure 7.1. Shows how algorithm works in first ranking and then second ranking

CHAPTER 8

EVALUATION

In previous chapter, we combined two methods of Doc2Vec as first ranker and Word2vec as second ranker. Both Doc2Vec and Word2Vec in previous chapter were trained on our own dataset. To evaluate our result, we used pre-trained Doc2Vec and Word2Vec models on Wikipedia as a large-scale open domain dataset. Training on such large datasets takes so much time so we preferred to use pretrained models. These pretrained models on large datasets like Wikipedia provides us with more accurate and justified weights in training so we could compare the result from them to our own dataset. We applied other scenarios in our training and compared the results.

The second scenario was to apply both pretrained Doc2Vec and Word2Vec models on Wikipedia. We used pretrained model of Doc2Vec and tried to gain all vectors corresponding to our training dataset (keys). For any given utterance in test set we also gained vector through Doc2Vec pretrained model. We gained top 10 most similar vectors from training keys by computing cosine distance between utterance vector and all training keys. For last comparison of Word2Vec we did same as previous explanation, we computed new vectors from pretrained Word2Vec for utterance, top 10 returned keys and their values. We computed cosine similarity between key-utterance and value-utterance for all 10 returned results. We averaged the distances and returned the value which their average distance is minimum as answer to the utterance.

The third scenario was to apply pretrained Doc2Vec on Wikipedia and trained Word2Vec model on my own dataset. We repeated all the steps as before for test set.

The forth scenario was to apply pretrained Doc2Vec on Wikipedia and simple string matching on our own dataset. After gaining top 10 most similar keys from dataset. We used simple string matching for new utterance from test set and both top 10 keys and values. Simple string matching is based on finding intersections between two strings. We computed the cosine distance and calculated average distance same as previous steps. We returned nearest value as answer to the utterance.

CHAPTER 8.1

AMAZON MECHANICHAL TURK

Human evaluation was used in this project. Amazon Mechanical Turk (Mturk.com, 2019) as a channel for crowd sourcing was used in evaluating the result of toxic speech generation. By using Amazon Mechanical Turk, we could have access to workers in all over the world and around the clock. Speed and accuracy are two important features for Amazon M Turk. Workers are able to work in 24 hours and in parallel.

We evaluated our four scenarios through Amazon M Turk. We randomly chose 100 samples from our test set and their corresponding responses from four scenarios. As the result per scenario we had 100 utterances from test set and their corresponding responses from that scenario. We defined each sample as a task and allocated each task to 3 workers, and paid them 60 cent. Workers were located in United states and were Master workers. We evaluated relevancy of our length 2 discussions by asking workers to rate the given response based on the given utterance between 1 to 5. We considered 1 as “Not related at all” means that response is not related to the utterance at all, and 5 as “Totally related” means that response is totally related to the utterance. The results are reported in Table 8.1 for all four scenarios. Our first scenario which was trained Doc2Vec and Word2Vec on our own dataset had highest relevancy score and with a little difference scenario of pretrained Doc2Vec and Word2Vec on Wikipedia had highest score in compare to the other scenarios.

Model	Relevancy score
Trained Doc2Vec and Word2Vec on our dataset	3.272
Pretrained Doc2Vec and pretrained Word2Vec	3.228
Pretrained Doc2Vec and Trained Word2Vec on our dataset	2.851
Pretrained Doc2Vec and self-trained simple string matching	2.743

Table 8.1 Shows relevancy score of all toxic creation models based on human evaluation.

CHAPTER 9

CONCLUSION AND FUTURE WROK

In this study we showed how toxic comments act in Twitter conversations. This study also was as a complementarity or confirmation to other results from other studies like Kwak et al., 2018, Zhang et al., 2018, Zhang et al., 2016.

We showed most of the time activity level of toxic commenters and non-toxic commenters are at the same level and it makes conversations neutral. But with a considerable difference after that, toxic commenters are more active than non-toxic commenters and as the result it makes conversations trend more negative.

We also showed perhaps having more user engagement helps the positivity of conversations, so, as the complementary to Zhang et al,.2016 study, which shows audience feedback on debates relates to debate outcome, perhaps user engagement or feedback can affect the flow of debate or trend of a conversation. We also showed maybe users stop in engaging in conversations which most of their tweets are toxic, so it shows hate comments affect conversations negatively, and they can reduce user engagement in conversations. We showed position of hate comment doesn't affect the length of conversation and regardless to the initial toxic index, after the first toxic comment, users either stop participating in the discussion, or if they do participate, they contribute positively.

Consistent with (Kwak et al., 2018) study, which shows players of online games are surprisingly not active in generating toxic, our study also shows just a limited number of users are interested in engaging in toxicity. We also showed the probability of being attacked by

toxic is highest in the first reply to the main tweet, even if main tweet is toxic or non-toxic. After first reply, main tweet has highest percentage of being toxic.

We also approved Zhang et al,.2016 which shows toxic engagement can be surprisingly increase by toxic demand, our results also show toxic comment can encourage other toxics in a conversation. Based on the result, portion of conversations that get non-toxic reply after the first toxic reply is less than the portion of conversations that get toxic reply after the first toxic reply. Also, this can show that toxicity is contagious, and one toxic reply encourages the next toxic replies. Initiating a conversation with toxic comment can affect the conversation trend to be toxic. We showed conversations which get toxic reply immediately after initial toxic have negative conversation trend.

In the last part of our research we worked on text generation tool to create toxic reply for an utterance or a tweet. We used Retrieval models of Doc2Vec as first ranker and Word2Vec as second ranker. We tried to re-rank returned candidates from first retrieval algorithm and come up with the most suitable answer. Our main focus was on re-ranking retrieved sentences to correctly finding toxic reply for a given utterance. As the result, we showed that its possible for an adversary to abuse text generating tools to create toxic comments with the goal of negatively affecting conversations.

As a future work, we plan to study the effects of followers and friends in conversations to see if they contribute to the positivity or negativity of a conversation. For this purpose, we will study if they engage in a conversation which started by a toxic comment or includes at least one toxic comment. Whether they initiate toxic in a conversation or try to make the

conversation positive after initial toxic. We will also recommend working on toxic creation through generative models. Generative models generate new replies, they create a new sentence word by word. (Song et al., 2018). Generative models may help to create more accurate and relevant reply to the given utterance.

REFERENCES

1. Ana Cristina Mendes and Lu´ısa Coheur. 2011 "An approach to answer selection in question-answering based on semantic relations." *In Toby Walsh, editor, IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, pages 1852–1857. IJCAI/AAAI, 2011. ISBN 978- 1-57735-516-8.*

2. Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." *International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 2012. https://doi.org/10.1109/socialcom-passat.2012.55.*

3. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. "Abusive Language Detection in Online User Content." *In Proceedings of the 25th International Conference on World Wide Web, pages 145–153. International World Wide Web Conferences Steering Committee.*

4. Council on Foreign Relations. 2019. "Council on Foreign Relations." *[online] Available at: https://www.cfr.org/ [Accessed 3 Dec. 2019].*

5. Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. "Reading wikipedia to answer open-domain questions." *arXiv preprint arXiv:1704.00051.*

6. Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. "Mean birds: Detecting aggression and bullying on twitter." *In Proceedings of the 2017 ACM on Web Science Conference, WebSci '17, pages 13–22, New York, NY, USA. ACM.*

7. Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying." *ACM Transactions on Interactive Intelligent System, 1–30. https://doi.org/10.1145/2362394.2362400.*

8. Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. "Hate Speech Detection with Comment Embeddings." *Proceedings of the 24th International Conference on World Wide Web - WWW 15 Companion, https://doi.org/10.1145/2740908.2742760.*

9. Gitari, Njagi Dennis, Zuping Zhang, Hanyurwimfura Damien, and Jun Long. 2015. "A Lexicon-Based Approach for Hate Speech Detection." *International Journal of Multimedia and Ubiquitous Engineering 10, no. 4 : 215–30. https://doi.org/10.14257/ijmue.2015.10.4.21.*

10. Haewoon Kwak, Jeremy Blackburn, Seungyeop Han. 2015. "Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, April 18-23, 2015, Seoul, Republic of Korea.*

11. Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. "Conversations Gone Awry: Detecting Early Signs of Conversational Failure." *In Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018. 1350–1361.*

12. Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In Proceedings of NAACL

13. Kayes, Imrul, Nicolas Kourtellis, Daniele Quercia, Adriana Iamnitchi, and Francesco Bonchi. 2015. "The Social World of Content Abusers in Community Question Answering." *Proceedings of the 24th International Conference on World Wide Web - WWW 15, https://doi.org/10.1145/2736277.2741674*.

14. Lee, Jinhyuk, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. "Ranking Paragraphs for Improving Answer Recall in Open-Domain Question Answering." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d18-1053.

15. Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth M. Belding-Royer. 2018. "Peer to Peer Hate: Hate Speech Instigators and Their Targets." arXiv, cs.SI.

16. Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. "Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media." arXiv preprint arXiv:1804.04257.

17. MediaSmarts. 2019. "MediaSmarts." *[online] Available at: http://mediasmarts.ca/ [Accessed 3 Dec. 2019].*

18. Medveď, Marek, and Aleš Horák. 2018. "Sentence and Word Embedding Employed in Open Question-Answering." *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, https://doi.org/10.5220/0006595904860492*.

19. Mehdad, Yashar, and Joel Tetreault. 2016. "Do Characters Abuse More Than Words?" *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, https://doi.org/10.18653/v1/w16-3638*.

20. Mturk.com. (2019). Amazon Mechanical Turk. [online] Available at: https://www.mturk.com/ [Accessed 11 Dec. 2019].

21. Novgorodov, Slava, Ido Guy, Guy Elad, and Kira Radinsky. 2019 "Generating Product Descriptions from User Reviews." *The World Wide Web Conference on - WWW 19, 2019. https://doi.org/10.1145/3308558.3313532*.

22. ProgrammableWeb. (2019). Perspective. [online] Available at: https://www.programmableweb.com/api/perspective [Accessed 11 Dec. 2019].

23. Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature 323, no. 6088: 533–36. https://doi.org/10.1038/323533a0.*

24. Sharp Bernadette, Delmonte Rodolfo, Acosta Olga., et al, 2015. Natural Language Processing and Cognitive Science. Proceedings 2014. Berlin, Boston

25. Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2017. "Evidence aggregation for answer re-ranking in open-domain question answering." *arXiv preprint arXiv:1711.05116.*

26. Shuohang Wang, Mo Yu, Xiaoxiao Guo, ZhiguoWang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2017. "R3: Reinforced reader-ranker for open-domain question answering." *arXiv preprint arXiv:1709.00023.*

27. 'Standard search API - Twitter Developers." (n.d.). *Twitter, Twitter, <https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets> (Dec. 1, 2019).*

28. Sood, Sara, Judd Antin, and Elizabeth Churchill. 2012. "Profanity Use in Online Communities." *Proceedings of the ACM Annual Conference on Human Factors in Computing Systems - CHI 12, 2012. https://doi.org/10.1145/2207676.2208610.*

29. Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, Animesh Mukherjee. 2018, "Opinion conflicts: An effective route to detect incivility in Twitter." *Proc. ACM Hum.-Comput.* Interact., 2(CSCW).

30. The Conversation. 2019. "The Conversation: In-depth analysis, research, news and ideas from leading academics and researchers." *[online] Available at: https://theconversation.com/ [Accessed 3 Dec. 2019].*

31. Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." *In Proceedings of the 11th International Conference on Web and Social Media (ICWSM), pages 512–515, Montreal, Quebec, Canada.*

32. Waseem, Zeerak, and Dirk Hovy. 2016. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." *Proceedings of the NAACL Student Research Workshop, https://doi.org/10.18653/v1/n16-2013.*

33. Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. "Character-level convolutional networks for text classification." *In Advances in neural information processing systems, pages 649–657.*

34. Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, Rui Yan. 2018. "An ensemble of retrieval-based and generation-based human-computer conversation systems,"

Proceedings of the 27th International Joint Conference on Artificial Intelligence, July 13-19, Stockholm, Sweden

35. Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, Ming Zhang. 2016. "Two are Better than One: An Ensemble of Retrieval- and Generation-Based Dialog Systems." *arXiv preprint arXiv:1610.07149.*

36. Zhang, Justine, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. "Conversational Flow in Oxford-Style Debates." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016. https://doi.org/10.18653/v1/n16-1017.*

37. Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. "Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?" *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018. https://doi.org/10.18653/v1/p18-1205.*

38. Zhang, Ziqi, David Robinson, and Jonathan Tepper. 2018. "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network." *The Semantic Web Lecture Notes in Computer Science, 2018, 745–60. https://doi.org/10.1007/978-3-319-93417-4_48.*