

DATA 698

Capstone Project Mid-Term

October 30, 2022

Prepared for:

Prof. Dr. Nasrin Khansari
City University of New York, School of Professional Studies

Prepared by:

Sie Siong Wong
Mario Pena
Joseph Shi

Introduction

Many of us spend several hours per week in social media platforms such as Facebook, Twitter, and Instagram to name a few. Although there are many advantages that social media provides e.g. serving as the fifth estate of power, a place to express opinions and such, it has also become a platform for someone to spread hate, cyberbullying, harassments and so on. Messages of hatred and spread of violence could have a high potential for influencing and motivating someone to commit a crime. According to CNN, hate crimes in New York City have increased 76% compared to last year. Many social media companies such as Facebook, have started to control the contents being posted by their users. Because there is a big concern in our society that offensive speech can result in more crimes and eventually become a serious threat to social, political, and cultural stability, we see the need to find a solution.

Some may argue that the First Amendment allows individuals to express anything they want. Instead of limiting everyone's freedom of speech openly in social media platforms, there are possible alternatives to deal with this issue. For example, a platform can allow users to turn on or off a switch to hide possible sensitive contents. In order to do this, we need some kind of solution that is able to identify inappropriate speech, tag them, and warn users that their post may not be visible to everyone due to containing sensitive language.

We believe that offensive speech has a high potential for causing more violent crimes to happen in the city. To make our city a better place to live, we need to find a way not only to identify inappropriate speech in social media platforms and reduce the spread of such speech, but also when to deploy more law enforcement to patrol areas crowded with high numbers of offensive speech in the hope we interrupt crime prematurely. Our research hypothesis is that higher offensive tweets from an area have strong correlation to the number of crimes committed in that area and minority groups are mostly the target.

The challenging part for our research is to classify the tweets metadata into offensive speech. Defining what is offensive speech such as hate speech can be hard because it's constantly evolving and often dependent on context. Until today, artificial intelligence still struggles when it comes to identifying hate speech. But we've found that quite a few research scientists have tried different approaches to identifying offensive speech in social media text data. The more advanced techniques, deep learning methods such as gated recurrent unit (GRU), convolutional neural networks (CNN), gated convolutional recurrent - neural networks (GCR-NN), and combination of transfer learning and weak supervision, are so far giving the best results compared to conventional classification approaches such as logistic regression, naive Bayes, decision tree, random forest, and gradient boosting.

There are many different types of hate speech that target different groups and individuals, and research has been done to differentiate and define them accordingly. Some approaches to the classification of hate speech from Tweeter use the hashtag as a variable for predicting violence and hate message from the tweet. While other approaches for classification of hate speech on social media may also include the development of neural language models. Additionally, there are algorithms that have been created to detect hate speech in digital micro-environments, which facilitate and reduce analysis tasks undergone by law enforcement agencies and service providers. We have also come across research that found there may not be a correlation between social media messages and crime, and others that consider "weather" to be an important environmental factor that affects the occurrence of criminal incidents. There have been many different approaches tried to find a link between social media posts and crime, all with varying results, and what makes this research so interesting is that we may yet find different results that may contribute to this solution.

Literature Review

The increase of using social media platforms such as Twitter and Facebook as a safe harbor to spread hate contents, and to cyberbully or harass an individual or a minority group because of their identities such as race, religion, sexual orientation, etc., has promoted social disharmony in which may bring local crime rate to rise. (Williams et al., 2019) In this research article, a Weka tool was used to develop a machine learning classifier to annotate racism and religion and non-hateful tweets. Besides tweets and crime data, census data such as no qualifications, age, long-term unemployed, and black and minority ethnicity, were considered part of the study. Statistical models were built to study a temporal and spatial association. Random and fixed effects, Poisson regression models, and negative binomial models, which all results show consistent of positive association between hate speech in Twitter and offline racially and religiously discriminated crimes. Other research has also shown that online hate speech has a strong correlation with significant events such as terror attacks, political votes, etc (Hanes & Machin, 2013; Williams & Burnap, 2015). A recent example of this was the Christchurch extreme right terror attacker’s post on 8chan. Although there is no direct causal link between online hate speech and offline hate crime, hate speech is still part of the cumulative process for a hate formula from social status, political context and geographical perspective that bring harm to the physical world.

Nowadays, hate speech detection remains a big challenge for artificial intelligence. (Matsaki L., 2018) Because the definition of hate speech is constantly evolving and often hidden within context, AI can be fooled easily. AI can be fooled by inserting typos, adding extra alphabets, and removing spaces. An example of altered text to evade AI detection would be something like this: “MartiansAreDisgustingAndShouldBeKilled love.” Humans can understand this message but the machine learning algorithms have trouble identifying it. Facebook CEO Mark Zuckerberg testified before congress in 2018 and mentioned he was optimistic that in 10 years AI would be able to identify hate speech more accurate. (Shead S., 2020) Only 24 percent of hate speech was able to be detected according to Facebook’s chief technology officer, Mike Schroepfer. Billions of users consume the Facebook product and platform, a cutting-edge machine technology to accurately spot hate speech is a must-have, to protect its users from exposure to harmful content.

Many previous researchers have tried different machine learning approaches to best identify hate speech. (Lee et al., 2022) Stacked ensemble Gated Convolutional Recurrent- Neural Networks (GCR-NN), a deep learning approach was used to detect racist speech from a tweets dataset. Its 0.98 accuracy outperforms other conventional machine learning models such as Random Forest, K-Nearest Neighbors, SVM, etc. A large scale dataset of tweets related to racism from the world were collected, preprocessed to such detail as to remove stop words, annotated using the TextBlob into positive, negative and neutral score sentiments before performing Term Frequency - Inverse Document Frequency (TF-IDF), and Bag of Words (BoW) for machine learning models development. United States has the highest number of racist tweets, 50%, followed by United Kingdom, 31%. More than 53% of people age between 15-30 years old in the United States is exposed to online hate material (Hawdon et al., 2016). Both Logistic Regression and SVM performs better than other models using BoW features on average with a 0.97 accuracy score. For a fair comparison, few single deep learning models such as Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), were also implemented and optimized through hyper parameter tuning. (Crabb et al., 2019; Starosta A., 2019) Combination of Weak Supervision and Transfer Learning approach is useful for identifying which data lacks labeling. Weak supervision is a method for building a labeled training set from a large amount of unlabeled data. Transfer learning is a method where it reuses already existing pre-trained models for new tasks. It will use the training set created from weak supervision stage to build a new classifier model.

Research has found that there is a contradiction between freedom of speech and hate speech. Hate speech creates an environment that tests the limits of free speech, and it violates fundamental rights of a human being. On social networks, we can observe how other forms of violence such as terrorism, extremism and hate crimes, may each have their own space, victims and aggressors (Chetty & Alathur, 2018). Miró-Llinares and Rodriguez-Sala (2016) describe the existence of many different types of hate speech that target different groups and individuals. Research has been done to differentiate and define each type accordingly. It is important to keep in mind classification as it will help us group Tweets with similar speech together in one

category. Some of the approaches to the classification of hate speech from Tweeter used by [Miró-Llinares and Rodríguez-Sala \(2016\)](#) take into account the hashtags as a variable for predicting violence and hate message from the tweet. While some of their other approaches for classification of hate speech on social media also include the development of neural language models. It is also important to underline that the authors warn the development of classification algorithms may not allow for a deeper understanding of the different nature of violent communications. The recommendation they make is to pursue a thorough study of the categorization of different expressions that are evident in these types of messages.

The study conducted by [Miró-Llinares et al. \(2018\)](#) focuses on designing an algorithm to detect hate speech in digital microenvironments. Its purpose is to facilitate and reduce analysis tasks undergone by law enforcement agencies and service providers. The algorithm used in this specific article uses machine learning classification techniques such as Random Forests. Furthermore, the study demonstrated that not all variables in the data relating to anonymity and visibility of users are applicable in distinguishing hate speech in tweets' content, and that tweet metadata proved to be more efficient in the classification process than account metadata. Compared to similar studies that have applied different classification approaches, the results obtained by this study slightly outperform the others. The Random Forest model applied reached a F1-score of 0.92, highlighting the accuracy of the model on the dataset. Previous attempts from other studies have obtained F-measures of 0.77, 0.90 and 0.76 according to the literature in this specific study.

We have also found research that is centered on the idea that social media metadata is a valuable input in the analysis of opinions and sentiment. From activism to detecting road traffic, access to these data is essential in today's comprehensive analyses. Using data from 18 Spanish-speaking Latin American countries, research found that similarly to mass media, social media suffers from a strong bias towards violent or sexual crimes. Simple models such as a linear regression were used in the research by [Curiel et al. \(2020\)](#) and found that countries with higher number of murders, murder rate and fear of crime are more likely to have crime-related tweets. However, it mainly represents the fears that people have of crime that may overemphasize certain types of crimes that are not as common as one would think. It was also found that there may not be a correlation at all between social media messages and crime, but rather demonstrate the reflection of the level of the fear of crime.

As explained by [Chen et al. \(2015\)](#), it is possible for weather to also be used as a feature for modeling crime around Tweeter data. Their work is also focused on predicting crime in order to maximize the allocation of scarce police resources. Their paper noted the limitations of past studies, suggesting that weather is a significant factor and it has been proved there is a correlation between weather and criminal activity. It is thought that certain environmental factors such as weather, should be considered as it may affect the occurrence of criminal incidents. Such data containing information about theft density, Twitter data, weather and geolocation points, have been modeled using logistic regression with recommendations for future analysis using support vector machine ([Chen et al., 2015](#)). More advanced techniques to predicting crime from Tweeter posts include Artificial Neural Networks approach that have achieved average accuracies of 0.903 on testing dataset and 0.933 on the training dataset as those demonstrated by [Sandagiri et al. \(2021\)](#).

To develop and train a hate detection model, we'll need a dataset with each record labeled with either hate or non-hate. This article ([Alnazzawi, 2022](#)) developed a corpus which has hate related context, hate crime type, and the motivation behind the hate crime. This dataset is available in the [Kaggle](#) website and freely available to download. For our research, we could leverage this dataset in part of our annotation process to classify hate and non-hate tweets using the key words in the given hate related context. There were two steps in developing the dataset, corpus construction and annotation. The author used a Hashtag tracking tool "[Hashtagify](#)" to search for hate related hashtags which were later to be used in the TweetsScraper tool to collect tweets related to hateful contents. Nine years (2010-2019) of tweets data related to these hashtags were collected. These hashtags list contain hate crime, racist, racism, Islamophobia, Islamophobic, sexism, disability, transgender, antisemitism, misogyny, and disabled. This hashtags list was found to be similar to what FBI used for the hate crime classification. For the annotation part, the author used the [COGITO Tech \(LLC\)](#) service to annotate each hate related tweets with 3 types of hate crime (physical assault, verbal abuse, incitement to hatred) and 5 types of motivation (racism, religion, disability, sexism, unknown) behind committing the crime. More than 60% of the 23,179 tweets, which hashtags' are hate crime related, are made up of physical assault and most of these physical assault crimes were motivated because of different

races and ethnicities.

Methodology

There are two sets of data we collect for this study, crime data and tweets data. The date range for these data is from year January 2019 to December 2021. We may extend the data collection a year back if needed for our analysis.

Data Collection

Crime Data:

The crime data was collected from the NYC Open Data source ([NYPD Complaint Data Historic](#)) by using the available API. The `read.socrata()` function available from the “RSocrata” library was used to send SQL query command through API to the site to get NYC crime data. The return data from the query has about 1.31 millions of crime data. Below are the variables and its values collected.

- `cmplnt_fr_dt` : Date of incident occurred
- `addr_pct_cd` : Precinct of incident occurred
- `ofns_desc` : Offense description
- `pd_desc` : Offense description (more granular)
- `boro_nm` : Borough name
- `susp_age_group` : Suspect’s age group
- `susp_race` : Suspect’s race
- `susp_sex` : Suspect’s sex
- `latitude` : Latitude coordinate
- `longitude` : Longitude coordinate
- `vic_age_group` : Victim’s age group
- `vic_race` : Victim’s race
- `vic_sex` : Victim’s sex

Tweets Data:

We applied for an academic research access developer account which has a higher tweets cap and got approved by [Twitter](#) for our research purpose to collect tweets data through API as well. The `get_all_tweets()` function available from “`academictwitteR`” library was used to query NYC tweets data. In the query, we use hashtags and keywords that are highly related to hate speech. We have collected about 1.2 million of tweets through API. From the data return, we select below variables which will be used for model building.

- `id` : Unique identification for each tweet
- `created_at` : Date the tweets created
- `text` : Text written by user
- `source` : Source used by user to tweet
- `possibly_sensitive` : True/False content is sensitive
- `type` : Type of coordinates
- `coordinates` : Longitude and latitude coordinates
- `retweet_count` : Count of retweet
- `reply_count` : Count of reply to the tweet
- `like_count` : Count of like to the tweet
- `quote_count` : Count of retweet with comments
- `longitude` : Longitude coordinate
- `latitude` : Latitude coordinate
- `borough` : Borough name

Data Analysis

We'll use a sentiment analysis approach to annotate tweets into positive (score>0), negative (score<0), and neutral (score=0) sentiments, to exclude non-hate related tweets. We consider tweets which have negative score are highly related to hate speech. Classification methods such as logistic regression, Naive Bayes, KNN, and SVM, will be used to develop hate speech detection models, by using a human-labeled twitter hate speech dataset taken from [Surge](#) website and then mix with some non-hate speech tweets. A best model is selected to label hate and non-hate for all identified negative attitude tweets. All labeled hateful tweets will then be joined to the crimes reported by using date and borough for spatial and temporal analysis. Both crime and hateful tweets count will be normalized using each place's population. A widely used regional statistics method, spatio-temporal Bayesian modeling, will be used to analyze spatio-temporal patterns, determine any developing trends, and test hypothesis. Below is the architecture of the proposed methodology.

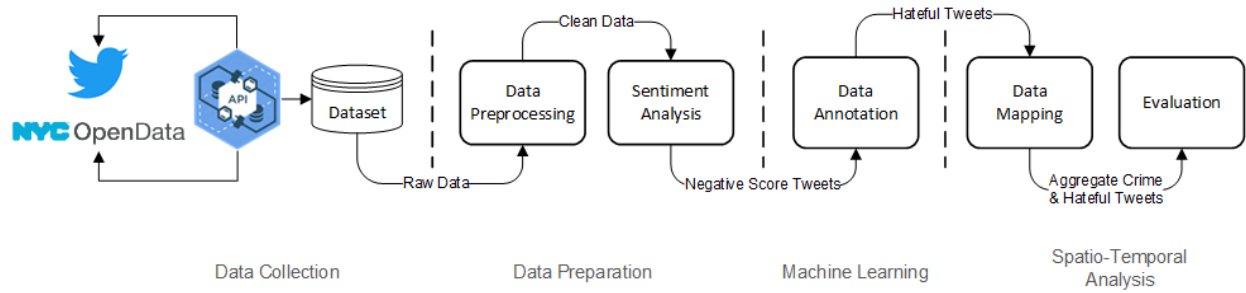


Figure 1: Architecture of the Proposed Methodology

Scope and Limitations

Our scope is to analyze the tweet's text created in the five boroughs of New York City: Manhattan, Brooklyn, Queens, Bronx, and Staten Island. But some tweets may have videos embedded and that's out of our capability to analyze the video's content. Only 32,224 tweets out of the 1.2 million tweets have geographic coordinates available; thus, we may not be able to fully perform spatial analysis at the zip code level. Another limitation to our study is that a trained binary classification model for labeling hundreds of thousands of negative attitude tweets into hateful or non-hateful tweets is unlikely to be 100% reliable compared to human moderation. With that being said, some of labeled hateful tweets could be simply expression of unhappiness, disagreement, sadness, disappointment, etc.

References

1. Williams et al. (2019, July 23). *Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime*. The British Journal of Criminology. Retrieved from <https://academic.oup.com/bjc/article/60/1/93/5537169>
2. Shead, S. (2020, November 19). *Facebook claims A.I. now detects 94.7% of the hate speech that gets removed from its platform*. CNBC. Retrieved from <https://www.cnbc.com/2020/11/19/facebook-says-ai-detects-94point7percent-of-hate-speech-removed-from-platform.html#:~:text=Facebook%20announced%20Thursday%20that%20artificial,and%20just%2024%25%20in%202017.>
3. Williams, M. & Burnap, P. (2015, June 25). *Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data*. British Journal of Criminology. Retrieved from <https://academic.oup.com/bjc/article/56/2/211/2462519>
4. Hanes, E. and Machin, S. (2013, September). *Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11*. Journal of Contemporary Criminal Justice. Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1038.9462&rep=rep1&type=pdf>
5. Matsaki L. (2018, September 26). *To Break a Hate-Speech Detection Algorithm, Try ‘Love’*. Wired. Retrieved from <https://www.wired.com/story/break-hate-speech-algorithm-try-love/>
6. Hawdon et al. (2016, July 13). *Exposure to Online Hate in Four Nations: A Cross-National Consideration*. ResearchGate. Retrieved from https://www.researchgate.net/publication/303480309_Exposure_to_Online_Hate_in_Four_Nations_A_Cross-National_Consideration
7. Lee et al. (2022, January). *Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model*. ResearchGate. Retrieved from https://www.researchgate.net/publication/357916429_Racism_Detection_by_Analyzing_Differential_Opinions_Through_Sentiment_Analysis_of_Tweets_Using_Stacked_Ensemble_GCR-NN_Model
8. Crabb et al. (2019, May 28). *Classifying Hate Speech: an overview*. Towards Data Science. Retrieved from <https://towardsdatascience.com/classifying-hate-speech-an-overview-d307356b9eba>
9. Starosta A. (2019, February 15). *CBuilding NLP Classifiers Cheaply With Transfer Learning and Weak Supervision*. Medium. Retrieved from <https://medium.com/sculpt/a-technique-for-building-nlp-classifiers-efficiently-with-transfer-learning-and-weak-supervision-a8e2f21ca9c8>
10. Chetty N. & Alathur S. (2018, May 8). *Hate speech review in the context of online social networks*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1359178917301064>
11. Miró-Llinares F. & Rodríguez-Sala J.J. (2016, July). *Cyber Hate Speech on Twitter: Analyzing Disruptive Events from Social Media to Build a Violent Communication and Hate Speech taxonomy*. ResearchGate. Retrieved from https://www.researchgate.net/publication/308487177_Cyber_hate_speech_on_twitter_Analyzing_disruptive_events_from_social_media_to_build_a_violent_communication_and_hate_speech_taxonomy
12. Miró-Llinares et al. (2018, November 15). *Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments*. Springer. Retrieved from: <https://link.springer.com/article/10.1186/s40163-018-0089-1>
13. Curiel et al. (2020, April 02). *Crime and its fear in social media*. Nature. Retrieved from <https://www.nature.com/articles/s41599-020-0430-7#Sec8>
14. Chen et al. (2015, June 8). *Crime prediction using Twitter sentiment and weather*. IEEE Xplore. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7117012/authors#authors>
15. Sandagiri et al. (2021, January 19). *Detecting Crime Related Twitter Posts using Artificial Neural Networks based Approach*. IEEE Xplore. Retrieved from <https://ieeexplore.ieee.org/document/9325485>

16. Alnazzawi, N. (2022, May 24). *Using Twitter to Detect Hate Crimes and Their Motivations: The HateMotiv Corpus*. MDPI. Retrieved from <https://www.mdpi.com/2306-5729/7/6/69>.
17. Hashtagify. Search And Find The Best Twitter Hashtags. Available online: <https://hashtagify.me/> (accessed on 15 March 2022).
18. Training Data for AI, ML with Human Empowered Automation. Cogit. Available online: <https://www.cogitotech.com/about-us> (accessed on 15 March 2022).
19. The HateMotiv Corpus. Kaggle. Available online: <https://www.kaggle.com/datasets/nohaalnazzawi/the-hatemotiv-corpus> (accessed on 3 October 2022).
20. NYPD Complaint Data Historic. NYC Open Data. Available online: <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i> (accessed on 17 October 2022).