Detection of Hate Speech on Twitter using AI Based Models

By: Ruba Othman

Course: 7CS084 Applying Artificial Intelligence

Tutor: Mr. Ahmed Khubaib

Table of Contents

Abstract

Online Social Networks have noticed a dramatic surge in user growth in the past years. As the number of people on these social media applications keep rising, it becomes increasingly difficult for humans to manually monitor user activity and if they are following guidelines of lawful behavior when it comes to the prohibition of hate speech. Artificial intelligence models offer a solution to this problem with the automatic detection of hate speech using machine learning and natural language processing techniques. This report will look into the literature of different models used for hate speech detection in Twitter, and will discuss the 'HaterNet' model in detail as an example of models used. The model has proved high accuracy and is currently the basis of many hate speech detection models nowadays. Future trends may explore how the model can better classify types of hate speech to aid in research and monitoring of cyberviolence.

1. Introduction:

Online social media networks or OSNs have grown tremendously in the past years. These platforms or applications were originally created to allow people to feel more connected in our expanding world. However, as technology evolved and users increased it has led to the misuse of these networks and put some people at risk for cyber-violence encouraged by specific prejudices.

A formal definition of hate speech by the UN Strategy and Plan of Action on Hate Speech is *"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor"* (Nations, 2021). As hate speech is usually motivated by agendas against a certain group of people, it becomes a topic of interest on applications that have a large number of users from different backgrounds where they are able to express their opinions or thoughts to a certain level of freedom.

Twitter is one of these applications that harbors millions of users and its estimated that it has an average of 500 million daily tweets (Shepherd, 2023). Though efforts have been made to combat offensive content on Twitter with the introduction of its hateful conduct policy, there is still much to be done when it comes to controlling online hate speech and preventing the harms it can have on society.

It is even more so difficult for humans to manually monitor online content on social networks due to their infinite size and continuous growth, hence artificial intelligence techniques offer the optimum solutions using automation and machine learning models that are able to analyse sentiments and detect hate speech to combat offensive behaviour and create safe spaces for everyone on social media.

Sentiment analysis using artificial intelligence is a widely researched topic. The use of natural language processing models in classifying the sentiment behind words or statements has been taken advantage of in a number of different applications. These include customer support chats, product research and social media monitoring. In addition, there have been numerous studies conducted in the specific area of online hate speech detection in social media applications such Facebook and Twitter. This research report will aim to focus on hate speech detection using AI on Twitter as that is the social media application where users tend to share their thoughts most due to their anonymity when compared to other social media applications (Burnap and Williams, 2015).

The report will also discuss as an example the popular model known as 'HaterNet' (Pereira-Kohatsu et al., 2019) that brought about a novel approach to hate detection on Twitter and is used as the basis of many hate detection models in the present day.

2. Problem Statement:

The risk of cyberviolence is fast-increasing as social media networks are growing. Though there are policies enforced to limit hate speech online, it is still almost impossible for humans to monitor all of the content on these social media applications. This report will look into how artificial intelligence models solve the problem of monitoring and detecting hate speech online, particularly on the social media application 'Twitter'.

3. Literature Review:

A study has been conducted by (Kouloumpis et al., 2011) to understand the sentiment behind the casual language used on Twitter through machine learning models. In this technique they have included hashtags as part of the training data to improve the performance in their model and it is considered one of the early approaches to sentiment analysis on Twitter.

Sentiment analysis is generally used understand whether a statement is negative, positive or even neutral. However, there is not enough clear definitions of what constitutes as hate speech on online social networks. In addition, there is also no uniform distribution of hate speech as hate speech in most cases erupts from a minority of biased individuals. These have put some constraints on natural language processing research of hate speech detection. (Waseem and Hovy, 2016) were able to interpret hateful speech in a corpus of 16,000 tweets and investigated features that would aid in the detection of hateful content, including demographics. They have also highlighted the words with highest indication of hate in a dictionary. The dataset has been used in many papers for hate speech detection using NLP.

A deep neural network was developed in (Zhang and Luo, 2019) study to efficiently draw out the features that can help in categorising hateful terms in tweets. The DNN has exceeded the performance of previous techniques by 5% of average F1.

The 'HaterNet' model presented by (Pereira-Kohatsu et al., 2019) combined LTSM+MLP techniques to detect hate speech on Twitter and they managed to obtain 0.828 AUC which when compared to studies in the past, it achieved better results. This same model is being implemented by the Spanish government to detect and monitor hate speech on Twitter.

A similar combination technique was also applied by (Lee et al., 2022) where they put together gated recurrent unit, convolutional neural network and recurrent neural network to form gate convolutional recurrent neural networks as their deep learning model that will identify racism in tweets and it attained an accuracy of 0.98.

(Yin and Zubiaga, 2021) however argued that models that are present nowadays do not generalise as well as expected on new datasets. In their research they mentioned that cross-dataset testing technique are a better fit to address this issue, however the study further elaborated on the problem of generalisation as it is closely connected to the construction of datasets, existing natural language processing models and how hate speech evolves with time. It was suggested that in further studies researchers should focus more on the techniques of gathering data and model building, as hate speech can come in various forms and contexts.

For example, hate-speech can sometimes be triggered by certain events. In a relevant study by (Tekumalla et al., 2022), it was noticed that tweets with negative sentiments have increased towards Asians after the outbreak of the Covid-19 pandemic. It particularly focused on attacks or negative comments that came from Americans in the United States towards American citizens with Asian descent. The research was able to characterise sinophobic slurs through the use of machine learning models.

Hate speech can also come in different languages, in (Roy et al., 2022) study an ensemble framework was used to detect hate speech in the Tamil and Malayalam language. An overall accuracy of 0.933 and 0.802 was achieved by their model.

4. Architecture:

The 'HaterNet' (Pereira-Kohatsu et al., 2019) model will be discussed in details as an example of what a hate speech detection model would generally involve from the process to features. The model has also presented a novel approach with the addition of the 'Social Network Analyzer'. Below Figure 1 shows the architecture of 'HaterNet'.

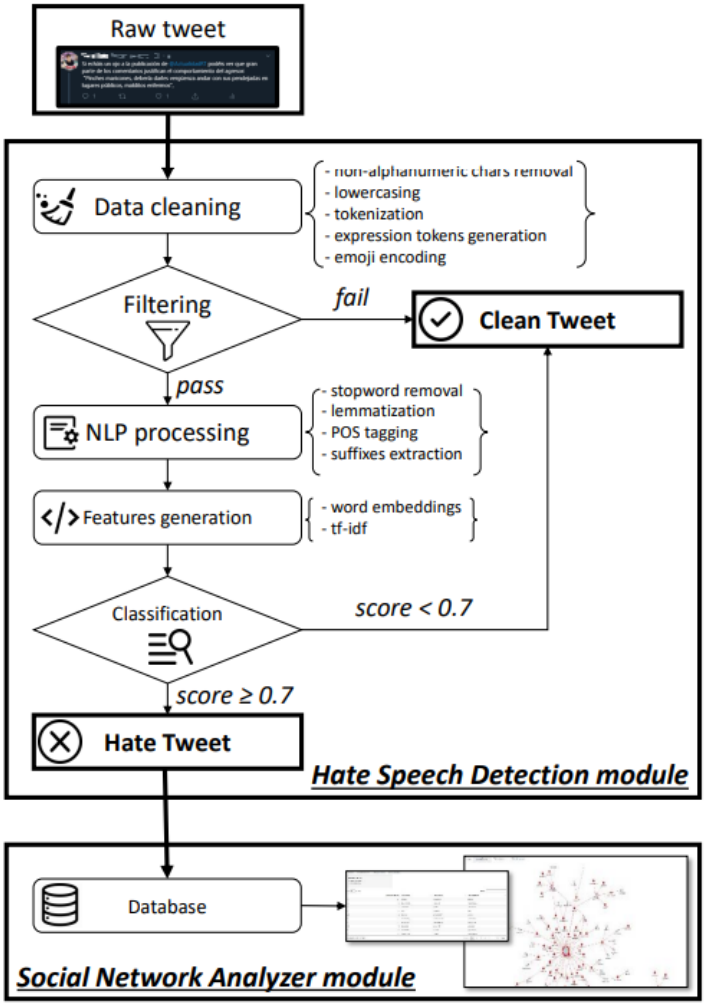

Figure 1. Overall architecture of HaterNet (Source: (Pereira-Kohatsu et al., 2019))

Twitter Rest API was utilized to gather different tweets from February 2017 to December 2017. The aim of collection and cleaning part was to gather all the latest tweets that were in the Spanish language or from Spain, through help of the UTC offset. The received text is transformed to lowercase and tokenized in order to steer clear of issues involving codification. Expressions that carry similar underlying meaning (anger and frustration, etc.) are grouped together to form a new expression. Any type of symbol that lacks value to the semantic analysis was removed. Emojis were kept and were also tokenized according to their meaning, since they also help with the semantic analysis process. There are rules that determine how certain symbols such as question marks will be differentiated from the remaining text found in a tweet. Below Figure 2 shows the table used for types of tokens for different semantics.

Semantic	Token Type
URL	TOKENURL
Mention	USER
Hashtag	HASHTAG
Question mark	TOKENQUES
Exclamation mark	TOKEXC
Laughing face: XD	TOKENXD
Quotation marks	TOKENCOMI
Laughs: jaja, ajaj, jajaj	TOKENLAUGH
Surprise: WTF, wtf	TOKENWTF

Figure 2. Table of semantics and token types (Source: (Pereira-Kohatsu et al., 2019))

Before the process of labelling is started, the data consisting of two million tweets is filtered. There are seven categories of hate speech through which the data gets filtered through. (Gender, race, religion, politics, disability and ethnicity and generic insults). There are two stages to the data filtering process, the first stage checks whether the tweet has words with absolute hate and the second stage checks for words with relative hate. In the Spanish language the word "negro" also means the color Black, so when a tweet is found with this word, the context would have to be checked to determine if it is relative hate. However, words like "feminazi" are regarded as absolute hate. Once a tweet has one word with absolute hate present then it would be filtered through as hate speech and becomes part of the training set, but if a word with relative hate is present, the other words in the tweet are checked for context. If there are generic insults present in the tweet, it would only then get filtered as hate speech otherwise it is disregarded.

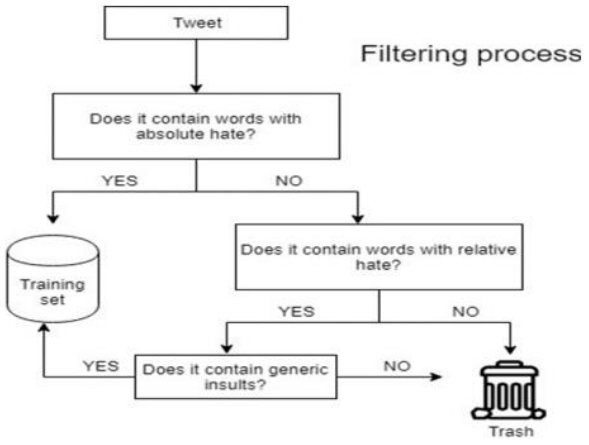

Figure 3. Filtering process of HaterNet (Source: (Pereira-Kohatsu et al., 2019))

The process of labeling in the ‘HaterNet’ model was done manually by four different types of professionals with various positions ranging in age and gender. The ruling opinion on whether a tweet was considered to be hate speech or not was decided based on the number of votes. If the number of votes was even, a fifth person would be consulted to determine the final ruling. Overall, only 26% of the filtered data was recognized as hate speech.

Natural Language Processing techniques are applied in the rendering stage which include POS tagging, lemmatization and stopword removal and results with tweets in a vector rendering format. To be able to identify the suffixes in the Spanish language, a special type of algorithm was created. Two forms of representation were also considered in the model, they are embedding-based and frequency-based. For the frequency-based representation there are four frequencies mapped to a single unigram including an additional four elements for sentences and word count. For the second type of representation, in order to allocate the semantic value for each word or expression found in a tweet, the word2vec model was applied. In an effort to expand the vocabulary of the embedding neural network, the unlabeled dataset was used in training to give a benefit in the case where HaterNet analyzer comes across a tweet with unlabeled terms.

Only in the case that frequency-based representation is utilized, feature selection would be implemented. The aim of this process is to rule-out vector elements that are scant or have minimal variation and if included in the classification stage would cause the model to overfit. A logistic regression model is then further implemented after the initial ruling-out stage using leave-one-out cross validation to finalise the overall number of features from 984 to 148.

The forms of classification models applied depend on the representation formats, the frequency-based representation will use a number of models that include SVM and Random Forest. The embedding-based representations follow a lengthier process since they are represented as matrices which would prevent a straightforward application of classification models due to the random unigrams number. The position of the words in tweets come with a high importance in hate speech detection so it is crucial to apply a classifier that would be used to comprehend and detect the order of words, for that reason a recurrent nueral network is applied. Typically for a tweet, the word limit is 33 and the tweets that are found with a lesser number are padded with zeros.

Below is the architecture of the neural network which consists of a two hidden layer multilayer perceptron. The position of words is sustained during the processing of the word embeddings which are then passed to the Tanh function. Each layer has a corresponding dropout probability which is used as a measure to determine whether or not there is hate speech, the tweet is considered to have hate speech if the probability measure is less than the output layer value.

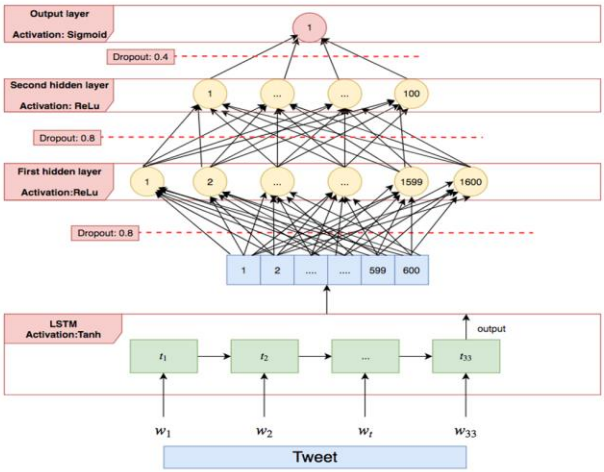

Figure 4. HaterNet Neural Network Architecture (Source: (Pereira-Kohatsu et al., 2019))

The last process in the architecture of the HaterNet model involves the "Social Network Analyzer" (Pereira-Kohatsu et al., 2019), which acts as an expert system that fetches certain type of data from the tweets to aid in the process of tracking and investigating them. The "Word Cloud Tab" (Pereira-Kohatsu et al., 2019) as the name indicates displays the most repeated terms with hateful content in a word cloud. Here the word embedding is made use of to show words that have similar semantics. The analyzer also uses a "User's Mentions Tab" (Pereira-Kohatsu et al., 2019) that observes how tweets with hateful terms spread across Twitter. It does so by creating a graph where hateful tweets with one person tagging the other are linked. This tab creates a way to realistically see how hateful content on Twitter is generated and reciprocated between users and if there is a certain group of people that are doing the attacking or being attacked. The "Terms Tab" (Pereira-Kohatsu et al., 2019) showcases the consistency of two or more words present in a tweet to see the relation between hate speech on Twitter and if specific occasions are a cause of a rise in them.

5. Suggestions and Improvements:

The are some suggestions on how to make the use of the model more specific, particularly in the feature extraction process. The tweets that are found to contain hate speech can be further categorised on the type of hate speech they hold such as if it is gender-based, racist or islamophobic. This type of classification can benefit competent organisations in their research or studies to combat cyberviolence.

The categorization of hate tweets could be even more advanced by adding a counter to the number of hate tweets sent for each classification. This would help in preventing hate crimes from happening by monitoring and observing if there is a surge in the number of hate tweets against a certain group after a specific event. The architecture could add another threshold and if the number of hate tweets exceed it then an alarm or trigger would be activated to notify police stations or high authorities of a possibility that a hate crime could occur in an area.

Besides improvements to the model, another area of research in hate speech detection that could be improved is in the real-time prevention of hate tweets being sent. As of today, Twitter does not include an automatic detection of hate terms before users send their tweets. For this

application it would be important to take into consideration the concurrency of words when typed, so as to check the whole context of a tweet before it is being prevented from being sent. Though it would be difficult to completely halt any form of hate speech in such platforms like Twitter where users are being granted their freedom, there are still some measures that can be taken to better enforce censorship.

6. Conclusion:

Hate speech detection on Twitter using artificial intelligence continues to be a topic of interest. Technology is advancing at a rapid pace and user growth is continuing on social media, it is important that artificial intelligence is made use of to maintain the safety of all on these applications. The model discussed was an example of what a standard hate detection model would include, with the overall architecture and features. Use of such models have shown high accuracy and more efficiency when it comes to controlling hate speech online. Future trends can explore more specific applications for hate speech detection for NGOs and government institutions in the research and monitoring of cyberviolence on social media.

References:

BURNAP, P. & WILLIAMS, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet,* 7**,** 223-242.

KOULOUMPIS, E., WILSON, T. & MOORE, J. Twitter sentiment analysis: The good the bad and the omg! Proceedings of the international AAAI conference on web and social media, 2011. 538-541.

LEE, E., RUSTAM, F., WASHINGTON, P. B., EL BARAKAZ, F., ALJEDAANI, W. & ASHRAF, I. 2022. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. *IEEE Access,* 10**,** 9717-9728.

NATIONS, U. 2021. *Understanding Hate Speech* [Online]. Available: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech [Accessed].

PEREIRA-KOHATSU, J. C., QUIJANO-SÁNCHEZ, L., LIBERATORE, F. & CAMACHO-COLLADOS, M. 2019. Detecting and monitoring hate speech in Twitter. *Sensors,* 19**,** 4654.

ROY, P. K., BHAWAL, S. & SUBALALITHA, C. N. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech & Language,* 75**,** 101386.

SHEPHERD, S. 2023. *22 Essential Twitter Statistics You Need to Know in 2023* [Online]. Available: https://thesocialshepherd.com/blog/twitter-statistics [Accessed].

TEKUMALLA, R., BAIG, Z., PAN, M., HERNANDEZ, L. A. R., WANG, M. & BANDA, J. Characterizing Anti-Asian Rhetoric During The COVID-19 Pandemic: A Sentiment Analysis Case Study on Twitter. Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media. Retrieved from https://doi. org/10.36190, 2022.

WASEEM, Z. & HOVY, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. Proceedings of the NAACL student research workshop, 2016. 88-93.

YIN, W. & ZUBIAGA, A. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science,* 7**,** e598.

ZHANG, Z. & LUO, L. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web,* 10**,** 925-945.