Effectiveness of Supervised Classification Models for Hate Speech on Twitter

Kunal Srivastava, Ryan Tabrizi, Ayaan Rahim

Abstract

The ceaseless connectivity imposed by the internet has made many vulnerable to offensive comments, be it their physical appearance, political beliefs, or religion. Some define hate speech as any kind of personal attack on one's identity or beliefs. Of the many sites that grant the ability to spread such offensive speech, Twitter has arguably become the primary medium for individuals and groups to spread these hurtful comments. Such comments typically fail to be detected by Twitter's anti-hate system and can linger online for hours before finally being taken down. Through sentiment analysis, our algorithm is able to distinguish hate speech effectively through the classification of sentiment.

1 Introduction

The United Nations defines hate speech as any kind of communication in speech, writing or behavior, 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, color, descent, gender or other identity factor. In an age where one can broadcast their beliefs to large sums of audiences through the internet, hate speech has inevitably become more ubiquitous and consequently leaves recipients with a decline in physical, emotional, and mental health [7]. Launched in 2006, Twitter has accumulated over 300 million users and has become the typical means to which people share their beliefs, particularly those regarding politics. Of these users, roughly 28 percent have experienced harassment. Twitter imposes a 280 character limit for each tweet, yet users can simply tweet multiple times to bypass the restriction. As of September 2020, Twitter has averaged 140 million tweets per day alone [1]. As of 2020, Twitter moderators must undergo training to identify what qualifies as hate speech as opposed to simple offensive comments [2]. With the aforementioned rate of 140 million tweets per day, the current anti-hate system is surely ineffective as there simply are not enough moderators to facilitate such hate speech prevention. Optimizing the automation of such an anti-hate speech system is quintessential to preventing such comments as the short-staffed moderation is no match for the overwhelming volume of tweets shared. The algorithm utilized a widely-used dataset from Kaggle containing 1.6 million tweets extracted using the Twitter API. The dataset contained six fields, specifically the sentiment score of the tweet, ranging from 0 to 4; the numerical id; the date of the post; the query; the username of the Tweeter; and the textual contents of the tweet. Similar works have addressed racist and sexist tweets, focused on hate trends, analyzed specific target groups for hate tweets, and have attempted to detect hate speech in multiple languages, but none have made the following effort of classifying hate speech into three categories. The developed algorithm can also be distinguished by its simple and efficient neural network, which makes training significantly less difficult, as well as its use of real time classification. The subsequent details of the findings are structured as follows. The related work section will provide findings that address similar topics and add further knowledge to the problem. Next, the method section will specifically explain the process used to analyze the data and create the neural network. The findings and insights section offers results and visuals to assess the findings of the topic. The discussion section brings up conflicts that arose during the research process, and how they were approached, as well as how we envision anti-hate Twitter to function. The conclusion provides a synopsis of the findings as well as concluding thoughts with regard to steps moving forward with the algorithm.

2 Related Work

In the past, there have been a few other studies pertaining to the reduction of hatred on Twitter. A lot of them have contributed their original data to the public, allowing for others to build upon their work.

[3] is an 8-month comprehensive study into the cross-behavior of abusive tweets and their effect on users. In this study, a dataset of 100,000 tweets was used. The work proposed an incremental and iterative methodology, attempting various approaches. Tools like matplotlib and seaborn were implemented in order to make clear and quantitative distinctions between abusive and non-abusive tweets. The originality of this work is that it examines a massive variety

of labeling schemes, covering different types of abusive behavior. This may range from sexual harassment, to racism, to outright hate speech. While it lacked efficiency and accuracy in classification, this study became a clear contribution to following works.

[4] is a study that implements a different approach to the problem of hate-speech on Twitter. This work is one centered around monitoring Twitter users, rather than the tweet's text itself. This implementation carries the advantage of preventing hate-speech from users entirely–making Twitter a much safer social network to be on. However, this approach carries the weakness in that it blocks out user reform–meaning that a theoretical user with 100 regular tweets and 5 hate-speech tweets would lose all 105 tweets. With this methodology, Twitter's popularity and usage will surely decrease due to the sheer number of users and hate on Twitter.

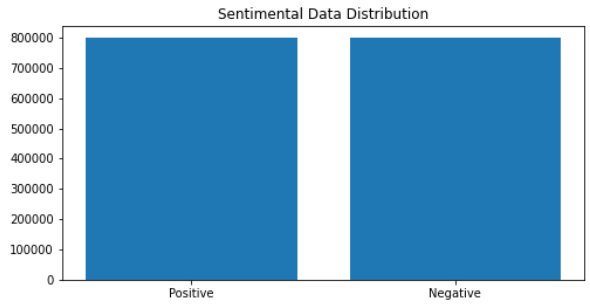
[5] pertains to the creation of an ontological classification model. However, the main and definitive contribution of this work is its new and original dataset of tweets that was created for analyzing the degrees of hateful Twitter text. The paper heavily stressed the importance of understanding the context of hate–that it was not just identifying keywords that distinguished hate speech, but taking the whole context into account. Taking the example "I hate black rhinos. One attacked us on a safari", a computer assigned to distinguish intent using only keywords would easily classify this as hate–while it is not. The paper correctly concluded that a much more comprehensive, context-understanding approach was needed.

[6] is a much more generic and simplified, yet meaningful, approach to the reduction of hate. The group conducting this research chose to not limit their data to Twitter, but also to another social networking platform called Whisper. The work encapsulated the skill of feature extraction into their methodology, effectively allowing them to analyze where and what topic hateful tweets fall into. In their findings report, they included a summary of the different groups of hate speech, as well as their intended targets (e.g. minorities, members of the LGBTQ+ community, and more).

3 Method

We aim to optimize classification via the implementation of original sentiment analysis algorithms through neural networks. The generic plan for the work is to create a strong, scaled, usable model for text classification on the Twitter architecture.

The first action item was data cleanup. Using list comprehension, each column was renamed into more descriptive keywords, rather than integer indexes. These included sentiment, id, date, query, userId, and text. Naturally, all that was needed for the process of classification was the sentiment and text features of the dataset. The extraneous columns were dropped. Next, the distribution of the data was measured. After utilizing visualization libraries to assess class counts, the data proved to be usable–approximately equally balanced among classes.



The next part of the process included text preprocessing–getting the text ready to be inputted to a custom neural network. The implementation of methods such as text stemming included removing useless suffixes from words (i.e. "adjustable" became "adjust"). In addition, text lemmatization (simplification of words in order for them to be more easily understood) was utilized. This included simplifying words like "better" to "good", or "running" to "run". The research group quickly realized that tweets often contained user mentions (containing the character '@') and hyperlinks. A scripting method to remove these characteristics (since they are useless to the purpose of sentiment analysis) was run. Finally, the algorithm effectively utilized the nltk library–a library containing different groups of words in english. One of these groups was entitled stopwords. These were words that were essentially useless to the purpose of understanding sentences. After effectively removing them, the text was fully clean and ready to be tokenized.

Tokenization is the process of splitting sentences into tokens, in order for them to be computationally understood. Unnecessary punctuation (i.e. periods, commas, and question marks) were also thrown away.

Word embedding is the process of vectorizing tokens in order for them to be used in natural language processing applications. Training embedding proved to be extremely computationally costly, and expensive time-wise. We decided to use transfer learning, and implemented GloVe Embedding from Stanford AI.

Finally, the data was fully ready for neural network implementation. The network architecture of Sequence Modeling

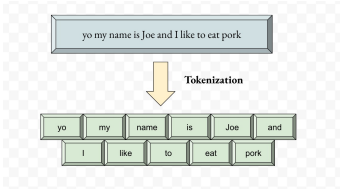


Figure 1: Process of Tokenization

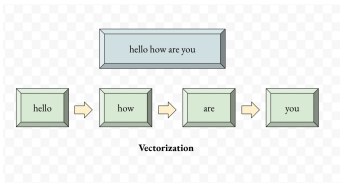


Figure 2: Process of Vectorization

was chosen, due to the fact that in this case, word vectors are either positive, neutral, negative, or hateful. Due to these clear distinctions, it was obvious that other architectures were not a good fit. The ternary classification proved to be difficult to train, so the team improvised by combining the labels of neutral and offensive tweets. After all, the purpose was to investigate the hateful tweets on twitter.

The model consisted of the following, in chronological order. An embedding layer, which generated an input vector to the network for each sentence. A convolutional 1-dimensional layer provided scaling of vectors into smaller feature vectors, enabling a higher accuracy. Next was the main layer: an LSTM, a variant of RNN (recurrent neural networks) that has the power to learn the context of words given those around it. Finally, a dense layer was implemented for the final stages of classification.

After training, the classification appeared to be quite effective. Overall, we are fairly content with the research design. The one major change that was made to the original research process was condensing the neutral and offensive tweets in order to make binary classification possible. The team is happy with the planning and execution of the research process.

4 Findings and Insights

After training, several pythonic tools were utilized in order to assess the performance of the neural network. The first step was to visualize some of the words assessed in the neural network. The python library wordcloud was implemented in order to provide a nice and clean visual representation of the top used words both in hateful tweets and non-hateful tweets. For academic appropriateness, only the neutral and negative tweet keywords are shown below, excluding the hateful keywords.

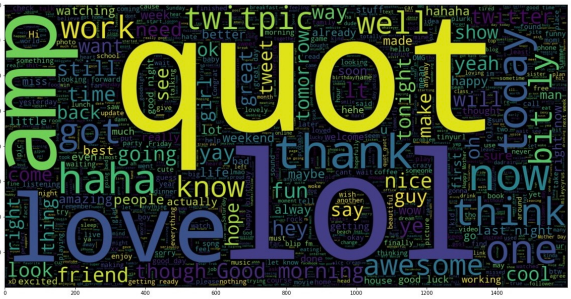
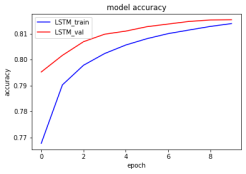
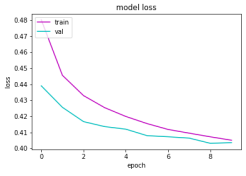


Figure 3: Popular Neutral and Negative Keywords in Tweets

The next step was assessing the model, as well as hyperparameter tuning. In order to optimize the accuracy of the model, as well as minimize the loss, the implementation of two callbacks were needed. Callbacks are special functions, built into model libraries, that are called after the end of an epoch (training loop) to perform specific operations. The first used was LRScheduler. This callback significantly reduced the learning rate (how fast the model is learning) exponentially, only if there was a plateau in the accuracy's increase. This essentially means that if the accuracy was not growing (signs of overfitting), the learning rate was reduced, in order to maximize the accuracy of the model. The next and final callback used was ModelCheckPoint. This callback essentially saves the best model achieved during training, making it easy to save record performances. In this case, we saved the model with the minimum validity loss.

The final model was ready to be assessed. Below are figures representing the training and loss processes.





A confusion matrix is a clean way to represent the overall and final classification of results. Alongside with the classification report, both of these visual aids provide simplistic approaches to viewing final statistics and metrics regarding the model's performance.

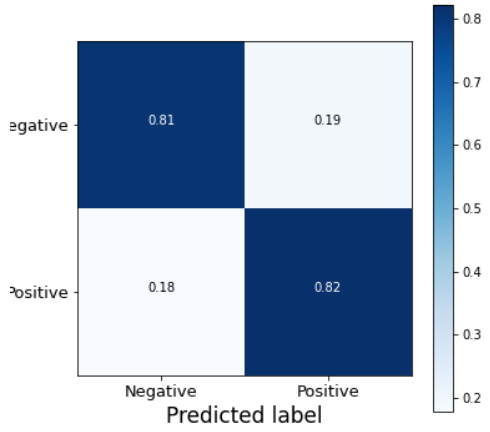


Figure 4: Confusion Matrix

	precision	recall	f1-score	support
Negative	0.82	0.81	0.82	160542
Positive	0.81	0.82	0.82	159458
accuracy			0.82	320000
macro avg	0.82	0.82	0.82	320000
weighted avg	0.82	0.82	0.82	320000

Figure 5: Classification Report

In summary, an accuracy of 82 percent was achieved. This is actually a very meaningful performance, and is definitely higher than the human's accuracy at recognizing hate speech. After all, humans are the ones who are creating hate speech, and thus some may be more prone to accepting it and thinking it is normal.

5 Discussion

Throughout the development of the algorithm, several conflicts arose that ultimately led to finding alternative paths towards completion of the system. Perhaps the most pressing issue was the nature of ternary classification with regard to the accuracy of the system. As mentioned earlier, there are three technical classifications regarding the offensivity of a message: non offensive, offensive, and hate speech. Upon completing the algorithm and sampling, it came to attention that ternary classification, as opposed to binary classification (hate speech vs non hate speech), produced significantly less accurate and efficient classification, prompting a change to binary classification. Once fitted towards binary classification, the algorithm demonstrated significantly higher efficiency and accuracy and thus became the primary approach towards classification. As mentioned earlier, another obstacle with the algorithm was getting it to classify whether a phrase would be considered hate speech within the context of the sentence as a whole.

The algorithm aims to promote anti-hate on Twitter to enable a platform where people can freely interact without receiving attacks towards their personal beliefs and identities. By doing so, users can secure their physical, emotional, and mental health while still enjoying the fruits of Twitter and ultimately all online platforms.

6 Conclusion

Many have become vulnerable to attacks on their identities, whether their physical appearance, political views, or religion, due to the ceaseless access enforced by the Internet. Twitter has arguably become the main outlet for individuals and organizations to distribute these hurtful messages from the many outlets that have the opportunity to share such offensive expression. Usually, such comments fail to be identified by the anti-hate mechanism of Twitter and can stay online for hours until they are eventually taken down. In conclusion, this work is a step forward into a dream Twitter, allowing users to enjoy one of the biggest social media networks happily and safely.

References

[1] *numbers*. URL: https://blog.twitter.com/official/en_us/a/2011/numbers.html.

[2] Kate Conger. *Twitter Backs Off Broad Limits on 'Dehumanizing' Speech*. July 2019. URL: https://www.nytimes.com/2019/07/09/technology/twitter-ban-speech-dehumanizing.html.

[3] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior". In: *CoRR* abs/1802.00393 (2018). arXiv: 1802.00393. URL: http://arxiv.org/abs/1802.00393.

[4] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgılio A. F. Almeida, and Wagner Meira Jr. "Characterizing and Detecting Hateful Users on Twitter". In: *CoRR* abs/1803.08977 (2018). arXiv: 1803.08977. URL: http://arxiv.org/abs/1803.08977.

[5] Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. "Degree based Classification of Harmful Speech using Twitter Data". In: *CoRR* abs/1806.04197 (2018). arXiv: 1806.04197. URL: http://arxiv.org/abs/1806.04197.

[6] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrıcio Benevenuto, and Ingmar Weber. "Analyzing the Targets of Hate in Online Social Media". In: *CoRR* abs/1603.07709 (2016). arXiv: 1603.07709. URL: http://arxiv.org/abs/1603.07709.

[7] *United Nations Office on Genocide Prevention and the Responsibility to Protect*. URL: https://www.un.org/en/genocideprevention/.