Use of Natural Language Processing Techniques for Detection of Cyberbullying Tweets

Prathamesh Kulkarni (Student ID: 36004026)
Dept. of Computer Science and Communication
Lancaster University, Lancaster, England
p.p.kulkarni@lancaster.ac.uk

Abstract

As people can be anonymous and or virtual on social media platforms like Twitter, cyberbullying has become much easier. There are many types of cyberbullying, depending on age, gender, ethnicity, and religion. There should be a system in place to deal with this issue, so it is easier to identify and flag such tweets. This paper attempts to develop models that can predict whether a tweet is a cyberbullying activity or not. According to the study, Random Forest, LightGBM, and Bidirectional LSTM had higher accuracy scores of 94 percent in classifying tweets.

1 Introduction

The term cyberbullying refers to bullying that occurs over digital devices such as a cell phone, computer, or tablet. Several social networks, such as Facebook, Twitter, Instagram, WhatsApp, etc., can be used to send emails and messages. Where a group of people are connected and sharing social information. By sharing abused photos, videos, or negative information about others, some people are abusing these facilities and bullying others. It poses a serious threat to society and needs to be combated and found. Cyberbullying consists of a series of communications, and it often involves a death threat or a threat to harm another individual (Subhendu Kumar Pani, Sanjay Kumar Singh, Lalit Garg, Ram Bilas Pachori, Xiaobo Zhang, 2021). Cyber bullying can often be a two-way street when both sides have access to technology, resulting in individuals reversing roles and becoming cyber bullies themselves. In extreme cases, cyberbullying can lead to mental health issues or suicide in its victims, just as real bullying can. Social media companies have to walk a fine line in protecting their users from harassment and bullying while restricting their freedom of speech (Carl Timm, Richard Perez, 2010).

Using natural language-based machine learning models, it is able to learn and adapt to the way humans communicate in text to become increasingly good at understanding slang, patterns of speech, regional language variations and turns-of-phrase (Bernard Marr, Matt Ward, 2019). Tweets are parsed by the models, and anything that is determined to be an abusive comment – insults about someone's race, gender, ethnicity, religion for example – is automatically classified. In this paper classical machine learning algorithms along with neural networks are implemented to compare the performance of models in classifying tweets based on their cyberbullying type.

2 Related Work

Pre-processing data is described in paper (Haddia et al., 2013). This might include removing hyperlinks, removing spaces, or transforming the data into feature matrices. The paper attempts to clean and convert the tweets in a way that would facilitate further analysis and modelling. A LSTM model is used to predict polarities in tweets in Paper (Wang et al., 2015). As a means of predicting the type of cyberbullying, this paper uses LSTMs to test the accuracy of the model. The papers (Hani et al., 2019) and (Qaiser et al., 2018) explain how TFIDF works and how it is applied to cyberbullying detection. It was for this reason that the TFIDF vectorizer was employed. A neural network and a support vector machine were used to detect cyberbullying in paper (Desai et al., 2021). They achieved an accuracy of 92 percent. This paper investigates several algorithms to determine which is more appropriate for the data and tries to expand on that. They obtained F1-scores of 64 percent for binary classification datasets in paper (Hee et al., 2018) which identifies social media texts that may be related to cyber-

bullying using SVM algorithm. In contrast, the current study tries to not only classify tweets based on whether or not they are cyberbullying cases, but also decide what type of cyberbullying they are, making it a multiclass classification problem.

3 Data

The data was obtained from kaggle. This dataset contains more than 47000 tweets labelled according to the class of cyberbullying:

1. Age
2. Ethnicity
3. Gender
4. Religion
5. Other types of cyberbullying
6. Not cyberbullying

The data has been balanced in order to contain 8000 of each class.

4 Methodology

4.1 Pre-Processing

Text data requires a lot of pre-processing before we can start analysing and modelling it. As machine learning models require appropriate data to be fed, i.e. numerical data, converting the tweets to a numerical format becomes vital. To analyse the text data, it was first cleaned by removing links, non-UTF8/ASCII characters, multiple spaces between words, punctuation marks, and numbers. The result is clean text. Duplicate tweets were later removed. Data for class "other_cyberbullying" was removed since keeping it resulted in poor model performance. All of the classes, such as "not_cyberbullying", "gender", "religion", "age", and "ethnicity", were encoded with 0 to 4 respectively. For neural network implementation, one-hot encoding is used. Finally, the data were split 80 per cent as training data and 20 per cent as testing data.

4.2 Tokenization

As part of preprocessing, the text is converted to the numerical format. The first step is to tokenize and then create a document term matrix (DTM). Tokenization is the process of breaking up large texts into smaller chunks. Tokens can consist of a word, a character, or a subword. Tokenization always fits on training data. In order to make all the sequences the same length, padding is applied. Usually, the longest sequence determines the length.

To obtain DTM based on tokens, we use the tfidf vectorizer and Keras tokenizer. TF-IDF represents Term Frequency Inverse Document Frequency. This method is used by machine algorithms for converting texts into meaningful representations of numbers that are then used for prediction. In the TF-IDF, the frequency of words in documents is compared with the inverse proportion of words in documents as a whole. Basically, this calculation determines the relevance of a given word in a document (Sahay et al., 2018) (Chaudhary, 2020). Tokenizing a text corpus is achieved using Keras' Tokenizer class. The text input will either be converted into an integer sequence or into a vector consisting of binary values corresponding to each token. TFIDF vectorizer and Keras tokenizer libraries perform tokenization and return DTMs.

4.3 Modelling

This data set was tokenized with TFIDF and Keras because they performed better. The TFIDF tokenizer was used for modelling classical machine learning algorithms. Moreover, Keras tokenizer was used for modelling neural networks.

This study looked at which algorithm would be the best at predicting whether a tweet is a cyberbullying act or not based on the data. It is a multiclass classification problem with 5 classes.

Initially, simple logistic regression was used to assess the performance of the simplest classification algorithm. Later on, classifiers such as random forest, gradient boosting, lightgbm, xgboost, Naive Bayes, and Support Vector Machines are implemented. (Sharma et al., 2018)

To measure how accurate each algorithm was, the algorithms were implemented without any parameters. Random Forest and LightGBM provided the most accurate predictions. Then, both of them were optimised using a grid search to examine if the performance of each one changed with optimization. A number of parameters, such as the learning rate, max_depth, and the GINI or entropy to calculate impurity, were tested in order to determine the best parameters. The accuracy, precision, recall, F1-score, and confusion matrix of these models were then evaluated to determine whether the predictions were correct.

The cleaned data was used with the Keras tok-

enizer to model neural networks. It was padded based on the maximum tweet length. An LSTM model was created based on the data since LSTMs are better models for natural language processing. It is composed of three layers: Embedding, Bidirectional LSTM, and Dense with 5 neurons representing the number of categories. Since there are multiple classes in the data, categorical cross-entropy is considered a loss function. Adam was selected as the optimizer. In light of a large amount of data and the limited resources, the model was trained over five epochs. The model performed almost as well as Random Forest and LightGBM algorithms.

5 Results

In order to analyze the data, things such as the number of data from a certain class, the tweet length, and the most frequent words were looked at.

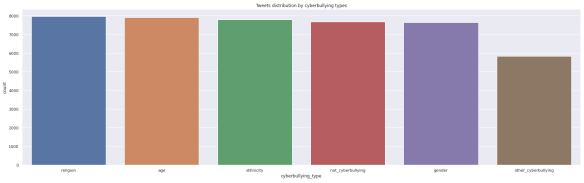

Figure 1: Number of tweet per cyberbullying type

It is clear from the data that the classes, i.e., different types of cyberbullying, are balanced. The plot shows that all classes have approximately 8000 data points associated with them. As an example, ethnicity has around 8000 tweets associated with it. This has the advantage of making the models unbiased since they will get an equal number of samples for each class as a result of the balanced data.

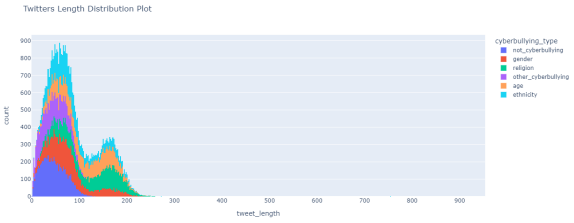

Figure 2: Tweet Length vs. Count Histogram

According to the tweet length plot, the majority of tweets in all classes have a length between 50 and 70 characters. The second peak can be seen between 150-180 tweet-length.

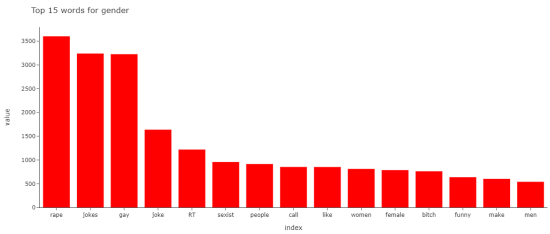

Figure 3: Top 15 words with gender as cyberbullying type

In order to identify the most frequently used words for each type of cyberbullying, a word frequency analysis was conducted. The plot above illustrates the 15 most common words associated with the class "gender". It is clear that words like "rape", "gay", "women", "sexist", and "bitch" appear frequently, which makes sense as they are associated with gender-based cyberbullying.

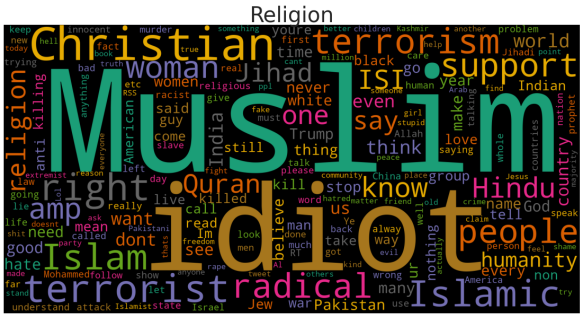

Figure 4: Word Cloud for Religion based cyberbullying

Likewise, the word cloud above shows that "Muslim", "Christian", "Islamic", "terrorism" and "radical" are associated with religion-based cyberbullying.

To predict the type of cyberbullying various algorithms were implemented.

Table 1: Performance of various models in terms of accuracy

Model	Accuracy
Random Forest	0.940626
LightGBM	0.940498
Gradient Boosting	0.935753
Support Vector Machine	0.934599
Logistic Regression	0.93
XGBoost	0.928315
Multilayer Perceptron	0.905104
Naive Bayes	0.855604

Table 1 shows the implemented algorithms and the accuracies of each of them.

The top-performing models were then trained separately to evaluate their overall performance of the models.

Table 2: Performance metrics of top performing models.

Metric	LightGBM	Random Forest
Accuracy	0.93	0.94
Precision	0.94	0.94
Recall	0.93	0.94
F1 Score	0.93	0.94

Table 2 shows the performance of the each of the top-performing models which are random forest and lightgbm.

From the confusion matrix of the random forest, it can be seen how the model performed in predicting each of the classes. It can be observed that both random forest and lightgbm had a good prediction overall but were better at predicting the “ethnicity” class.

The neural network gave an accuracy of 0.93 which is par with random forest and lightgbm. Though there are some limitations to this model as the training was performed for only 5 epochs hence the performance of the model over many epochs cant be predicted. But the aim of implementation was to check whether the neural network model would perform better than traditional algorithms, it doesn’t seem to be the case.

The confusion matrix of the neural network shows that it has a lot more predictive power with class “gender” but performed poorly on classes “religion”, and “age”.

6 Limitations and Future Work

This study has some limitations. There are probably more efficient and better ways of cleaning the data. For example, taking emojis into account. Tokenization techniques can be varied and used in specific ways to improve the understanding of text and modelling. Various parameters can be changed and experimented with to find even more optimal parameters that would lead to a more accurate model. It might be possible to improve performance by adding more layers or changing the loss function and optimizers in neural networks.

Perhaps future work will involve trying to figure out the exact meaning of the tweets and whether or not they should be flagged. A better prediction model can be developed. The extraction of text from an image can also assist in understanding cyberbullying types, since images can be used as a form of cyberbullying.

7 Conclusion

The purpose of the paper was to analyse and create models to predict whether or not a tweet is cyberbullying. It has been implemented successfully in this paper. Several techniques to process the textual data have been discussed, from cleaning the data to analyzing and modelling.

From the data, we can conclude that we can determine some key information on what a tweet is conveying from the tone and words used. When it comes to cyberbullying, it becomes increasingly important to determine what kind of bullying is happening. Words such as "bully", "dumb", "nigger", "Muslim", "rape," etc., are some examples of words that can indicate whether a tweet is related to age, gender, religion or ethnicity types of cyberbullying. In the various models discussed in the paper, random forest, lightGBM, and neural networks performed better at predicting the types of cyberbullying.

Social media platforms can use these models to identify or flag tweets as well as accounts that are frequently involved in cyberbullying. It can be used to take action against their accounts and, in severe cases, against the people.

8 Acknowledgements

I would like to acknowledge

1. The creators of the dataset (J. Wang, 2020) and the various libraries used.

2. The professor for providing the various steps involved in conducting the experiments.

3. Stefan Rares Niculae for providing the Code for pre-processing.

References

Bernard Marr, Matt Ward. 2019. *Artificial Intelligence in Practice*. Wiley.

Carl Timm, Richard Perez. 2010. *Seven Deadliest Social Network Attacks*. Syngress.

Mukesh Chaudhary. 2020. Tf-idf vectorizer scikit-learn. *https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a*.

Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, and Rashmi Dhumal. 2021. Cyber bullying detection on social media using machine learning. *International Conference on Automation, Computing and Communication*, 40.

Emma Haddia, Xiaohui Liua, and Yong Shib. 2013. The role of text pre-processing in sentiment analysis. *International Conference on Information Technology and Quantitative Management*.

John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Ammar Mohammed. 2019. Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications*, 10(5).

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *arXiv:1801.05617*.

C.T. Lu J. Wang, K. Fu. 2020. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. *Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020)*.

Qaiser, Shahzad, Ali, and Ramsha. 2018. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181.

Kshitiz Sahay, Harsimran Singh Khaira, Prince Kukreja, and Nishchay Shukla. 2018. Detecting cyberbullying and aggression in social commentary using nlp and machine learning. *International Journal of Engineering Technology Science and Research*, 5.

H. Kumar Sharma, K. Kshitiz, and Shailendra. 2018. Nlp and machine learning techniques for detecting insulting comments on social networking platforms. *International Conference on Advances in Computing and Communication Engineering (ICACCE)*.

Subhendu Kumar Pani, Sanjay Kumar Singh, Lalit Garg, Ram Bilas Pachori, Xiaobo Zhang. 2021. *Intelligent Data Analytics for Terror Threat Prediction*. Wiley-Scrivener.

Xin Wang, Yuanchao Liu, Chengjie Sun, Baoxun Wang, and Xiaolong Wang. 2015. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1343–1353, Beijing, China. Association for Computational Linguistics.