IMT 575

APPLYING MACHINE LEARNING ALGORITHMS TO DETECT CYBERBULLYING

By Chinmay Yalameli, Nikhil GB, Pallavi Banerjee, Shouyang Wang

TEAM QUATTRO

Table of Contents

MOTIVATION

The motivation for our project is simple and straightforward - we wanted to work in a socially relevant space such that the results of our work can have an impact that improves the current state of the society, in any particular aspect.

Machine Learning has great potential to address several socially relevant issues. There is a lot of ongoing work and research happening in the areas of criminal justice. Many ethical issues of bias in ML models are also being addressed by newer age ML algorithms. In this spirit of such groundbreaking and impactful work, we as a team wanted to work in an equally socially important area. We also feel it is our duty to leverage technology to try and understand the cause, impact, and severity of the several ill manifestations of technology and innovation. This made us ponder about what kind of technological progress is omnipresent but also has hidden negative impacts.

One particular problem on the rise over the past few years is Cyberbullying. To cite some obvious and glaring instances of cyberbullying, comments and opinions posted on various social networking websites like Facebook, Instagram, Twitter, Reddit target individuals of all gender, race, ethnicity, and age, etc. Activist Greta Thunberg was cyberbullied by Trump via Twitter, school children are body-shamed by their classmates, and such examples are never-ending.

With the global pandemic where everything moved online, the amount of cyberbullying and harassment has only ever increased. Given this context, we felt that the space for tackling cyberbullying is not just socially relevant but quite a need of the hour.

QUESTION

While narrowing down our scope and formulating a question, we took inspiration from two major works in this field-- Hinduja et al. and Wang et al. The former argue how there is an immediate need to address cyberbullying but establishing a link between cyberbullying and suicide, while the former works on applying the more structured Machine Learning model to a dataset comprising of labeled tweet instances of cyberbullying.

Given the fact that cyberbullying is not limited to one particular area, it is important to be able to recognize and differentiate not only between cyberbullying and not cyberbullying but also to identify the type of cyberbullying.

Given this premise, we chose to direct our efforts to answer this one question in our project:

> ***How do we accurately predict the label of a tweet as one of the pre-defined fine-grained categories related to cyberbullying?***

If we are able to understand and predict the type of cyberbullying that occurs, we can better understand which direction society is moving as a whole, and which human biases and/or prejudices are most visited. This helps direct policy measures, amend laws, and overall implement correctional measures.

LITERATURE REVIEW

Bullying, Cyberbullying, and Suicide

In this paper, the author Hinduja et al. explain to what extent victimization and cyberbullying relate to suicidal ideation among adolescents. The author examined a study conducted in 2007 for a random sample of 1,963 middle-schoolers and concluded that students who experienced traditional bullying or cyberbullying had more inclination towards suicidal thoughts. The author also concluded that victimization was more strongly correlated to suicidal thoughts than offending **(Hinduja, 2010).** The author utilized various machine learning libraries to conduct analysis. From logistic regression analysis (const = -2.58, Exp **β** -0.79), it was concluded that

> *"Traditional bullying victims or students were 1.7 times more likely, and traditional bullying offenders were 2.1 times more likely to have attempted suicide compared to those who were not traditional victims or offenders. Similarly, cyberbullying victims were 1.9 times more likely, and cyberbullying offenders were 1.5 times more likely to have attempted suicide than those who were not cyberbullying victims or offenders (Hinduja, 2010)."*

This paper is a strong foundation against cyberbullying as the author defends all the premises with quantitative research and analyses. It was one of the papers which acted as motivation for us to pursue our topic for this research, and it helped us identify the initial results for our study. It also outlined the context for quantitative analytics used in our work.

SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection

This paper contributed a significant amount of research and changed the perspective of utilizing artificial intelligence to combat cyberbullying. The author investigated the possibility of utilizing an automatic multiclass cyberbullying detection mode which was able to classify if the cyberbully was targeting the victim's ethnicity, gender, age, religion, or other qualities **(Wang & Fu, 2020).** The biggest problem the previous researchers faced, especially dealing with extensive cyberbullying data, was the class imbalance. The popular methods used in handling class imbalance problems, i.e., Upsampling, Downsampling, and SMOTE, had drawbacks. They both are sensitive to outliers, and they reduce the separation between classes. The authors utilized Dynamic Query Expansion (DQE) to combat this problem as it helped process and extract more raw data points of a specific class, and it worked very well with sparse matrices **(Wang & Fu, 2020).** Thus the combination of Dynamic Query Expansion (DQE) along with Graph Convolutional Network(GCN) represented a new path to solve these problems. The author utilized various text embedding techniques ranging from simple Bag of Words, Tf-Idf , BERT, Glove, and fasttext to advanced Sentence BERT methods. The author also implemented various machine learning algorithms, namely logistic regression, K nearest neighbors (KNN)(LR), Naive Bayes (NB), support vector machine (SVM), XGBoost (XGB), and multi-layer perceptron (MLP), The highest accuracy of 94.38% was achieved from XGBoost algorithm with Bag of words embeddings. The tweet set containing 47000 tweets was made available to the public in Kaggle, which we are using as the data set for our project detailed below.

The above two papers laid the foundation for the context and direction of our research. More data and motivation were gathered from stopbullying.gov, which states that students who experience cyberbullying are twice as likely to attempt or commit suicide. More than 65% of parents agree that cyberbullying is one of their primary concerns in today's generation. Thus with the series of articles, literature reviews, and context, we decided to approach this topic with novice methods explained below.

DATA AND ANALYSIS

In this study, we leverage a dataset (Wang et al., 2020) containing labeled textual data. The dataset aggregates several previous datasets that studied online harassment and cyberbullying. There are 47,000 tweets within the dataset, and the tweets are divided evenly based on the category of cyberbullying. Each tweet is classified as a specific type of cyberbullying, including ethnicity, religion, gender, age, other types of bullying, or neutral (i.e., not cyberbullying). There are approximately 8000 tweets within each category. The data was first introduced at the 2020 IEEE International Conference on Big Data and made available on Kaggle for public access.

Data Cleaning and Feature Selection

Twitter data is known for its messiness. Removing stop words alone is often not sufficient for social media mining. Before data modeling, we removed URLs, Twitter mentions, emojis, numbers, reserved Twitter words (i.e., RT – "retweet"), punctuations, and HTML encoding characters.

We also explored Part-of-Speech tagging techniques, such as filtering out nouns and verbs, for feature selection purposes. This was not implemented in the final analysis to prevent over-cleaning of data. We observed that the models generally perform worse after filtering based on speech tagging results.

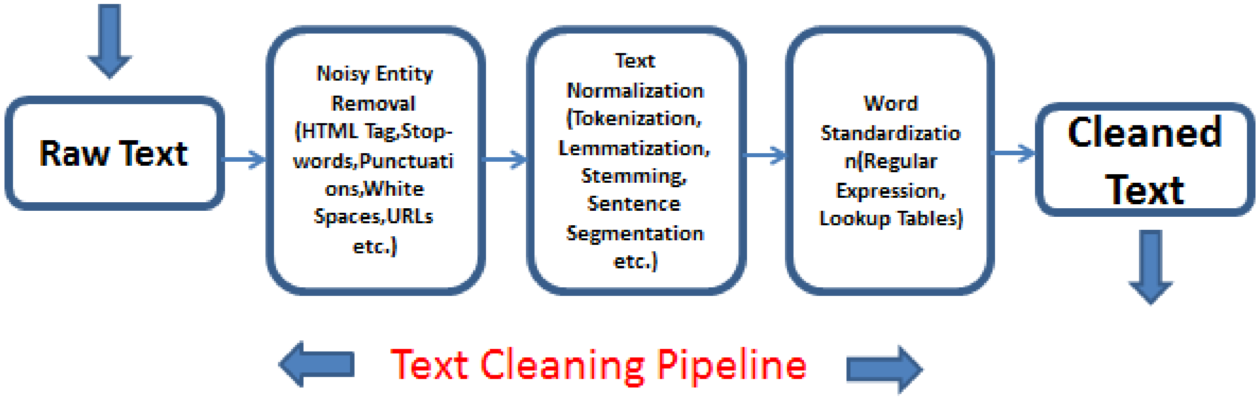

Raw Text
Noisy Entity Removal (HTML Tag,Stop-words,Punctuations,White Spaces,URLs etc.)
Text Normalization (Tokenization, Lemmatization, Stemming, Sentence Segmentation etc.)
Word Standardization(Regular Expression, Lookup Tables)
Cleaned Text
Text Cleaning Pipeline

Fig 1 Text cleaning pipeline
(Source: P M, Ashok Kumar, et al., 2021)

Exploratory Data Analysis

Considering the number of words as tweet length, we performed an in-depth descriptive analysis to understand how tweet length varied for the entire dataset, cyberbullying vs. non-cyberbullying tweets and for different categories of cyberbullying. After performing the analysis on the cleaned dataset, we observed that neutral tweets tended to be shorter in length compared to cyberbullying tweets. In addition, we also observed that the tweet length for each of the cyberbullying categories is roughly normally distributed.

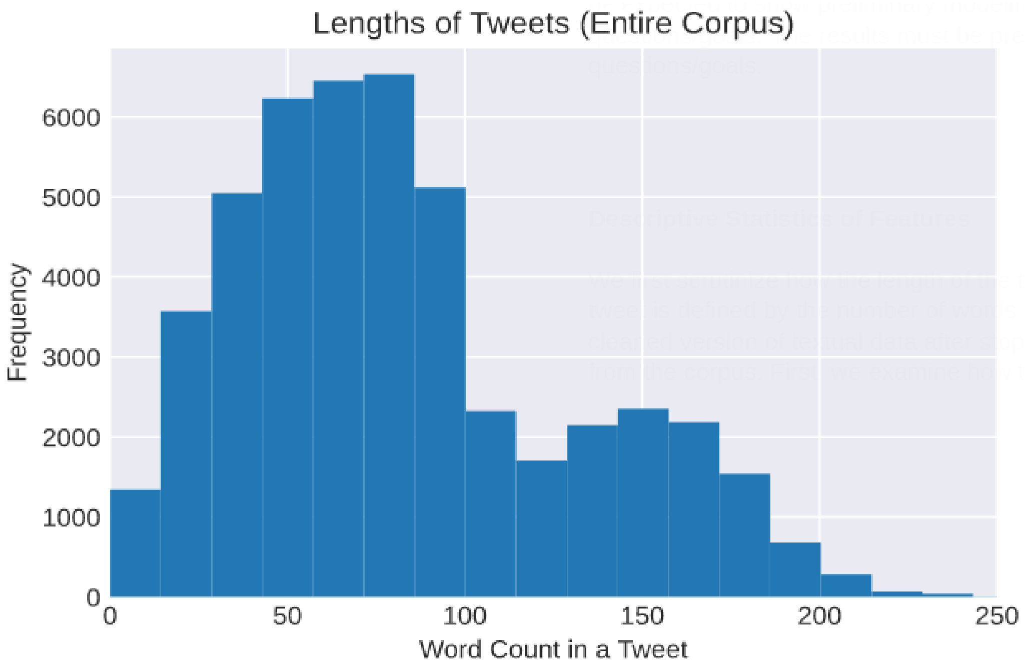


Fig 2 Lengths of Tweets (Entire Corpus)

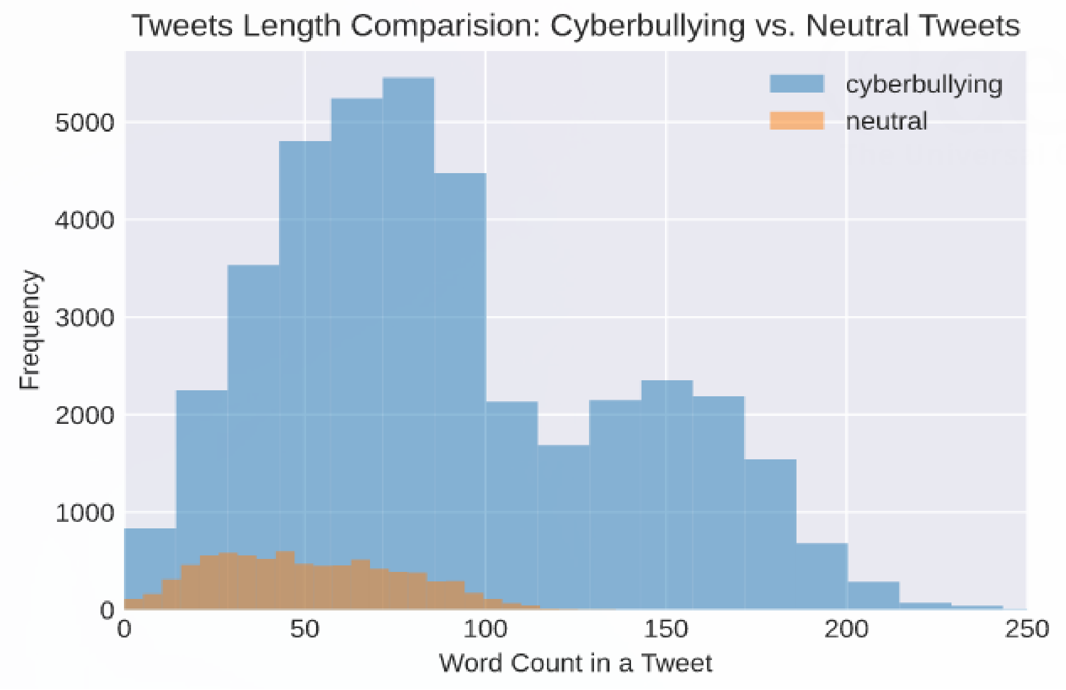


Fig 3 Tweets Length Comparison: Cyberbullying Vs. Natural Tweets

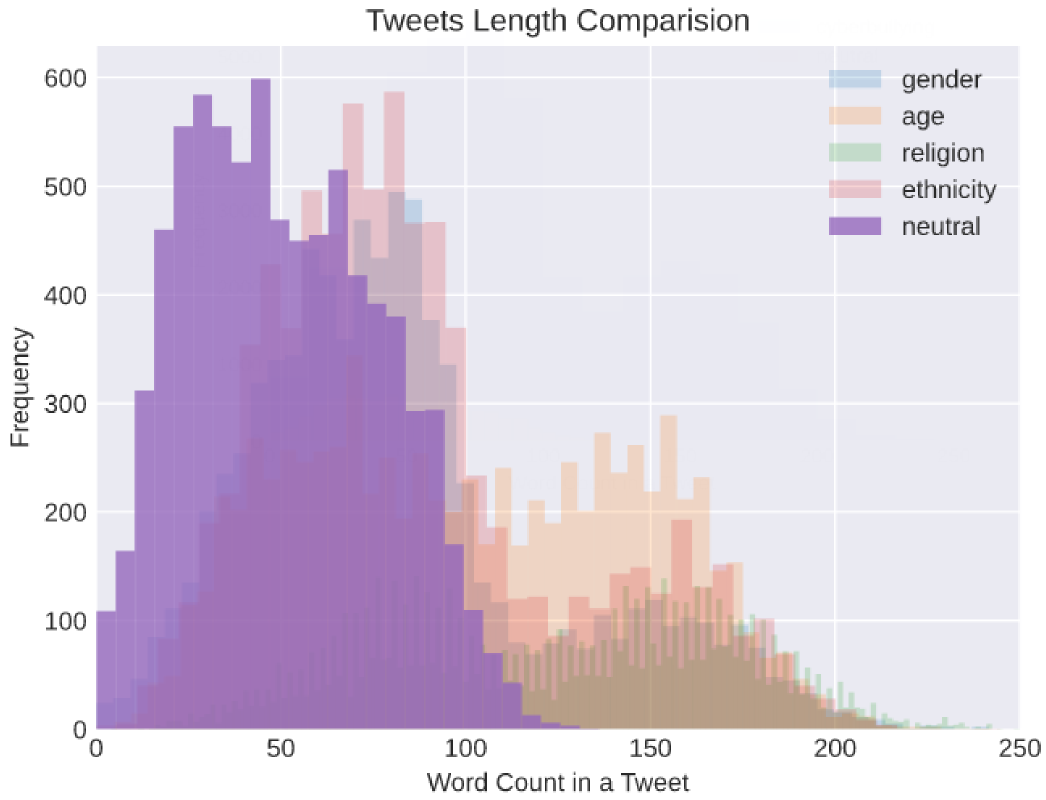


Fig 4 Tweets Length Comparison

Analysis Techniques

The known algorithms for text classification can be roughly split into three classes:

	Classical ML	Deep Learning	Transformers
Description	Bag-of-words-based	RNN-based	Transformer-based

This report scrutinizes the first two approaches.

Classical Machine Learning approach

The bag-of-words-based approach, including BoW and TF-IDF, naturally leads to sparse data. Such an approach works reasonably well, and many classical ML algorithms can be implemented using this vectorization approach. Some of the commonly used algorithms for text classification include Naïve Bayes, Logistic Regression, SVM, and tree-based algorithms (i.e., various types of decision trees with or without ensemble methods, such as gradient boosted trees and Random Forest). TF-IDF, a BoW-based vectorization technique, is often used in this type of approach to vectorize textual data. The TF-IDF equation assigns high scores to words with high discriminating power and low scores to terms in every document.

One major limitation with the BoW or TF-IDF-based approach is that word sequence is often discarded. Textual data is sequential data because the order of words matters in human languages. Ignoring word sequence is often a simplified approach, and it can lead to lower prediction accuracy in many cases. One rudimentary or rather crude solution to the word sequence problem is to implement N-grams (i.e., bigram, trigram) during the vectorization process. The limitation of the N-gram method is that it significantly increases the dimensionality of data and computation time. Since words are being used as features for prediction purposes, we already have an extremely high dimensionality without the implementation of N-grams in most text classification tasks. In our analysis, we took both unigrams and bigrams into account.

RNN-Based Approach

One way to tackle the word sequence problem is the usage of Recurrent Neural Networks (RNN). In a vanilla RNN architecture, the output of a layer is added to the next input and fed back into the same layer, doing so in a recursive manner (Debasish, 2022).

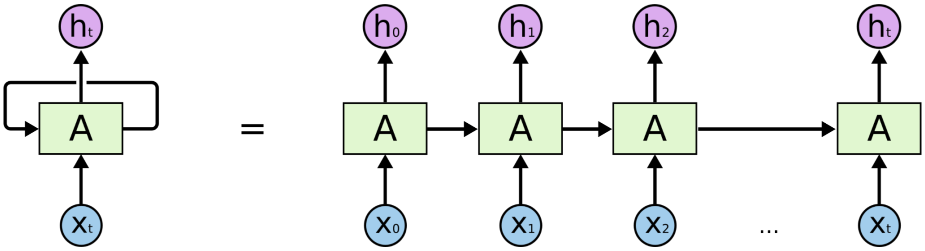


Fig 5 High-level RNN architecture

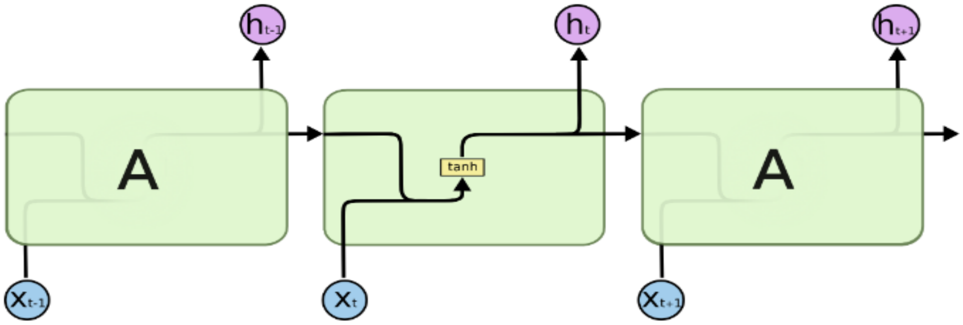


Fig 6 RNN Cell
(RNN architecture. Source: Source: Olah, 2015)

The approach works significantly better on sequential data. However, one main limitation with a vanilla RNN algorithm is that when the weights are getting updated during the backpropagation phase using the chain rule, sometimes the weights become extremely small to the point where they are insignificant. In deep learning, this is known as the vanishing gradient problem (or sometimes the weights can get extremely large – exploding gradient problem). Because of this limitation, a vanilla RNN only works well on very short sequences of words.

Long Short Term Memory (Hochreiter and Schmidhuber, 1997), LSTM, a special flavor of RNN, was introduced to tackle the short memory problem (aka vanishing gradient problem) mentioned above. An LSTM cell contains the basic structure of a vanilla RNN cell, but the cell is being modified to make it more sophisticated to handle long-term dependencies and patterns.

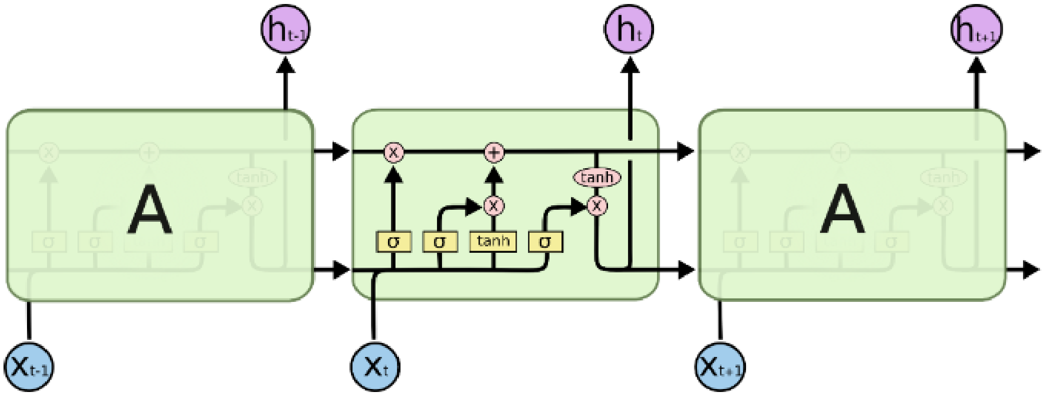


Fig 7 LSTM cell.
(Source: Olah, 2015)

An LSTM cell can be divided into three gates: forget gate, input gate, and output gate. The cell state, which is the horizontal line on top of the cell that contains multiplication and summation pointwise operations, is responsible for dealing with long-term dependencies and patterns. The forget gate, which is the vertical line at the very left, decides what information to filter from the long-term memory state, whereas the two vertical lines in the middle (input gate) determine what new information needs to be added to the cell state.

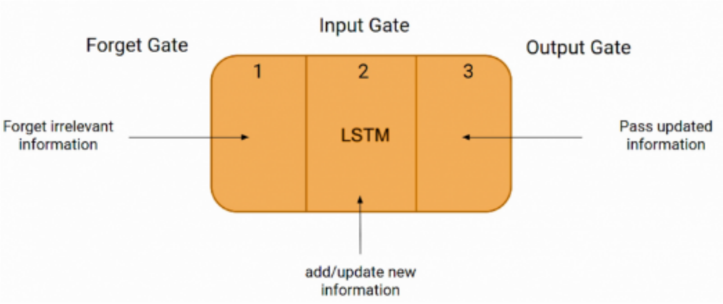


Fig 8 LSTM cell. High-Level illustration
(Source: Saxena, 2021)

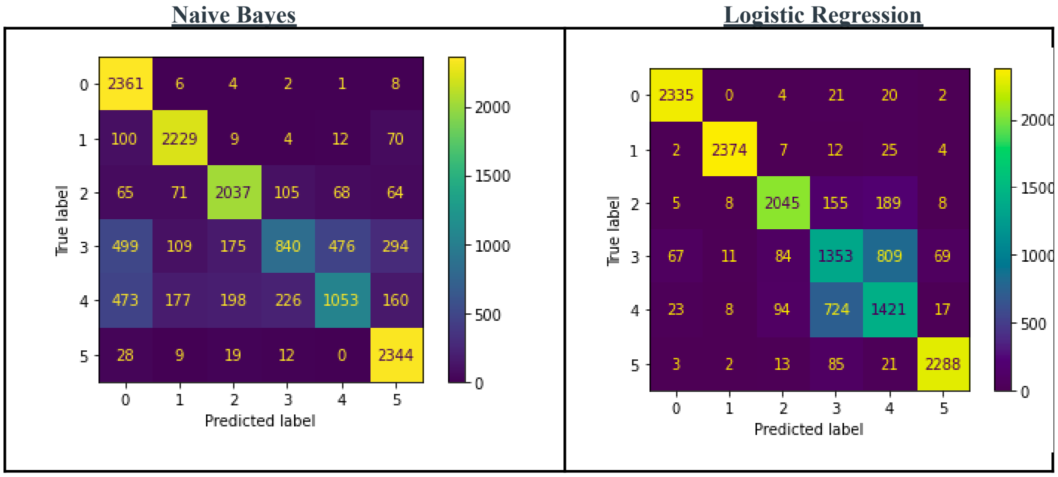
Intuitively, the cell state keeps tracking what further information needs to be filtered or added to long-term memory. Training in an LSTM network can be interpreted as understanding which patterns to forget because they're not essential and which patterns to track because they are crucial for text classification purposes.

RESULTS

The classification models were evaluated based on 10-fold cross-validation procedures on the balanced testing dataset, which contains approximately 2300 observations equally distributed among the classes. We then assessed and evaluated the effectiveness of these classifiers based on performance metrics using the average accuracy, precision, and recall scores.

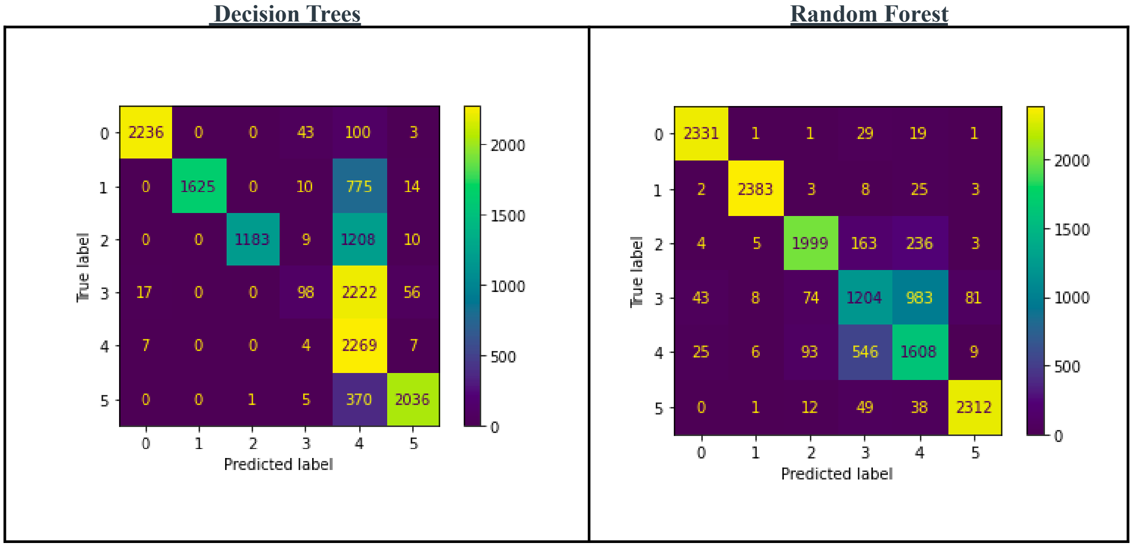
	Performance Metrics		
Models	**Accuracy**	**Precision**	**Recall**
Naïve Bayes	0.76	0.75	0.76
Logistic Regression	0.83	0.83	0.82
Support Vector Machines	**0.83**	**0.83**	**0.83**
Decision Trees	0.66	0.81	0.66
Random Forest	0.83	0.83	0.82
XGBoost	**0.85**	**0.86**	**0.85**
CatBoost	**0.84**	**0.86**	**0.84**
LSTM	0.78	0.78	0.77

From the models used, we observed that gradient boosting tree-based classifiers (XGBoost and CatBoost) performed exceedingly well compared to other classifiers. We believe that this is due to their ability to control overfitting through regularization and split finding through linear search. Similarly, they also are highly usage efficient for bigger datasets due to parallel learning, out-of-core computation, and cache-aware access techniques which minimizes their complexity in training. Similarly, we also observed that linear SVM performed exceedingly well due to their ability to handle tasks with high-dimensional feature spaces like text classification effectively.


Naive Bayes

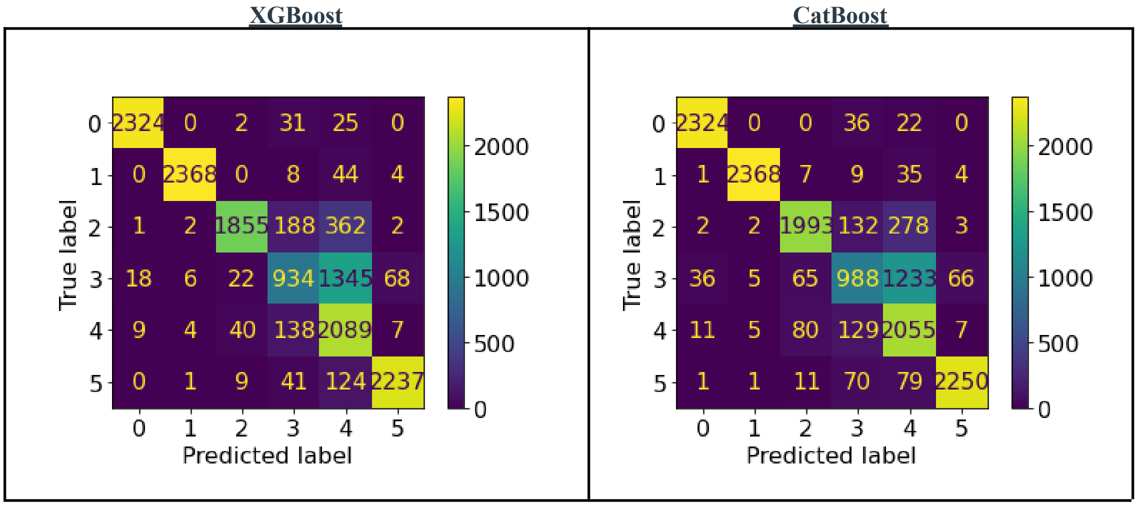
Logistic Regression

Fig 9 Confusion Matrix


Decision Trees

Random Forest

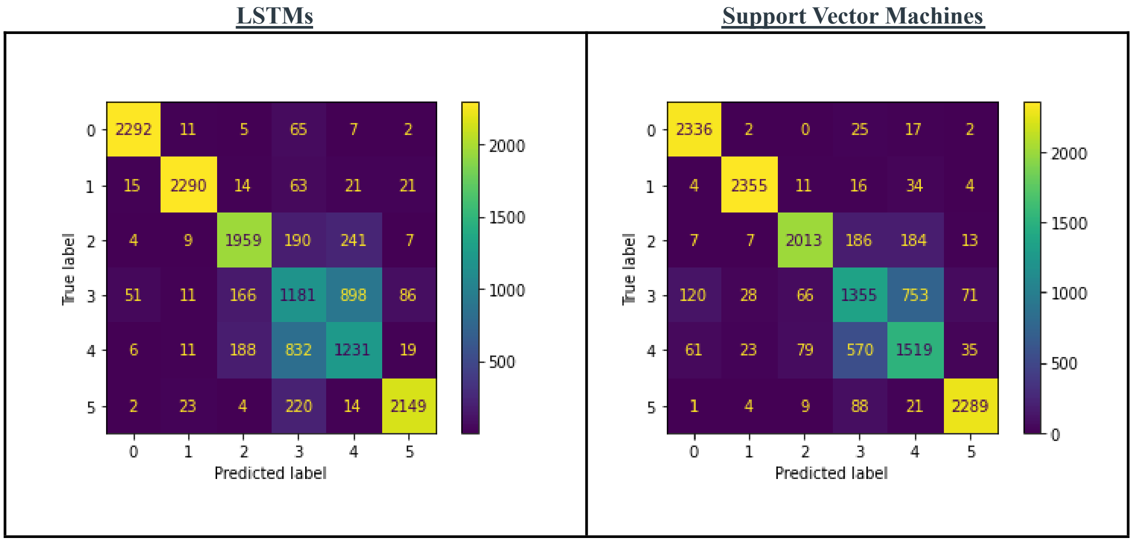
Fig 10 Confusion Matrix



XGBoost

CatBoost

Fig 11 Confusion Matrix



LSTMs

Support Vector Machines

Fig 12 Confusion Matrix

Observing the confusion matrices for all the models, we observed that the number of observations predicted correctly for class 3 i.e., “Not Cyberbullying” was the most difficult to classify. This can be attributed to the fact that tweets that are not related to cyberbullying could contain textual references to age, gender, ethnicity etc which might appear in tweets labeled for the other classes. Therefore, we realize there is further scope of improvement where the relationship between words can be understood further which will help in better prediction and labeling of non-cyberbullying tweets.

FUTURE DIRECTIONS

Based on the results that we observed from modeling analysis, we believe that there is further scope of improvement in terms of tweet classification accuracy, understanding the finer details between different indicative cyberbullying elements in a tweet as well as increasing the scope our tweet classification from the initial 6 classes and expand into other cyberbullying categories like "sexual inclination", "political preference" etc.

A few of the limitations that we observed in our research study included:

- The dataset used by us for classification purposes contained tweets in textual format along with emojis depicted using specific characters. As part of our data exploration and cleaning process, we performed iterative processes to remove non-useful characters along with emojis that were present in the corpus as our initial scope was to analyze cyberbullying limited to textual content present in the tweet. However, we realized that there could be tweets that contain only emoji content and could specifically be used to propagate bullying of a specific type. Given a larger time period, we would have looked towards replacing emojis with their text descriptions (Singh et al., 2019) which would aid us in creating an in-depth understanding of the context of the tweet and its association with cyberbullying/non-cyberbullying content.

- Considering the sensitive nature of the text contained in tweets especially pertaining to cyberbullying, it becomes pertinent to explore the relationships between various words within the tweet. Through our data cleaning methods, we ensured to remove words that didn't have any relevance to our research purpose. However, the further scope of improvement would have been to not exclude specific words from our corpus that have a strong association with cyberbullying. We would have also wanted to focus more on feature selection methods like Part-of-Speech tagging to understand the nouns, verbs, and adjectives present based on their definition and context, Named Entity Recognition methods like chunking and Conditional Random Fields (CRF), which uses word sequences as opposed to just words for entity recognition.

- Since the nature of text classification is quite complex, we would also have explored the implementation of bi-directional transformer models like BERT for our research case study for pre-training over a lot of unlabeled textual data to analyze and learn a language representation which would be utilized for specific purposes. BERT operates on the basis of two pre-training tasks, namely, Masked Language Model(MLM) and Next Sentence Prediction(NSP), which allows it to form associations and relationships between words and sentences. Through MLM, unstructured text is converted into tokens which are further used as input and output for training. MLM allows the model to perform bidirectional learning from the textual information as opposed to left-to-right/right-to-left unidirectional learning and thus allows it to understand and learn the meaning and context of every word from the words appearing before and after it. Additionally, an arbitrary number of tokens are masked(hidden) during training, and the aim is to predict their correct identities after the learning phase. Subsequently, in the NSP phase, the model formulates relationships between sentences by predicting the true next sentence for a pair. Here again, the model is trained using 50% random pairs with the remaining 50% correct pairs. BERT is also available for multilingual implementation and, therefore, would be useful for identifying bullying in tweets that contain non-English-based text.

- Since the dataset we performed our analysis was strictly limited to the content of the tweet and the labels that were associated with it, we weren't able to explore the metadata analysis of the tweet with respect to the time of the tweet, whether these types of tweets originated from specific locations and such similar factors.

- Finally, we would have also wanted to expand our present study, which was focused only on twitter-based text, to content posted on other social media platforms like Facebook etc. to analyze and identify cyberbullying behaviors and harassment. Through the usage of APIs, we would implement a similar model that would be reliable in detecting and flagging content that negatively impacts the mental health of individuals using these platforms.

ACKNOWLEDGMENT AND CONTRIBUTION

We would like to take this opportunity to list the contribution of each team member:

All of us took equal responsibility for finishing the team assignments. The equal contribution and positive reinforcement from every member made this assignment come together very well and was a wonderful learning experience. We brainstormed together, fine-tuned different models and analyzed them, and then collaborated to get everything together in a coherent manner.

REFERENCES

Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

Kalita, Debasish. “A Brief Overview of Recurrent Neural Networks (RNN).” Analytics Vidhya, 11 Mar. 2022, www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/.

Olah, Christopher. “Understanding LSTM Networks -- Colah’s Blog.” Github.io, 27 Aug. 2015, colah.github.io/posts/2015-08-Understanding-LSTMs/.

P M, Ashok Kumar & A, Anitha & H, Verma & M, Laxmannarayana. (2021). Finding State of Mind Through Emotion and Sentiment Analysis of the Twitter Text. 10.3233/APC210146.

Staudemeyer, Ralf C., and Eric Rothstein Morris. “Understanding LSTM -- a Tutorial into Long Short-Term Memory Recurrent Neural Networks.” ArXiv:1909.09586 [Cs], 12 Sept. 2019, arxiv.org/abs/1909.09586.

Saxena, Shipra. “LSTM | Introduction to LSTM | Long Short Term Memor.” Analytics Vidhya, 16 Mar. 2021, www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/.

Singh, A., Blanco, E., &amp; Jin, W. (2019). Incorporating emoji descriptions improves tweet classification. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1214

S, H., & Patchin. (2010). *Bullying, cyberbullying, and suicide*. Archives of suicide research: official journal of the International Academy for Suicide Research. Retrieved June 5, 2022, from https://pubmed.ncbi.nlm.nih.gov/20658375/

Wang, J., & Fu, K. (2020). *SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection*. IEEE Xplore. Retrieved June 5, 2022, from https://ieeexplore.ieee.org/document/9378065