Cyber Bullying Classification on Twitter

Yukti Rao, Parmesh Yadav, Nakul Periwal, Ojasva Singh
Department of Computer Science and Mathematics
Indraprastha Institute of Information Technology, Delhi

yukti20352@iiitd.ac.in,parmesh20319@iiitd.ac.in,nakul20316@iiitd.ac.in,ojasva20318@iiitd.ac.in

ABSTRACT

Cyberbullying classification on platforms like Twitter holds paramount significance in fostering a safe and inclusive digital environment. It serves as a crucial line of defense against the insidious spread of online harassment and vitriol. By employing advanced algorithms and machine learning techniques, we aim to identify the instances of cyberbullying, thereby protecting individuals from emotional distress and potential harm. Ultimately, the implementation of robust cyber bullying classification on Twitter contributes to a more respectful, tolerant, and ultimately more constructive digital discourse. After an initial examination of the gathered data, we noticed several redundant and unnecessary features (like emoticons, links, mentions, special characters,etc.). These were subsequently eliminated through the application of various pre-processing techniques (tokenization, stemming, and term weighting), resulting in the creation of a refined dataset. Then, we used classifiers (Naive Bayes classifier and SVM classifier) to make the predictions.
After scrutinizing a range of performance metrics, precision emerged as the most suitable criterion for evaluating their performance.

1. INTRODUCTION

Twitter is a widely used social media platform that enables users to share short messages, called "tweets," with their network of followers. Each tweet can contain up to 280 characters. These messages can include text, images, GIFs, videos, links, and more. Twitter is known for its real-time nature, making it a popular platform for news updates, discussions, and conversations on a wide range of topics.
Cyberbullying is when people say mean or hurtful things online, like on Twitter. It can be hard to know if cyberbullying is happening unless someone tells you about it. Identifying cyberbullying tweets means figuring out which tweets have mean or hurtful words in them. This helps keep the internet a safer and kinder place.
Due to COVID-19 and increased online activity, UNICEF warned about the heightened risk of cyberbullying, which can lead to academic struggles, emotional distress, and self-harm. Alarming statistics reveal that around 36.5% of middle and high school students have been affected by cyberbullying, with 87% witnessing it. The dataset used for cyberbullying analysis consists of over 47,000 tweets categorized by types of cyberbullying, including age, ethnicity, gender, and religion. The dataset has been balanced, with approximately 8,000 tweets per category, and includes both cyberbullying-related tweets and unrelated tweets.
This report provides a summary of the literature we've reviewed, the machine learning models we've developed and trained, and the resulting analysis and findings from these models. Additionally, we formulate conclusions based on our study.
After examining the outcomes, we noticed that a significant portion of tweets were part of cyberbullying of different classes, which shows a need to classify the cyberbullying on twitter.

2. LITERATURE SURVEY

Research Paper 1 [1]
The paper addresses the problem of cyberbullying in online communication and proposes a system that integrates the detection of cyberbullying words and rumor tweets on Twitter into a single application. It uses machine learning algorithms, such as Naïve Bayes and Random Forest classifiers, to detect cyberbullying words in tweet contents and comments. The proposed work also incorporates type and topic specific classification and Twitter speech-act classifier to detect rumor tweets on Twitter. Preprocessing techniques, such as removing

stop words, are applied to improve the accuracy of the input data. The training dataset consists of a list of cyberbullying words, which are used for detection. Intelligent text mining techniques are proposed to overcome the limitations of existing techniques and provide accurate results with a lower error rate. The proposed approach shows promising results in accurately detecting cyberbullying and rumor tweets on Twitter, providing accurate results with less error rate compared to existing techniques.

Research Paper 2 [2]

The paper aims to develop a machine learning model to automatically detect cyberbullying on Twitter, using SVM and Naïve Bayes classifiers. The proposed model utilizes the Twitter API to fetch tweets, which are then passed through a preprocessing step and feature extraction using the TF IDF vectorizer. The SVM and Naïve Bayes classifiers are trained on the extracted features to detect bullying content with varying accuracies. The machine learning model suggested in the paper attained a 71.25% accuracy in identifying cyberbullying with the SVM classifier, while achieving a 52.70% accuracy with the Naïve Bayes classifier. SVM demonstrated superior performance compared to Naïve Bayes when applied to the same dataset. The SVM algorithm attained the highest precision score of 71, whereas Naïve Bayes achieved a precision of 52. Additionally, SVM outperformed Naïve Bayes in terms of recall and f-score values. The research underscores the significance of creating models capable of automatically identifying and stopping instances of social media bullying. This is especially critical in light of the escalating prevalence of cyberbullying alongside the surge in popularity of social networking platforms. Although, the paper does not discuss the specific features or criteria used to identify and classify cyberbullying content, which may affect the accuracy and effectiveness of the model. It also does not provide information about the size or diversity of the dataset used for training and testing the machine learning models, which could impact the generalizability of the results.

Research Paper 3 [4]

The paper uses a feature engineering method to detect cyberbullying tweets in twitter, by sorting tweets in different groups based on gender, age, personality and feelings expressed they measure the severity of cyberbullying. In classification problems, class imbalance is a common problem, in which some categories have more instances as compared to others. To resolve this, SMOTE (syncthetic minority over-sampling technique) is used to basically increase the instances in smaller groups by 300 perc., by duplicating and adding them to the dataset to increase their representation. Along with SMOTE, weighted costs method is used for smaller group's misclassification. Different machine learning algorithms (Naive Bayes (NB), Support Vector Machine with Radial Basis Function (SVM with RBF) kernel, Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN)) were tried to figure out the right classifier for the problem of detecting cyberbullying in twitter. The methods which work better during cross validation are preferred. The multi-class classification method yielded 93% accuracy and 92% F-measure, proving the efficiency of the model.

3. DATA PREPROCESSING AND VISUALIZATION

3.1 Dataset Description

We obtained our dataset from Kaggle. UNICEF raised an alarm on April 15th, 2020, concerning the heightened threat of cyberbullying during the COVID-19, prolonged screen exposure, and reduced face-to-face interactions. In light of all of this, this dataset contains more than 47000 tweets.[3]

3.2 Data Visualisation

We plotted a histogram (Figure A) for the label distribution, to know the distribution of tweets under different labels/classes of cyberbullying, and not cyberbullying, and another histogram (Figure B) for the distribution of tweet length. For the outliers, we plotted box plots (Figure C) for tweet lengths per class, according to which we fixed a range for tweet lengths that we would consider for our model, and a box plot for word length per class (Figure D). Below are the histogram and box plots.

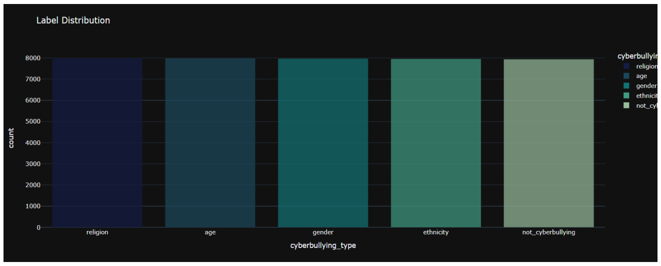

Figure A

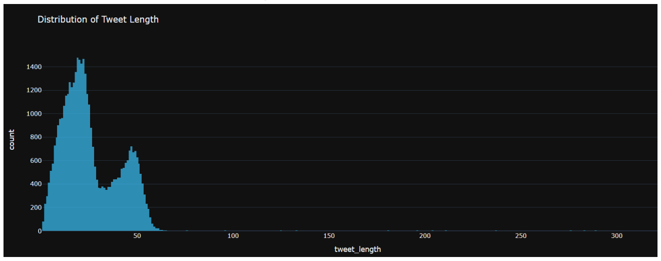

Figure B

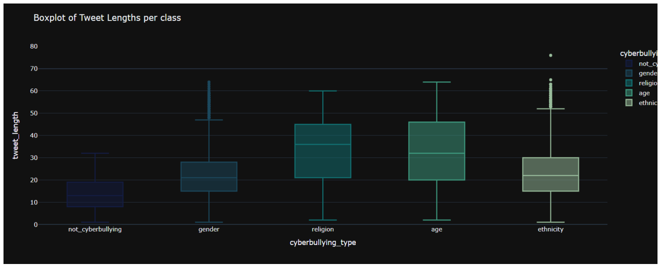

Figure C

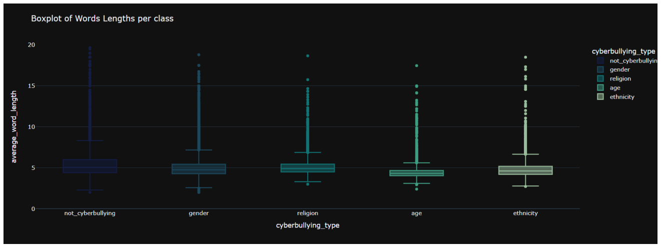

Figure D

4. PREPROCESSING

4.1 Feature Selection and Cleaning

We removed all the unnecessary emojis, space, links, mentions, non-utf8/ASCII characters, hashtags at the end of the sentence, multiple sequential spaces, some special characters, and words longer than 14 characters. We also did stemming by reducing the words to their base form. We also removed the duplicates. We removed tweets with less than four words and more than 100 words as they can be outliers.

4.2 Encoding

We converted the text into a numerical format that can be used for machine learning algorithms. The TF-IDF scores will represent the importance of each word in the documents, which can be used as features for training a machine learning model.

5. METHODOLOGY

5.1 Model Details

We have split the above dataset into a training and testing set. 20% of data is allocated to the testing set, while the remaining 80% will be used for the training set. We then trained our data on the following models, to predict the tweets in classification for cyberbullying:
Naive Bayes: a supervised machine learning algorithm, mainly used for classification tasks
SVM (Support vector machines): supervised machine learning algorithm, mainly for classification and regression tasks
KNN (K-Nearest Neighbours): supervised and non-parametric machine learning algorithm, for classification and regression tasks
Random Forest: supervised machine learning algorithm, mainly used for classification and regression, and also includes decision trees during training phase
ANN (Artificial neural networks): modelling method, based on the neural network inside the human brain, mainly used for classification, regression, etc.

5.2 Performance metrics

The metrics used by us to analyse our models were accuracy (i.e. it assesses the overall effectiveness of a classifier), precision (i.e. The proportion of correct positive predictions made by our model out of all the positive predictions it made), recall (i.e. The proportion of correct positive predictions made by our model in comparison to the total number of actual positives within the class), and FI-score (i.e. the harmonic mean of precision and recall).

6. RESULTS AND ANALYSIS

Upon analyzing the provided dataset with the model, the table below showcases the computed values for precision, recall, accuracy, and F1 score.

	SVM	Naive Bayes	KNN	Random Forest	ANN
Accuracy	0.92	0.83	0.84	0.94	0.90
Precision	0.93	0.84	0.84	0.94	0.89
Recall	0.92	0.83	0.84	0.94	0.90
F1 score	0.93	0.82	0.83	0.94	0.89

The SVM, Naive Bayes, KNN, random forest and ANN classifiers have provided the precision 93%, 84% 84%, 94% and 89% respectively. The random forest classifier gives the best results for all the performance metrics.

The SVM classifier provides higher precision because it chooses a decision boundary carefully and looks for the best line or curve that separates different categories.

The Naive Bayes classifier assumes all features to be independent, which is not always true in real-world data, hence the lower precision.

KNN provides low precision since it is a non-parametric, instance-based algorithm, it classifies data points based on the majority class among their K-nearest neighbours. It can perform well when the decision boundary is non-linear.

Random Forest performs the best among all the models since it is robust, handles non-linearity well and reduces overfitting by averaging multiple trees. Also, it is an ensemble learning method which builds multiple decision trees and merges their predictions.

ANN could've performed better if there was more data and computational resources, which would've helped in modelling the complex relationships in data through the interconnected layer of neurons.

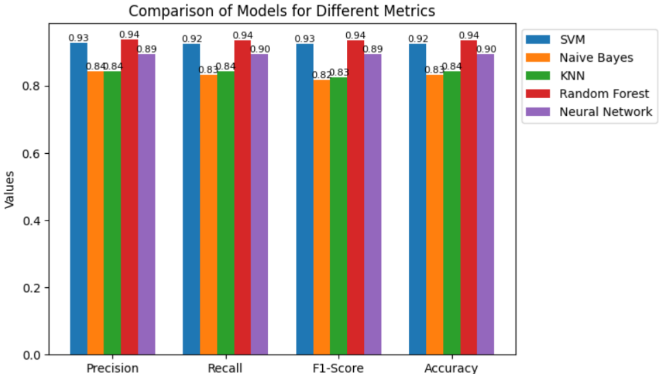

Figure E

7. CONCLUSIONS

We learned the significance of pre-processing the dataset to have a more understandable dataset and more precise and accurate results.

Visualizing the data makes the information easier to understand by converting the text into graphs. Even complex data seems easier to understand in graphs and can help analyse the data.
The SVM and Random Forest classifiers have given the best precision and accuracy results.

We have done all the proposed work, including training data on five classifiers: SVM, Naive Bayes, KNN, Random Forest and ANN, and analyzing the performance metrics.
The final presentation is also ready for the project.

8. INDIVIDUAL TASKS

Team Member	Tasks
Parmesh	Data Visualization, Data Preprocessing, Performance Evaluation and Analysis
Nakul	SVM and KNN Training, End

	Sem Report
Ojasva	Naïve Bayes and ANN training, End Sem Report
Yukti	Data Visualization, Data Preprocessing, Random Forest training, End Sem Report

9. REFERENCES

[1] https://www.researchgate.net/profile/Sheeba-Immanuvelrajkumar/publication/333320174_AUTOMATIC_DETECTION_OF_CYBERBULLYING_FROM_TWITTER/links/5ce6a24ba6fdccc9ddc93238/AUTOMATIC-DETECTION-OF-CYBERBULLYING-FROM-TWITTER.pdf

[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9120893&casa_token=2-8RL7yb-tUAAAAA:pZUKZeBKfiuOZ8G7dd0JxRuLwL69GEdsW2hDa39jhG0fqH0wstFLhZs-tlxBVCXhJprny4hDsiUf&tag=1

[3] https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

[4] https://www.mdpi.com/2227-9709/7/4/52