Spring 2024

Machine Learning Approaches for Cyberbullying Detection

Roland Fiagbe
University of Central Florida, ro210333@ucf.edu

Machine Learning Approaches for Cyberbullying Detection*

Roland Fiagbe
Department of Statistics and Data Science
University of Central Florida
Orlando, United States
fiagberoland@knights.ucf.edu

***Abstract*—Cyberbullying refers to the act of bullying using electronic means and the internet. In recent years, this act has been identified to be a major problem among young people and even adults. It can negatively impact one's emotions and lead to adverse outcomes like depression, anxiety, harassment, and suicide, among others. This has led to the need to employ machine learning techniques to automatically detect cyberbullying and prevent them on various social media platforms. In this study, we want to analyze the combination of some Natural Language Processing (NLP) algorithms (such as Bag-of-Words and TF-IDF) with some popular machine learning algorithms (such as Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbor (KNN), and Extreme Gradient Boosting(XGboost)) to detect cyberbullying on Twitter. The NLP methods were employed to extract features from tweets and convert them to numerical vectors and these features were analyzed with the machine learning algorithms. Comparing their performances and accuracy, the Extreme Gradient Boosting(XGboost) model emerged as the best-performing classifier irrespective of whether it uses features from bag-of-words or TF-IDF.**

***Index Terms*—Cyberbullying, Twitter, classification**

I. Introduction

In recent years, the use of social media has been very popular among young people around the world. It has been a popular medium where people discuss societal issues, communicate and share ideas and knowledge. Social media has made this possible through the use of texts, images, audio, videos, etc. Although social media have advantageously impacted lives, however, it also comes with some disadvantages. One of these major problems is the aggressive intentional act or behavior that is carried out by people, via electronic forms of communication, continuously against victims who are unable to easily defend themselves. This act is being performed via the use of social media and is referred to as "Cyberbullying" [6]. Some types of these social bullying are physical, verbal, relational, age, sex, and also indirect (eg. rumor spreading). Cyberbullying has gained much popularity on social media platforms like Facebook, Instagram, and Twitter among others.

Psychologically, cyberbullying negatively affects one's emotions and has the long-run effect of even committing suicide. Research has shown that cyberbullying can lead to adverse outcomes like depression, anxiety, harassment, and suicide, among others [8]. Therefore, in recent times, there has been a need for researchers to come up with ways to automatically detect this bullying and prevent them on various social media platforms. In response to this challenge, the study by Wang et al [8] presents a significant advancement in our understanding of cyberbullying and cyberviolence through a novel User-Activity-Content (UAC) triangular view. This perspective underscores the critical importance of analyzing the interplay among users' characteristics, their activities, and the content they produce to effectively detect and combat cyberbullying. Many studies have been done to address this problem. The goal of most of these studies is to formulate an online cyberbullying detection algorithm using machine learning techniques, particularly classification methods. This is done by applying machine learning algorithms to search abusive words in online textual communications, posts, and discussions. Dinakar et al [5] applied classification techniques such as SVM and Naive Bayes to model the detection of textual cyberbullying and compared the performance of these two techniques. Moreover, Dadvar et al [4] discovered that using SVM to model cyberbullying detection can be improved by using gender information. Chavan and Shylaja [3] began by weighting the comments based on the probability of being offensive and generated features using the skip-gram technique. Then they applied SVM and Logistic Regression to the generated features to detect cyber-aggressive comments by peers on social media networks. In alignment with these efforts, Emil Agbemade [1] study [8] employs machine learning to automate the detection of ethnic and religious cyberbullying, further highlighting the field's growing sophistication. Additionally, Amir Alipour Yengejeh [2] demonstrated the efficacy of Logistic Regression, Multinomial Naive Bayes, K-Nearest Neighbor, and Extreme Gradient Boosting (XGBoost) for detecting cyberbullying on Twitter, emphasizing the critical role of feature extraction techniques.

In this project, we seek to find the best-performing machine learning approach among Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbor (KNN), and Extreme Gradient Boosting(XGboost) in detecting cyberbullying on Twitter. To achieve this, we will first use text feature selection techniques like Bag-of-Word (BoW) and TF-IDF to convert the text data into mathematical vectors (features). Basically, we will combine Bag of words and TF-IDF with different classification algorithms such as LR, NB, K-NN, and XGboost to classify ethnicity-based cyberbullying, religion-based cyberbullying, or no cyberbullying.

II. Data

The data used in this project is a popular cyberbullying dataset available to the general public on IEEEDataPort https://ieee-dataport.org/open-access/fine-grained-balanced-cyberbullying-dataset. The dataset was generated by J. Wang et al [7]. The dataset is text data extracted from Twitter and it contains $8,000$ tweets for each type of cyberbullying such as religion, ethnicity, age, other, and not-cyberbullying but this project will focus on only ethnicity, religion, and not-cyberbullying. Hence, the total number of tweets used in this project is $24,000$.

III. Text Data Cleaning

In this section, we performed text data cleaning on the Twitter data to remove unhelpful parts and noise from the data. The dataset contains $8,000$ tweets for three labels; ethnicity, religion, and not-cyberbullying. The following labels with codes were created for the tweets, not-cyberbullying - 0, ethnicity - 1, and religion - 2. Following that, all the tweets were converted to lower cases to help in pre-processing the data. Stop words such as is, no, at, the, not, etc, URLs and mentions such as @username were removed since they do not provide any useful information in the models. Furthermore, numeric data in the tweets, punctuation marks, and emojis were removed. Tokenization and lemmatization were performed to split paragraphs and sentences into small units that can be assigned meaning and group together different forms of the same word. Below is a sample of the cleaned data.

	Tweets	Labels	LabelsEncoded
0	word katandandre food crapilicious mkr	notcb	0
1	aussietv white mkr theblock imacelebrityau tod...	notcb	0
2	classy whore red velvet cupcake	notcb	0
3	meh p thanks head up not concerned another ang...	notcb	0
4	isi account pretending kurdish account like is...	notcb	0
...	...	...	...
23995	imagine christian came together like time day ...	religion	2
23996	support justice initial problem morphed became...	religion	2
23997	rt harbour doubt muslim believe sharia note da...	religion	2
23998	one thing muslim want exterminate everyone not...	religion	2
23999	quran precludes woman human right adherent did...	religion	2

24000 rows × 3 columns

Fig. 1. Cleaned Text Data

IV. Exploratory Data Analysis

In this section, we performed some EDA by analyzing the frequency of words in the tweets using a word cloud. The word cloud visually gives a summary of the frequency of words by sizing the most frequently mentioned words in the tweets. The plots below show the word cloud of not-cyberbullying (Figure 2), cyberbullying by ethnicity (Figure 3), and cyberbullying by religion (Figure 4). The most frequently mentioned words in the tweets are shown by size.

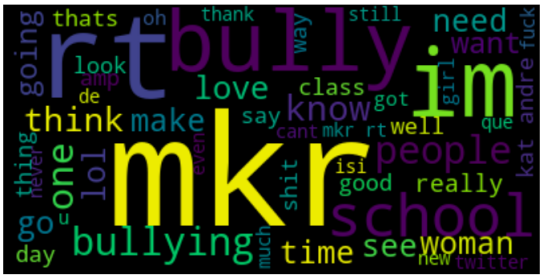


Fig. 2. Word Cloud for Not-Cyberbullying Tweets

Fig. 3. Word Cloud for Ethnicity Cyberbullying Tweets

V. Modelling

In this chapter, we will combine Bag-of-words (BoW) and TF-IDF with different classification algorithms such as Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbor (KNN), and Extreme Gradient Boosting(XGboost) to model the tweet data. Finally, we will analyze and compare their performance in multi-label classification.

A. Features Extraction

In other to model the text data with machine learning algorithms, the data needs to be converted to numeric data. There are various models used to extract features such as Bag-of-Words, TF-IDF, Word Embedding, N-grams, etc., however, this study will focus on Bag-of-words and TF-IDF to extract the relevant features.

1) Bag-of-Words (BoW): Bag-of-words is a basic and simple approach to extracting features from text data. It is defined as the representation of text that illustrates the frequency of words in a text document. The concept of bag-of-words is creating a set of vectors showing the word count occurrence and disregarding its grammatical details.

2) Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF is an improved version of BoW by discovering the meaning of sentences a word is contained in. A breakdown of the term is explained below.

TF: represents the frequency of each term/word occurring in the data. It calculates the weight $(tf(t,d))$ of every single term/word (t) in a document (d). It is computed as

$$tf(t,d) = \frac{\text{the count of t in d}}{\text{number of words in d}} \tag{1}$$

Fig. 4. Word Cloud for Religion Cyberbullying Tweets

DF: this is the count of terms/words (t) in a document of a whole corpus which represents the importance of terms/words. It is computed as

$$df(t) = \text{frequency of t in the document} \tag{2}$$

IDF: Due to the presence and frequency of stop words, the inverse document frequency (IDF) is computed to diminish the weight of stop words and measure the weight of useful words. It is computed as

$$IDF = \log\left(\frac{N}{df+1}\right) \tag{3}$$

Putting it all together;

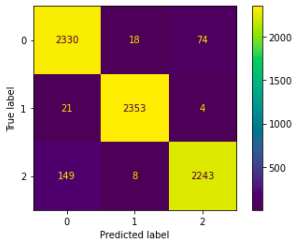
$$TF - IDF = tf(t,d) \times \log\left(\frac{N}{df+1}\right) \tag{4}$$

B. Model Building

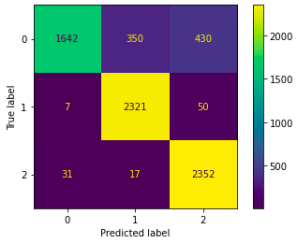
In the model-building stage, we feed the extracted features to the various multi-class classification algorithms to compare their performances in detecting cyberbullying. The models were trained using the training set from BoW and TF-IDF features and tested with a test set. The following classification algorithms were implemented.

1) Logistic Regression (LR)
2) Naive Bayes (NB)
3) K-Nearest Neighbor (KNN)
4) Extreme Gradient Boosting(XGboost)

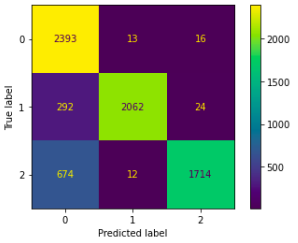
1) Classification with BoW Features: Here, we evaluate the performance of the above-mentioned classification algorithms using bag-of-words extracted features. The classification was done in a multi-class fashion since the data has three labels. The graphs below show the confusion matrix for the three-class classification algorithms.



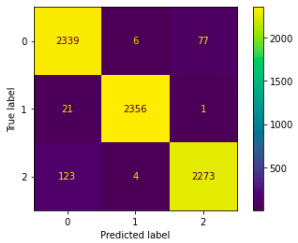
Logistic with BoW



Naive Bayes with BoW

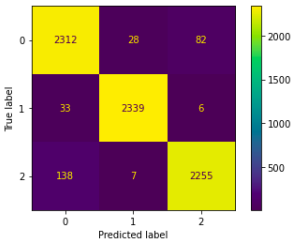


KNN with BoW

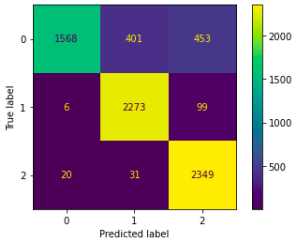


XGBoost with BoW

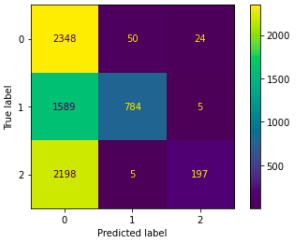
2) Classification with TF-IDF Features: Here, we evaluate the performance of the above-mentioned classification algorithms using TF-IDF extracted features. The same process for BoW features was repeated in this section. The graphs below show the confusion matrix for the three-class classification algorithms.



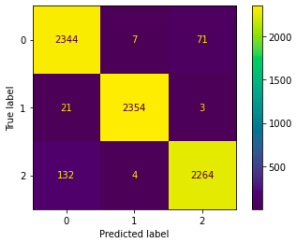
Logistic with TF-IDF



Naive Bayes with TF-IDF



KNN with TF-IDF



XGBoost with TF-IDF

C. Results

The table below (Table I) reports the accuracy measures of the combination of the Natural Language Processing algorithms and machine learning algorithms. The models were compared via their accuracy rates and F1- scores computed using the confusion matrix.

From the results, we could see that the Extreme Gradient Boosting (XGboost) model is the best-performing classifier irrespective of whether it uses features from bag-of-words or TF-IDF. The accuracy and F1- score reported were 97% and 97% respectively. The K-Nearest Neighbor (KNN) algorithm had the worst performance when it is fed with features extracted from TF-IDF. The reason for this drastic decline in performance could be a further study from this project.

VI. CONCLUSION

In this project, we have explored the performance of some well-known machine learning algorithms in detecting cyberbullying using Twitter data. We analyzed the combination of

Models	Accuracy Measures	Bag-of-Words	TF-IDF
Logistic Regression (LR)	Accuracy	0.96	0.96
	F1-Score	0.96	0.96
Naive Bayes (NB)	Accuracy	0.88	0.86
	F1-Score	0.87	0.85
K-Nearest Neighbor (KNN)	Accuracy	0.86	0.46
	F1-Score	0.86	0.40
Extreme Gradient Boosting (XGBoost)	Accuracy	0.97	0.97
	F1-Score	0.97	0.97

TABLE I
RESULTS SHOWING ACCURACY MEASURES OF THE COMBINATION OF NATURAL LANGUAGE PROCESSING ALGORITHMS AND MACHINE LEARNING ALGORITHMS

some Natural Language Processing (NLP) algorithms (such Bag-of-Words and TF-IDF) with some popular machine learning algorithms such as (Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbor (KNN), and Extreme Gradient Boosting(XGboost)). The NLP methods were used to extract features from tweets and convert them to numerical vectors and these features were analyzed with machine learning techniques. Comparing their performance and accuracy, the Extreme Gradient Boosting (XGboost) model emerged as the best-performing classifier irrespective of whether it uses features from bag-of-words or TF-IDF.

REFERENCES

[1] Emil Agbemade. *Silent Agony: Automated Detection of Ethnic and Religious Cyberbullying Using Machine Learning*. https://stars.library.ucf.edu/data-science-mining/13. STARS, (2023).

[2] Amir Alipour Yengejeh. *Combating Cyberbullying on Social Media: A Machine Learning Approach with Text Analysis on Twitter*. https://stars.library.ucf.edu/data-science-mining/15/. STARS, (2024).

[3] Vikas S Chavan and SS Shylaja. "Machine learning approach for detection of cyber-aggressive comments by peers on social media network". In: *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2015, pp. 2354–2358.

[4] Maral Dadvar et al. "Improved cyberbullying detection using gender information". In: *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent. 2012.

[5] Karthik Dinakar, Roi Reichart, and Henry Lieberman. "Modeling the detection of textual cyberbullying". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 3. 2011, pp. 11–17.

[6] Peter K Smith et al. "Cyberbullying: Its nature and impact in secondary school pupils". In: *Journal of child psychology and psychiatry* 49.4 (2008), pp. 376–385.

[7] Jason Wang, Kaiqun Fu, and Chang-Tien Lu. "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection". In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 2020, pp. 1699–1708.

[8] Shuwen Wang et al. "Cyberbullying and Cyberviolence Detection: A Triangular User-Activity-Content View". In: *IEEE/CAA Journal of Automatica Sinica* 9.8 (2022), pp. 1384–1405.