Detection of Cyberbullying Using Machine Learning

[1]Neel Gheewala, [2]Yash Shah, [3]Sakshi Patel, [4]Prof. Brijesh Vala

[1,2,3]Students, [4]Exam Coordinator
Computer Science and Engineering Department,
PIET, Parul University. (NAAC A++ Grade)

Abstract:
Nowadays, social media has become the best leadership conference in recent years. With the widespread use of social media, cyberbullying, cyberbullying and cybercrime have increased, which has had a positive impact on people's worldview. People's health can be negatively affected by cyberbullying and can sometimes lead to mental health problems. Explicit messages about sex and rumors spread by many users are two examples of how this affects relationships. The number of researchers interested in cyberbullying has increased in recent years. One of our goals is to use natural language processing (NLP) and random forest algorithms to create a system that can identify the nature of online abuse. The rapid spread of the COVID-19 virus has changed the culture, leading to cyberbullying, especially among young people. Most teenagers follow this model. As some of the platforms that facilitate online dating grow, so does social media for bullying. The COVID-19 virus has not only caused cyberbullying but also changed the nature of human relationships online. Bullying is becoming an issue as many people start working from home. The planning process is divided into data maintenance, text search, word embedding, regression analysis and other methods. The paper mining model uses lemmatization technology, which helps improve the accuracy of the model. TF-IDF is used for word embeddings and Tf-idf gives good mathematical meaning to the model. The random forest algorithm is used for the distribution of data in the conceptual model; This will help reduce overfitting of the data in the model.

Keywords: Machine learning, Analysis and Detection.

1. Introduction

Given the growth of the internet and other media, it is not surprising that young people use all forms of social media to hurt and offend each other. However, there is still a lack of research on various phenomena and measures of cyberbullying in Arab societies. Cyberbullying is also a problem in many other civilizations due to social and cultural barriers.

When it comes to daily life, social media has become the most influential platform in modern history. The proliferation of social media has not only had a positive impact on people's worldview, but has also led to the emergence of online bullying, harassment and bad acts in cyberspace. Cyberbullying not only affects people's physical health but also their mental health; Sometimes it can also be the cause of the victim's psychological problems.

Moreover, the social media environment is influenced by many factors, including sexist comments and rumors from many users. In recent years, more and more researchers are interested in how to identify cases of online bullying. One of our goals is to use natural language processing (NLP) and random forest algorithms to create the ability to find cases of online bullying.

Because social life transcends physical barriers to human interaction and includes contact with strangers, there is a need to define and investigate the context of cyberbullying. Cyberbullying makes victims feel like they are being attacked anywhere because the Internet is just a click away. It can have psychological, physical and emotional effects on the victim. Cyber bullying usually occurs on social media in the form of text or images. If it can distinguish bullying text from non-bullying text, the system contact accordingly.

2. Literature review

This literature review examines existing research on cyberbullying investigations and crime investigations, providing insight into their benefits, limitations, and advancements. Highlights of the document include:

1. RR Dalvi et al:

In this study, real-time Twitter API was used to collect tweets and create data. The proposed model was tested on collected data Support Vector Machine and Naive Bayes. They use TFIDF vectorizer to extract features. The results show that the accuracy of the support vector machine-based cyberbullying model is close to 71.25%, which is better than the negative Bayes accuracy of close to 52.75%

2. J. Yadav et al:

In this study, The model has been trained and tested on the Form spring forum and Wikipedia repositories. The proposed model achieved 98% accuracy for the Spring Paper dataset and 96% accuracy for the Wikipedia dataset, which is higher compared to previous models. The proposed model gives better results to the Wikipedia dataset due to larger g dimension without oversampling, while the Spring Paper dataset requires oversampling.

3. Dinakar et al:

In this study, Break down cyberbullying into different contexts such as race, gender, culture, and intelligence. Therefore, they used some controversial videos on YouTube as data to classify ads using four different methods (Naive Bayes (NB), rule based on Jrip, tree based on J48, and SVM). The data contains approximately 50,000 analyzes split into 50% training, 30% validation, and 20% testing. However, according to Jrip, the best accuracy rate achieved by the law still does not exceed 80%.

4. Hee et al:

In this study, proposed a method to identify the best types of cyberbullying, such as insults and threats. The cyberbullying terms used by the author have similar characteristics to those found in OSN; This content (in English and Dutch) is provided by the Ask.fm website. The authors divide the content of cyberbullying conversations into three categories: perpetrators, victims, and bystanders. The jury is divided into two groups: representatives who protect the victim (defender) and representatives who support the bully (defense eight). Then, support vector machine was used to analyze the difference. However, in this article, we will focus on exploring cyberbullying on Twitter. It is more difficult to detect bullying in tweet content.

5. Ozel et al:

In this study, Use of Turkish Research in Cyberbullying Investigation. They collected streaming data from Twitter to create benchmark data for their experimental study. Using the bag-of-words approach, each tweet is given its own vector and parsed using various machine learning methods (Support Vector Machine, Naive Bayes, C4.5, and KNN) to determine whether it is most related to torture. In terms of

F-measure, Naive Bayes classifier out performs other classifiers with 79% accuracy.

6. Reynolds, kontostathis, Edwards:

In this study, The authors used C4.5, k-nearest neighbor (KNN), and support vector machine (SVM) classification methods and tested them on a dataset containing words from the Formspring.me platform. According to the test results, the C4.5 decision tree algorithm outperforms KNN and SVM classifiers with a detection rate accuracy of 78.5%.

7. Yin et al:

For the first time, research has been conducted on automatic detection of online cyberbullying. The authors use three different data sets to analyze online bullying in three different ways. The Konggate platform is used to collect a set of data, another set of data is collected from social networking sites such as Reddit. A standard classification system and various extraction methods (N-gram and Inverse Term Frequency (TF-IDF)) are used for the classification function. Although the results of their experiments were not clear, the research became a starting point for further research.

In summary, the review of the literature shows the evolution of cyberbullying detection, emphasizes the effectiveness of modern methods, while recommending the reduction of cyber stalking and online harassment. This information informed further analysis and formed the basis of

Twitter's current harassment investigation.

3. Current problem

The following problems have been identified in the literature review of various methods and techniques used to create models, and this section also explains how the disadvantages are addressed with the help of the plan. In the process of text mining, stem extraction is used to remove several characters in a word, in other words, to remove the stem. In literature, stem extraction techniques are often used to find the meaning of a word, but these techniques have many disadvantages, such as causing incorrect meanings or spellings. Also output incorrect results can reduce the accuracy of the prediction.

In the conceptual model, lemmatization is used to improve the meaning of words by analyzing them as part of speech to create a more accurate translation. For quality analysis, the lemmatization method is a better choice and the algorithm provides better performance. The TF-IDF method is used in the preparation of word embeddings, which is a statistical measurement used to determine the mathematical meaning of a word in a document. The TF-IDF method helps provide a way to associate each word in a document with multiple representations by assigning larger values to rare words and lower values to frequent words. the truth of every word in the document. For prediction, many studies use the logistic regression algorithm based on the design of random variables. Logistic regression has some limitations, for example logistic regression assumes a positive relationship between variables, accurate results cannot be predicted when the relationship between variables is not linear. A forest matching algorithm based on decision trees is used to solve problems with the proposed model; this algorithm reduces data overfitting and thus improves the overall accuracy of the model. It is also suitable for categorical and continuous data. The algorithm can also remove missing values from the given data.

4. Problem Statement

During data research on various methods and algorithms used to build models, the following problems related to these methods are listed below, this section also explains how they can be solved with the help of process related questions. Text mining techniques, stem extraction is used to extract several characters from a word or other words used to remove the stem. Literature often uses stemming techniques to find the exact meaning of a word, but these techniques have many disadvantages such as meaning extraction or spelling errors. Additionally, the output contains erroneous values that reduce the accuracy of the prediction. In the application model, lemmatization is used to improve the meanings of words by making them a part of speech based on analysis and thus create a more accurate translation. For a good analysis, the lemmatization method is a better choice and the algorithm provides better performance.

For word embedding, many writers use the word2vec algorithm, which is based on neural networks and learns to associate a word in a document or text. The main problem with using word2vec for word embedding is that it does not handle words other than word quality. It also relies on local knowledge provided in the database, and new language teaching methods cannot be shared. It also requires a lot of text to illustrate the model. To overcome the shortcomings of Word2vec, the TF-IDF technique was used in the word embedding application. Word embedding is a measure used to determine the mathematical meanings of the words in the paper. By assigning larger values to rare words and lower values to frequent words, the TF-IDF method helps provide a way to associate each word in a document with more representations and check the relevance of each word in the document.

For prediction, many studies use logistic regression algorithm based on modeling of random variables. Logistic regression has some limitations, for example, logistic regression assumes linearity between variables, accuracy is not possible when the relationship between variables is not linear. To solve problems related to logistic regression, the proposed model adopts random forest algorithm based on decision tree; this reduces overfitting of the data and increases the overall accuracy of the model. It is also suitable for categorical and continuous data. The algorithm also gave values that were not significant in the given data.

5. Methodology

This article presents a method for identifying cyberbullying on social media that is not only based on a theoretical analysis, but also takes into account the gathering, cleaning, tokenization, lemmatization vectorization and TF-IDF of the sentence before classifying it as hate speech. To achieve our goals, we began with traditional theoretical analysis, a content search of the text to identify and extract important information from the material to understand the meaning of thoughts, feelings, or thoughts.To achieve our goals, we began with traditional theoretical analysis, a content search of the text to identify and extract important information from the material to understand the meaning of thoughts, feelings, or thoughts. We will introduce the process of “engagement” that can influence and guide the process of detecting cyberbullying.

We divided all results into 6 categories: religion, age, gender, ethnicity, other types of cyberbullying and non-bullying. All of these features are classified based on existing data analysis and each is unique. Special properties define text. Data, description and selection of independent features is an important step, for the performance of algorithms in pattern recognition and, classification problems.

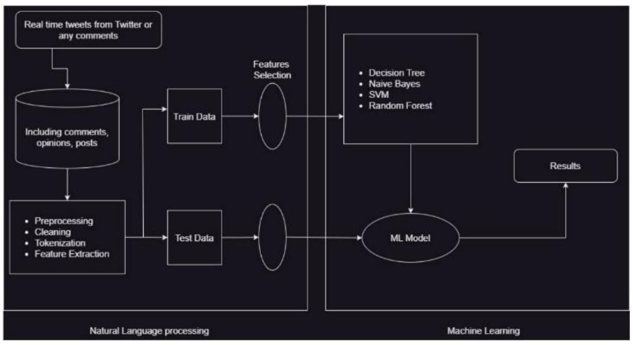


Fig. 1. Flow of the System

- Data Gathering:

Cyberbullying occurs in many forms. For example, this can be used by sharing or posting inappropriate video content; sharing hashtags using vulgar language; or sharing hashtags without the owner's permission, etc. However, the most common form of cyberbullying involves the forwarding of text messages. So Twitter Api can be used to immediately write ads available on. To do this, Tweepy, a library written in Python, will extract information from posts containing relevant hashtags. We get data from Kaggle in the application system. Import libraries for the application process All libraries are listed and submitted here help identify cyberbullying including:

1. Nltk: Utilizing the tools and techniques provided by natural language processing in order to conduct an analysis of sentiment.
2. Sklearn: Provides many useful tools for machine learning and statistical models, including classification, all of which will be used in the system we demonstrate.
3. Pandas: It is a fast, powerful, flexible and easy-to-use, open source data analysis and management tool developed in Python programming language.

- Data Cleaning:

It is the process of correcting or removing inaccurate, incomplete, incomplete, duplicate or incomplete data from the database. When combining data from multiple sources, data can easily be duplicated or mislabeled. If the data is wrong, the results and algorithms will not be reliable, even if they appear correct. There is no way to document the exact steps in the maintenance file because the maintenance file varies from file to file. But it's important to create a model for your data cleansing process so you know you're always doing it correctly.

- Tokenization:

With a standard tokenization process, the text is split into pieces and converted into meaningful word tokens. so algorithm can easily understand meaning of sentence and identify about category .

- Lemmatiztion:

This process is speech-based and is used to help people define their words as a block and determine the lemma (lexical form) of the words with dictionary meaning.

- Vectoriztion:

Finally, vectorization is an NLP method used to assign a weight, i.e. a probability, to each word in the dataset. This can be used to find and take into account predictive words.

- TF-IDF:

It stands for Term Frequency and Inverse Document Frequency, is a metric that measures how related words are in a document. This is done by giving the following two expressions:

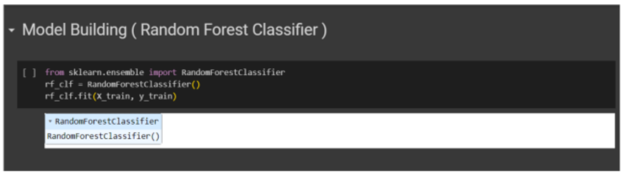
1. The frequency of a word in the document/text.
2. Information changes the frequency of a word in the document/text. It shows how few or many words are in the document. The closer the value is to 0, the more words there are.

Put these two together to get the TF-IDF score of each word in the file.

6. Algorithm

- Random Forest Classifier:

Random forest is an ensemble learning algorithm that uses multiple decision trees to make predictions. Decision trees are a simple machine learning algorithm that can be used to classify data. Random forest works by training multiple decision trees on different subsets of data. Predictions from decision trees are combined to create the final prediction. Random forests can be very efficient, but training them can be computationally expensive. Random forest algorithm uses less data than neural networks and SVM. The random forest algorithm returns output with multiple distributions. Although the accuracy of random forest classification is lower than neural network and SVM, due to the greater classification ability of random forest algorithm, we can get more detailed information from neural network only binary output 0 or SVM is taken as 1 [false or true].



Model Building (Random Forest Classifier)

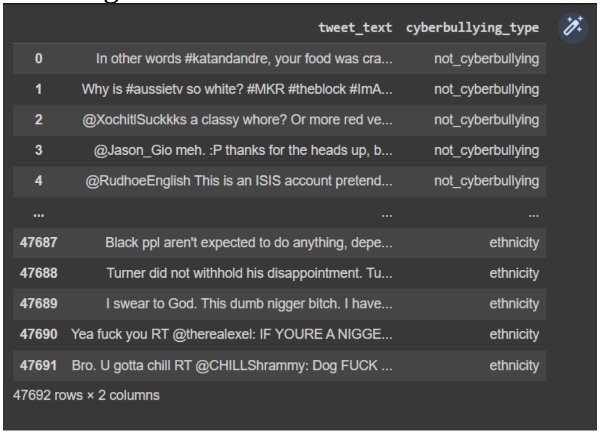
```
from sklearn.ensemble import RandomForestClassifier
rf_clf = RandomForestClassifier()
rf_clf.fit(X_train, y_train)
```

```
▾ RandomForestClassifier
RandomForestClassifier()
```

Fig. 2. Implementation of Random Forest

7. Result Analysis

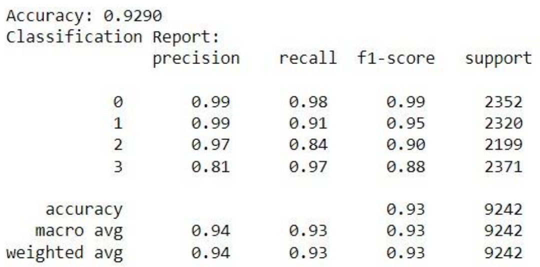
Figure 3 shows the processing strategies and predictions we made during testing. We took a tweet from Twitter containing signs of bullying and applied it to our model. Figure 4 shows the distribution based on our test data. Here the letters 0 and 1 do not represent bullying and bullying. Figure 5 represents the confusion matrix based on the results of our test data. Table 1 shows that the accuracy of random forest is 92.90. When applying the Twitter dataset for sentiment analysis, it can be thought that it provides better results than traditional machine learning models on similar data.



	tweet_text	cyberbullying_type
0	In other words #katandandre, your food was cra...	not_cyberbullying
1	Why is #aussietv so white? #MKR #theblock #ImA...	not_cyberbullying
2	@XochitlSuckkks a classy whore? Or more red ve...	not_cyberbullying
3	@Jason_Gio meh. :P thanks for the heads up, b...	not_cyberbullying
4	@RudhoeEnglish This is an ISIS account pretend...	not_cyberbullying
...	...	...
47687	Black ppl aren't expected to do anything, depe...	ethnicity
47688	Turner did not withhold his disappointment. Tu...	ethnicity
47689	I swear to God. This dumb nigger bitch. I have...	ethnicity
47690	Yea fuck you RT @therealexel: IF YOURE A NIGGE...	ethnicity
47691	Bro. U gotta chill RT @CHILLShrammy: Dog FUCK ...	ethnicity

47692 rows × 2 columns

Fig. 3. Dataset

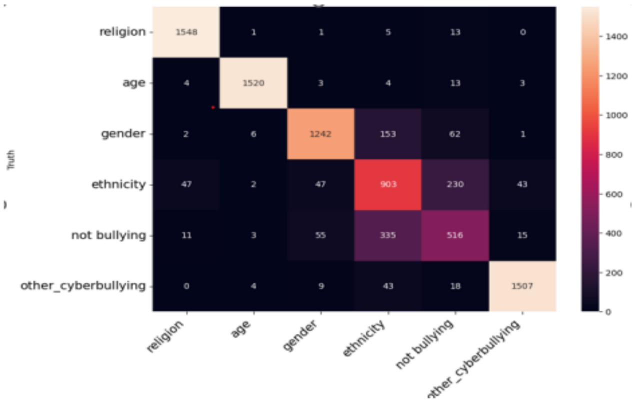


```
Accuracy: 0.9290
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      2352
           1       0.99      0.91      0.95      2320
           2       0.97      0.84      0.90      2199
           3       0.81      0.97      0.88      2371

    accuracy                           0.93      9242
   macro avg       0.94      0.93      0.93      9242
weighted avg       0.94      0.93      0.93      9242
```

Fig. 4. Classification Report



	religion	age	gender	ethnicity	not bullying	other_cyberbullying
religion	1548	1	1	5	13	0
age	4	1520	3	4	13	3
gender	2	6	1242	153	62	1
ethnicity	47	2	47	903	230	43
not bullying	11	3	55	335	516	15
other_cyberbullying	0	4	9	43	18	1507

Fig. 5. Confusion Matrix

8. CONCLUSION

- The Predictive analytics is a field of analysis used to control the extraction of data. Forecasting methods use this information to predict patterns and designs. This research is based on the estimation of the problem of online bullying. A random forest method is used to classify important data as offensive or non-invasive. If it is a crime, separate the type of crime. Compared to Naive Bayes, Random Forest classification models perform better in terms of accuracy, precision, recall, and f-measure. When using the random forest algorithm, we achieved an overall accuracy of 0.92% multidistribution.

ACKNOWLEDGMENTS

I would like to thank Prof. Brijesh Vala for guiding me during my BTech studies. His support and advice were crucial to the success of this project.

I am grateful to the IT department and Parul Institute of Engineering and Technology for providing a platform where I can present my research. I'm grateful for this opportunity.

REFERENCES:

1. Google Scholar (https://scholar.google.com)
2. IEEE Xplore (https://ieeexplore.ieee.org)
3. ResearchGate (https://www.researchgate.net)
4. Google Bard (https://bard.google.com/)
5. Git Hub(https://github.com)
6. Kaggle(https://www.kaggle.com/)

RESEARCH PAPER

1. R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi:10.1109/ICICCS48265.2020.9120893. (2020)
2. J. Yadav, D. Kumar and D. Chauhan, Cyberbullying Detection using Pre-Trained BERT Model, ICESC, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700. (2020)
3. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.

4. Hee, C.V.; Lefever, E.; Verhoeven, B.; Mennes, J.; Desmet, B.; Pauw, G.D.; Hoste, V. Automatic detection and prevention of cyberbullying. In Proceedings of the International Conference on Human and Social Analytics (HUSO 2015), Nice, France, 18–22 July 2015; pp. 13–18.
5. Ozel, S.A.; Saraç, E.; Akdemir, S.; Aksu, H. Detection of cyberbullying on social media messages in turkish. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 366–370.
6. reynolds, K.; Kontostathis, A.; Edwards, L. Using machine learning to detect cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, Washington, DC, USA, 18–21 December 2011; pp. 241–244.
7. Yin, D.; Xue, Z.; Hong, L.; Davison, B.D.; Kontostathis, A.; Edwards, L. Detection of harassment on web 2.0. In Proceedings of the Content Analysis in the WEB, Madrid, Spain, 21 April 2009; pp. 1–7.