Nano World Journal

https://doi.org/10.17756/nwj.2023-s4-091

Research Article

Open Access

Cyber Threat Detection: Machine Learning Algorithms for Advanced Security

Ajmeera Kiran* **and B. Veda Vidhya**

Department of Computer Science and Engineering, M.L.R. Institute of Technology, Hyderabad, Telangana, India

***Correspondence to:**
Ajmeera Kiran
Department of Computer Science and Engineering,
M.L.R. Institute of Technology,
Hyderabad, Telangana, India.
E-mail: kiranphd.jntuh@gmail.com

Received: September 19, 2023
Accepted: December 06, 2023
Published: December 11, 2023

Citation: Kiran A, Vidhya BV. 2023. Cyber Threat Detection: Machine Learning Algorithms for Advanced Security. *Nano World J* 9(S4): S533-S539.

Copyright: © 2023 Kiran and Vidhya. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY) (http://creativecommons.org/licenses/by/4.0/) which permits commercial use, including reproduction, adaptation, and distribution of the article provided the original author and source are credited.

Published by United Scientific Group

Abstract

On social media platforms, it has become increasingly common to persecute people online. It has led to negative outcomes including suicide and despair as a consequence. The policing of content on social media platforms is becoming an increasingly relevant topic. The work that is presented here makes use of Natural Language Processing and machine learning methodologies in order to construct a cyber harassment detection model. This model is built with the help of data obtained from instances of cyber bullying, tweets containing hate speech from Twitter, and personal assaults from Wikipedia. In order to establish which strategy is more effective, we are looking into two feature extraction approaches and six classification algorithms (SVM (Support vector machine), random forest, XGBoost, MLP (Multi-layer perceptron), and logistic regression). This article provides a synopsis of the overall process for identifying instances of cyberbullying as well as the methodology that is most important.

Keywords

Cyber bullying, Machine learning, Support vector machine, Random forest, XGBoost, Logistic regression, Feature extraction

Introduction

Machine learning is a group of computer algorithms that can learn from their experiences and improve without being explicitly coded, and the name "machine learning" is used to characterize this group of algorithms. The field of artificial intelligence known as machine learning combines data analysis with statistical techniques to make predictions about future outcomes, which can then be put to use to deliver insightful interpretations. It provides a literal interpretation of human thought. The same way that people learn from their mistakes, machine learning learns as well. This implies that the better and more accurate the data that is provided, the higher the level of efficiency that can be achieved. When we apply the same scenario to the machine, it possesses the ability to foresee how things will turn out. When presented with an example that it has seen before, the machine, just like a person, has difficulty making predictions. One of the most essential responsibilities of a data scientist is to select the information that will be input into the computer in an accurate and thoughtful manner. A collection of attributes that are utilized to find a solution to a problem is called a feature vector. One way to think of the data that is utilized in the process of finding a solution to a problem is as a feature vector [1].

The machine makes use of complex algorithms in order to reduce the complexity of reality and construct a model based on this result. The data are therefore characterized and summarized into a model during the learning step of the process. The importance of technology in our everyday lives is greater than it has ever been. As a result of the development of the internet. These days, every-

one is talking about how popular social networking sites are. However, just like there are people who abuse anything else, there will be people who show up late or early at certain times. Cyberbullying has become increasingly common in today's society. Social networking platforms make personal interaction much easier to establish and maintain. Even if the usage of social networking has become increasingly widespread over the years, a significant number of people continue to exploit it in unethical and immoral ways. This is a trend that we notice among young adults and those in their teen years. One of the unfavorable things that they do is engage in online harassment and bullying of one another. In most cases, they will laugh it off with a simple joke or tell you "not to take it so seriously." This fight takes place online, but it often escalates into threats in real life for specific individuals [2].

There have been cases of persons killing themselves by their own hand. It is absolutely necessary to put an end to these actions as soon as possible. Any measure might be taken to prevent this from happening; for instance, if a person's tweet or post is deemed offensive, that person's data could be destroyed, or the account could be suspended for a certain amount of time. When someone is harassed, intimidated, embarrassed, or targeted online for the purpose of entertainment or through well-planned tactics, this behavior is known as cyberbullying.

Nanotechnology enhances cyber threat detection by integrating with machine learning through nanosensors for improved intrusion detection, robust encryption using nanoscale materials, secure communication channels with nanostructured materials, tamper-resistant nanoscale memory devices, boosted processor performance for machine learning algorithms, nanostructured filters for network traffic analysis, nanodevices monitoring behaviors for anomaly detection, and exploration of nanotechnology in quantum computing components. This integration offers innovative solutions for evolving cybersecurity challenges, emphasizing ongoing interdisciplinary research and development to realize its full potential in enhancing cyber threat detection [3].

Technology to repeatedly target an individual with the intent to intimidate, humiliate, or embarrass them is an example of cyber persecution. Most fights that begin on social media frequently escalate into threats made in real life. There are a lot of people who have messed up their lives by either killing themselves or becoming a victim of mental harassment. They have used this point to identify the genuine authors of the tweets that have caused the nuisance, and in order to find them, they have employed machine learning techniques. This is due to the fact that the false profiles and the original one are closely tied to one another. Existing systems have a number of drawbacks, including reduced accuracy, techniques that only search for patterns that already exist in the data, and the majority of these techniques being more manual processes that rely on human participation and decision making. It has been discovered via studies conducted on the topic of cyberbullying that the majority of adolescents use the internet for roughly three hours every day. As they become increasingly dependent on social media platforms, adolescents are making themselves more susceptible to being bullied online.

Cyberbullying has repercussions not only for their personal lives but also for their social lives. It has a direct impact on an individual's mental health, which can result in feelings of despair, impaired capacity to make decisions, and many other negative outcomes. The government of the country ought to take action to address this problem in order to save the lives of the people, i.e., save people from being subjected to cyberbullying. The gravity of this issue can be understood by looking at the recent tragedy in Russia involving a blue whale that claimed the lives of hundreds of people, the majority of whom were youngsters. This is not a trivial issue, despite how it may sound; it has been responsible for the loss of the lives of a great number of people; as a result, it should be resolved as quickly as possible by cultivating a safe and non-hazardous atmosphere on social networking networks. However, it would appear that people are using social media in a destructive way because to the personalities that are associated with it. As well as being a serious problem for society as a whole, cyberbullying is a kind of online abuse [4].

With this history and this desire, implementing pertinent strategies to identify cyberbullying on social media can help to prevent the incidence of cyberbullying. [Cyberbullying] Because of this, the results of this research present a strategy for identifying instances of cyberbullying that is founded on social network analysis and data mining. Sentiment analysis, word matching methodology, and social website analysis are the three primary approaches that will be investigated by this research in order to discover instances of cyberbullying. In addition to the tactic, we will talk about the experimental design that was used to assess how well it worked.

Cyberbullying is one form of unacceptable behavior that carries the potential for extremely negative outcomes. As a consequence of this, it is extremely important to identify instances of cyberbullying as rapidly as is practicable by reading tweets. The most frequent approaches to identifying instances of cyberbullying are generally stand-alone and, as a result, laborious to carry out. In this investigation, the fundamentals of collaborative computing are applied to improve the performance of the detection task. This study presents and discusses a wide range of different models for collaborative work. The detection method appears to be both more efficient and more accurate than the stand-alone paradigm, according to the preliminary findings.

The act of cyberbullying can have a broad variety of negative and dangerous consequences, including the suicide of young people. One of the realistic options for reducing and putting an end to cyberbullying is to use techniques based on machine learning and natural language processing to automatically detect content that constitutes bullying. However, the majority of the currently available methods in the research are merely text categorization algorithms that don't take into account bullying. In this study, we introduce a representation learning approach to the problem of identifying instances of cyberbullying. The obtained representation is then combined with Bag-of-Words and latent semantic traits before being fed into an SVM classifier.

We employ word embeddings to extend a list of insulting words and provide various weights to bullying aspects. Our methodology is evaluated against a variety of standard text

representation learning models and cyberbullying detection algorithms through the use of an experiment that uses a dataset from Twitter. Because of this test, we were able to establish that our method is superior.

The way in which people work together has been radically transformed, which has added a new facet to the process of communication. However, even if innovations benefit us in many facets of life, they also have a number of unintended side effects that have an impact on people in a variety of different ways. Cyberbullying is one of the worst side effects of the internet. Cyberbullying is a form of online harassment that can have detrimental effects on a person's mental health, relationships, and even their physical well-being. It is a form of online bullying that occurs when one person harasses another online. To address this issue, we proposed a ground-breaking method that makes use of deep neural networks to identify instances of cyberbullying. It is possible to improve the performance of already-in-place systems by utilizing the Convolution Neural Network.

Experimentation

Proposed system

This problem, which falls under the category of Natural Language Processing, is described in this article, which is titled cyber threat detection. This is a problem of binary classification in which two forms of cyberbullying, namely hate speech on Twitter and attacks on Wikipedia, are recognized, and it is necessary to classify them as either containing or not containing cyberbullying respectively. The suggested system makes use of a number of different machine learning algorithms, including SVM, random forest classifier, Bernoulli Naive Bayes, logistic regression, MLP, and XGBoost. Comparisons of the accuracies of the different models were made so that the "best" model could be chosen. A support vector classifier is utilized in order to plot a hyperplane in an N-dimensional space that differentiates between individual data points. For the purpose of optimizing the marginal values, a loss function known as the hinge function is applied. Since the data can be partitioned linearly, a linear SVM, which is the most effective method for such data, is utilized.

A random forest, as its name suggests, is a forest that is made up of numerous trees. Similarly, this classifier is made up of a number of decision trees, each of which makes a prediction regarding a different category. for a set of inputs and using an aggregate of all results predicted by an individual decision tree, a final prediction is obtained; this means that the final prediction is the forecast that received the most votes from the decision trees. A prediction can be obtained from a decision tree by the use of a set of rules known as decision rules, which are learned from feature vectors.

Logistic regression is frequently used because it can accurately predict binary outcomes. The ease with which this classifier can be built and educated contributes to its utility. Since Bernoulli Naive Bayes produces more accurate answers with smaller datasets, it is most commonly utilized in these situations. In comparison to other models, it is both quick and accurate in its forecasting abilities.

As can be seen in figure 1, the suggested system demonstrates that SVM has an accuracy of approximately 96% when it comes to detecting cyberbullying content, which is an improvement over the solutions that are already in use. Individuals will be shielded from the barrage of cyberbullying that occurs on social media platforms by our approach. The proposed approach, rather than merely searching for patterns, goes beyond what has occurred in the past to anticipate future outcomes based on the data that already exists. The outcomes of the suggested approach will be more accurate when compared to the way that is now being used.

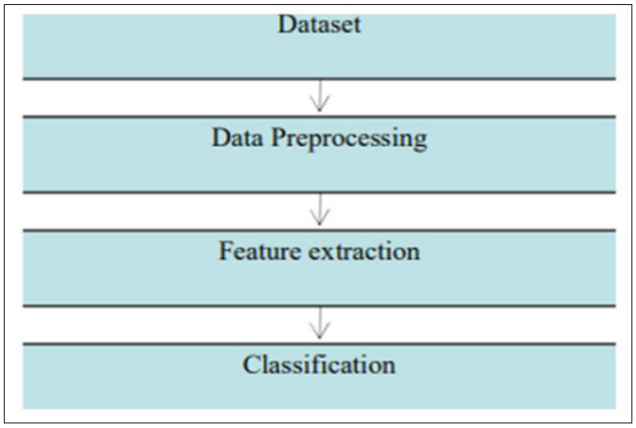

Figure 1: Methodology of proposed technique.

Implementation setup

Data collection

Twitter dataset

This is the first and most important step in the process of building a machine learning model. Collecting data is an important step that will determine how accurate the model will be. The more and better data that we use to train and test the model, the more accurate it will be. Data collection can be accomplished via a variety of methods, such as web scraping, manual interventions, and many more. In order to identify instances of cyber persecution on social media platforms, machine learning is utilized. The data set has been acquired from Kaggle dataset (Figure 2).

The dataset consists of 7614 individual data. The dataset has three columns, which are detailed below. Id: unique id. Labels: 1: offensive 0: non offensive 3. Tweet : comment.

Wikipedia dataset

The data set from Wikipedia includes approximately one million comments that are tagged as being personal attacks. Only a restricted number of comments, which is approximately 40000, have been taken into consideration because using 1 million comments requires a significant amount of time. Of these, 13000 remarks are categorized as being a personal attack. The same data set is subdivided into a train dataset and a test dataset, with respective percentages of 70% and 30%

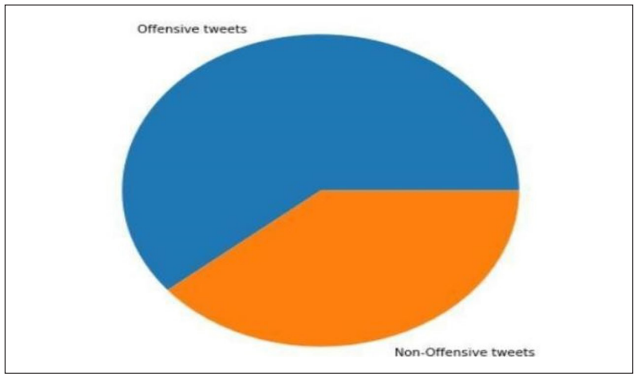

Figure 2: Distribution of tweets in Twitter dataset.

(Figure 3).

Data pre-processing

Processing the data is necessary in order to get it to the point where it may be used. That is to say, the data have to be put through some kind of procedure before they are ready for further use. The practice of assigning tokens to individual words in a sentence is known as tokenization. For example: After the raw data has been transformed to lowercase, it is then cleaned by eliminating any punctuation marks or special characters that may have been present. The following methods of natural processing are utilized in order to finish the pre-processing process, also known as getting the data ready to be used in subsequent steps.

Tokenization

Tokenization is the process of allocating tokens to a sentence's words. For instance: Consider the sentence "He will do it", it is tokenized as "He", "will", "do", "it". Tokenization is just splitting the words of a sentence. The tokens are generated based on certain rules, in the case of regex tokenizer, the regular expression are the rules, it tokenizes based on the matching regular expression provided. Stemming: It is common knowledge that a single word can take many different forms; however, since doing so would only serve to slow down the process, the focus is instead placed on the most common form of the word, which is then used as a root. When the word appears in the text, it is compared to the root word in order to determine its meaning. For example, the word "work" is the root of the two words "working" and "work." Because the two branch words derived from the root word 'labor' signify the same meaning, they ought to be dealt with in the same manner. The Porter Stemmer, the Lancaster Stemmer, the Snowball Stemmer, and the Regexp Stemmer are the four different kinds of stemmers that are included in the Natural Language ToolKit. Port Stemmer is utilized in the completion of this project. Take out the word "halt," and then read: Stop words are words that do not contribute anything meaningful to the meaning of a sentence. Stop words are common in the English language and include terms such as what, is, at, and an. It is possible to omit these words without changing the significance of the sentence in any way. NLTK provides a list of stop words in the English language that can be applied to filter out any and all tweets. The preprocessing steps for the dataset are depicted in the image below. The preprocessing of the dataset is depicted in the figure below, which applies to both the Twitter and Wikipedia datasets (Figure 4).

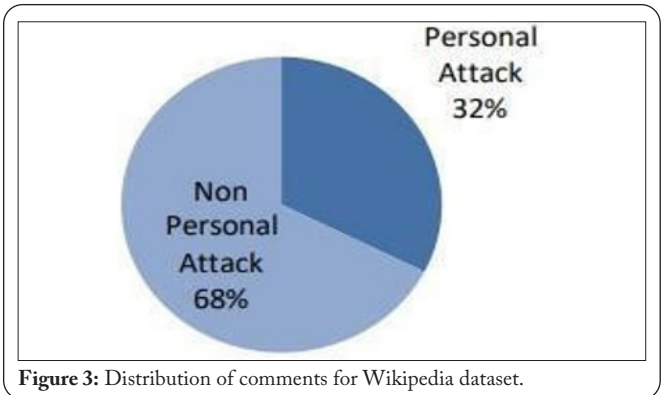

Figure 3: Distribution of comments for Wikipedia dataset.

Since classifiers are unable to use text in its original form to make classifications, the text must first be translated into numerical data before the classification process can begin. The process of removing particular features is referred to as "feature extraction," and it is a technique. The extraction of features is an essential part of the problems that are associated with natural language processing. The conversion of text into numerical data is accomplished through the use of feature extraction methods. Every piece of content—a tweet, for instance, or a comment—has the potential to be converted into a vector, which can then be put to use in the classification process. This study looks into two different aspects. The following are some methods for extracting features: In order to establish which approach yields the most reliable and accurate model, we have utilized two distinct strategies for feature extraction.

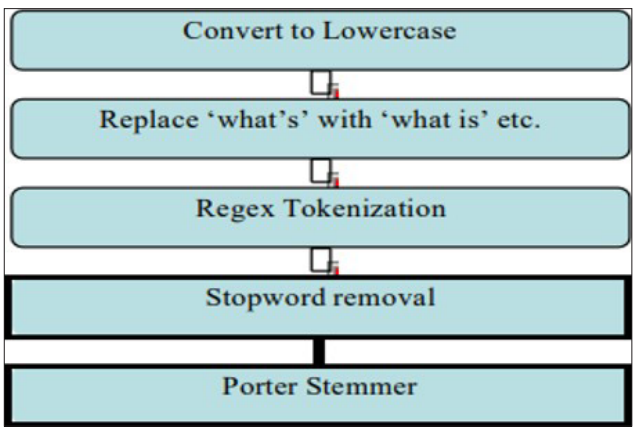

Figure 4: Data pre-processing pipeline.

Bag of words

TF-IDF vectorizer

It is a technique that isolates characteristics of texts based on the frequency with which particular words appear in a given document. The lexicon of words, which are typically referred to as tokens, are derived from every single document, which brings us to the first of this method's two essential and primary components. This model is referred to as a bag because it just maintains track of the occurrences of the words

and not the order in which they appear. The following is the technique that should be followed when using this model. The tokens, also known as the vocabulary, are produced from all of the inputs, also known as the paperwork. The lexicon may have all of the tokens found in all of the texts, or it may only contain some of the tokens that are used the most frequently. After the design of the vocabulary was completed, the next step was to alter the document by incorporating measurement characteristics that were dependent on the vocabulary. The most common use cases involve two distinct kinds: the first kind is a binary one, in which the characteristics are either true or false based on whether or not they are present. Taking the example into consideration, "simply good" and "extremely good" are not interchangeable terms. It will not take into account the sequence in which the words are presented, which might make a significant impact in certain circumstances. The process of building vocabulary can become difficult for large datasets due to the increased number of attributes.

TF-IDF model

The TF-IDF technique is somewhat comparable to the bag of words model in that it generates a vocabulary in a manner that is analogous to that of the model. TF-IDF rises as a function of, or in proportion to, the frequency of the document will drop in tandem with the decrease in the frequency of a phrase within the same text.

TF stands for term frequency. It does what its name suggests, which is to determine how frequently a given word appears in a given document. It is the ability to search for and locate a certain text phrase within a given document. The frequency of a single word (Wi) in a document (Rj) is calculated by dividing the total number of times that word appears in the text by the total number of words in the document.

tf (Wi, Rj) = (No. of times Wi appears in Rj)/(Total no. of documents in Rj)

The index document frequency (IDF) analyzes how often or infrequently a particular word appears in a given text. The value of the IDF is inversely related to the frequency of the word; that is, it is higher for words that are used less frequently and lower for terms that are used more frequently. The IDF is determined by dividing the total number of words contained in document D by the total number of words contained in a file that contains document D. The phrase t, and then computing the log value of the resulting.

$$IDF(d,D) = \log \frac{|D|}{\{d \in D : t \in D\}}$$

Since the denominator that denotes number of docs that contains term t cannot be 0 as it results in divided by 0 exception, it is added with 1 to ensure such cases are avoided and $|D|$ denotes no. of documents.

TF-IDF (t, d, D) = TF (t, d) x IDF (d, D)

Classification

A feature vector is generated as a result of the use of two distinct methods of feature extraction. The test data are modified first, then used, before the final data are used directly. The following classifiers will be put through training and testing with the help of the data from the train in order to identify which one is superior to the others.

SVM

The support vector classifier creates a map of a hyperplane in a space with (N) dimensions that differentiates between data points. For the purpose of optimizing the marginal values, a loss function known as the hinge function is applied. Due to the fact that the data can be organized linearly. The usage of linear SVM, which works well with these kinds of data, is recommended. assuming there was no error in categorisation, that is. When our model correctly predicts the class of data point, we do not need to apply the reduction with the gradient update regularization. If our model is incorrect, we apply the regularization to fix the error (Figure 5).

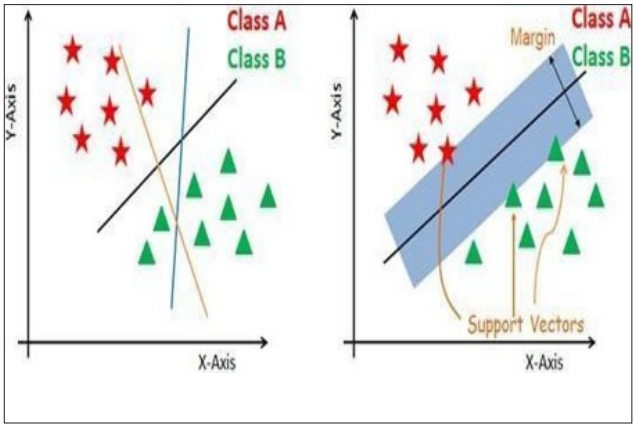

Figure 5: SVM classification.

Logistic regression

Logistic regression is a classification model that use the sigmoid function as a probabilistic function to provide the problem's needed output.

sig(x) = 1/{1 + exp(– x)}

A = LT + C

T(x) = sig(A)

Where L is weights resulted by the classifier, T is feature vector (input), The bias calculated by the classifier is C and T(x) is the hypothesis function. h(x) > 0.5 which can be basically described as a threshold function or value. If it If the value is more than 0.5, the class is 1, else the class it belongs to is 0. The sigmoid is useful for categorization since it sits between 0 and 1.

Random forest

A random forest, as its name suggests, is a forest that is made up of numerous trees. Similarly, this classifier is made up of a number of decision trees, each of which makes a prediction regarding a different category. The final prediction is obtained by taking the aggregate of all outputs predicted by each individual decision tree and calculating them for the given in-

puts. Another way to think of this is that the final prediction is the forecast that received the most votes from the decision trees. The future is foretold by decision tree through the application of a set of rules referred to as decision rules (Figure 6).

MLP

Artificial neural networks known as MLPs are constructed with three distinct layers: an input layer, a hidden layer, and an output layer. Back propagation is used for training purposes, whereas forward propagation is utilized for the purpose of calculating the activation value of each node. the many weights that are used in the neural network. The idea that data cannot be separated linearly is one that is used rather frequently. Activation functions based on the Relu or Sigmoid families could be used. It can be shown that the sigmoid function, which takes values between 0 and 1, is identical to the tanh function, which takes values between -1 and 1. Rectified linear unit can be calculated using the formula f(x) = max (0,x). It is possible to create and train MLP with the help of the Keras framework.

The MLP has been trained using a total of a thousand epochs, and the dropout rate was only 0.2 percent. The sigmoid function serves as the activation function, and Adam is the optimizer (Figure 7).

Bernoulli Naive Bayes

Training and classification using Naive Bayes for data that has been distributed using multivariate Bernoulli distribution techniques; that is, there may be a large number of features, but each one is a binary-valued (Bernoulli, Boolean) variable. As a consequence of this, samples have to be shown as feature vectors that contain binary values. Any other kind of data will have its format transformed into binary via a BernoulliNB instance (depending on the binarize parameter).

XGBoost

XGBoost is a gradient boosted trees implementation that is widely used and highly effective. It is a supervised machine learning technique that combines the predictions of a number of smaller models in an effort to get as accurate a forecast as possible.

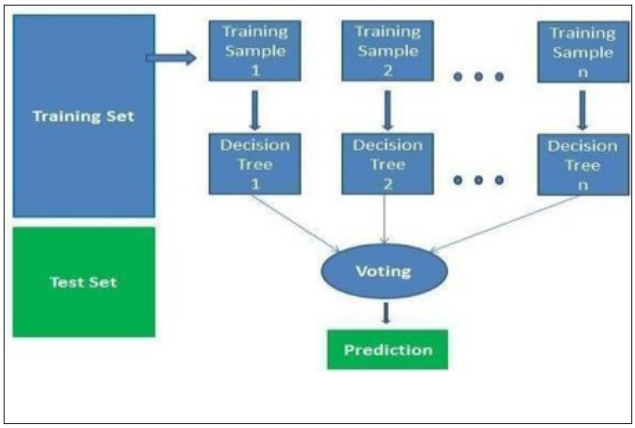

Figure 6: Random forest classification.

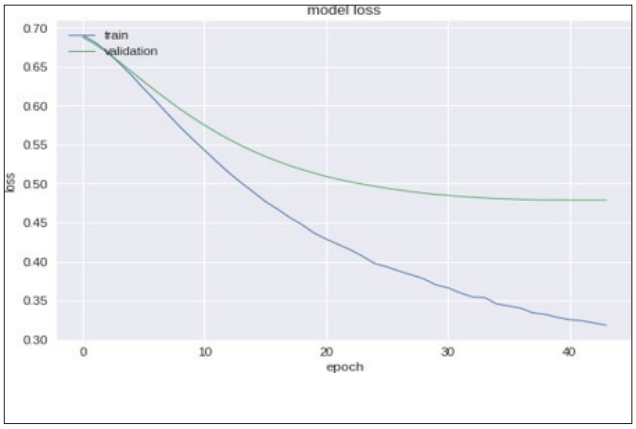

Figure 7: Model loss of MLP.

Results and Discussion

Performance evaluation

Google Colab was utilized throughout the process of training and validating the models. As a result of using a number of different classifiers, it is necessary to identify the one that most accurately completes the required task. The following criteria will be utilized in the evaluation.

Accuracy (A)

It is a metric that can be used to determine how accurate the prediction is; A = (True positives)/(Size of dataset).

Precision (P)

It is a metric that identifies real positives among all positives. P = (True positives)/(True positives + False positives).

Recall (R)

Positive predictions out of all positive inputs. R = (True positives)/(True positives + False negatives).

F-measure (F)

Used to compare precision and recall by computing the harmonic mean of precision. F = (2 x P x R)/(P + R).

The results that were produced through the use of the Bag of Words model. The results of evaluating numerous different classifiers using the Bag of Words model are summarized in table 1.

The figure 8 illustrates the degree of accuracy achieved by utilizing several classifiers in conjunction with the Bag of Words model.

Table 1: Comparison of classifiers based on the measures.

Measure	SVM	Random forest	XGBoost	Logistic regression	Naïve Bayes
Accuracy	0.95	0.97	0.73	0.93	0.87
Precision	0.96	0.78	0.78	0.94	0.78
Recall	0.95	0.74	0.74	0.93	0.74
F-Measure	0.96	0.81	0.71	0.93	0.71

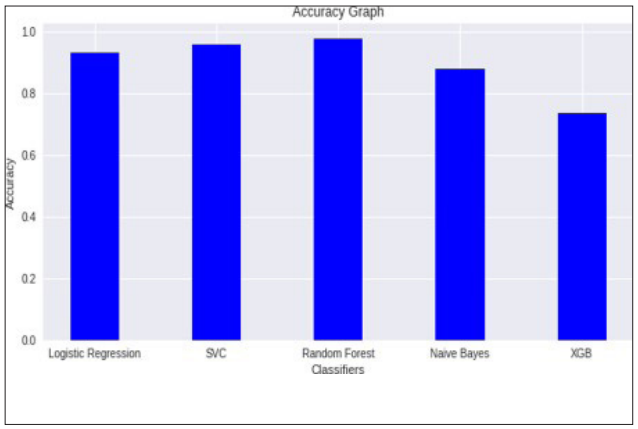

Figure 8: Accuracy graph of classifiers used with Bag of Words model.

The following are the findings achieved by utilizing the TF-IDF model: The table that follows illustrates every facet of the evaluation of the several classifiers that can be employed with the TF-IDF model. The accuracy of each classifier when combined with the TF-IDF model may be seen in figure 9.

For Wikipedia dataset, random forest classifier and the TF-IDF feature extraction technique are both utilized in this research due to the fact that their combination produced more precise findings. The report that classifies the aforementioned model is presented in figure 10.

Conclusion

The problem of cyberbullying in social media has grown to become a huge and severe issue that has to be addressed as quickly as possible in order to prevent people's lives from being negatively impacted by depression and other mental illnesses. Within the scope of this study, we propose a framework that has the potential to assist in the mitigation of this issue. The architecture that is being designed can identify offensive comments. When applied to problems requiring Natural Language Processing, strategies that were integrated with basic machine learning techniques produced results that were more accurate. The TF-IDF model was the more promising of the two feature extracting models, despite the fact that both of them produced results with similar degrees of accuracy. The fundamental model that has been developed to identify inappropriate comments made on social media networks looks somewhat like this. This concept may be implemented in the form of an extension that, when activated while using social networking websites, would emphasize both abusive and non-offensive comments. This portion of the extension can be utilized in future works to put an end to cyber persecution and establish a safe and non-hazardous atmosphere within the various social media websites.

Acknowledgements

None.

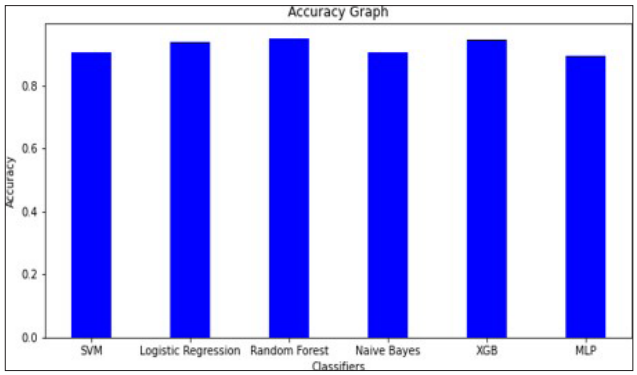

Figure 9: Accuracy graph of classifiers used with TF-IDF model.

	precision	recall	f1-score	support
0.0	0.68	0.79	0.73	17120
1.0	0.75	0.64	0.69	17640
accuracy			0.71	34760
macro avg	0.72	0.71	0.71	34760
weighted avg	0.72	0.71	0.71	34760

Figure 10: Wiki data random forest classifier with TF-IDF feature extraction model produce accurate results. Here's the model's categorization report.

Conflict of Interest

None.

References

1. Galán-García P, Puerta JGDL, Gómez CL, Santos I, Bringas PG. 2016. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. *Logic J IGPL* 24(1): 42-53. https://doi.org/10.1093/jigpal/jzv048
2. Chowdary MK, Turaka R, Alabduallah B, Khan M, Babu JC, et al. 2023. Low-power very-large-scale integration implementation of fault-tolerant parallel real fast fourier transform architectures using error correction codes and algorithm-based fault-tolerant techniques. *Processes* 11(8): 2389. https://doi.org/10.3390/pr11082389
3. Arulananth TS, Chinnasamy P, Babu JC, Kiran A, Hemalatha J, et al. 2023. Edge detection using fast pixel based matching and contours mapping algorithms. *PLoS One* 18(8): e0289823. https://doi.org/10.1371/journal.pone.0289823
4. Ali L, Azim MI, Ojha NB, Peters J, Bhandari V, et al. 2023. Integrating forecasting service and Gen2 blockchain into a local energy trading platform to promote sustainability goals. *IEEE Access* 12: 2941-2964. https://doi.org/10.1109/ACCESS.2023.3347432