INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework

P Vivekananth[1]*, **Navneet Sharma**[2]

1 PhD Research Scholar, Computer science, IIS Deemed to be University, Gurukul Marg, SFS, Mansarovar, Jaipur, India
2 Associate Professor, IIS Deemed to be University, Gurukul Marg, SFS, Mansarovar, Jaipur, India

OPEN ACCESS

Received: 03-07-2024

Accepted: 09-10-2024

Published: 18-02-2025

Citation: Vivekananth P, Sharma N (2025) Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework. Indian Journal of Science and Technology 18(5): 380-389. https://doi.org/10.17485/IJST/v18i5.1491

* **Corresponding author**.

Vivekanandhan2024@outlook.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To accomplish better cyberbullying classification accuracy through Modified Term Frequency and Inverse Document Frequency (MTF-IDF) with Machine Learning (ML) technique. The cyberbullying classification is improved by modifying the parameter by hypertuning the concept in MTF-IDF using the Optuna method. **Method:** To categorize the bullying text, this study discusses the Ensemble model that integrates MTF-IDF and ML methods with sophisticated feature extraction approaches. This research has considered 47692 tweets along with the label of cyberbullying classes, and 1000 tweets is the sample size considered for testing proposed MTF-IDF with various ML algorithms. From the text data, MTF-IDF with ML has assisted in identifying the cyberbullying feature patterns for feature extraction by hyperparameter tuning the MTF-IDF method using Optuna technique for the accurate classification of cyberbullying tweets. **Findings:** In text classification, MTF-IDF with ML has assisted in identifying the cyberbullying feature patterns for feature extraction techniques by hyperparameter tuning the MTF-IDF method by Optuna technique for the accurate cyberbullying tweets classification. The proposed hyperparameter-tuned MTF-IDF is evaluated with traditional Natural Language Processing (NLP) methods by confusion matrix metrics like accuracy and balanced accuracy as 83.63% and 83.42%, respectively which are seen to be high when compared to traditional NLP methods. **Novelty:** The proposed framework focuses on cyberbullying detection in social media more precisely by text mining sentimental words using Modified NLP methods with hyperparameter tuning and improving the classification t using Optuna with various ML methods and modified entropy concept of TF-IDF as Modified TF-IDF (MTF-IDF).

Keywords: Social Media; Cyberbullying; Term Frequency Inverse Document Frequency (TF IDF); Natural Language Processing; Machine Learning; Sentiment Analysis (SA); Optuna

1 Introduction

Modern communication and information-sharing methods have been replaced by social media networks. Users may interact, discuss, and contribute to specific issues while exchanging ideas and opinions on social media platforms like Twitter by sending brief updates or tweets that are limited to 140 characters. According to a study, the examination of publicly available data on social media may be utilized to analyze transformations in people's beliefs, actions, and psychological states as more individuals participate in these platforms [1]. Thus, sentiment analysis utilizing Twitter data has gained popularity. Cyberbullying is one of the main issues preventing the current society from growing sustainably and growing at a quicker rate than the rest of them. Psychological distress and various aspects of an individual's life may be impacted by cyberbullying.

Sentiment Analysis (SA) approaches

According to a research study, the incorporation of sentiment evaluation with ambiguous management can account for more precise outcomes and provide extensive descriptions of emotions, including happiness, enthusiasm, rage, sorrow, and anxiety [2]. However, SA is typically performed on text data. The application of NLP for a variety of purposes is made possible by multimodal SA, which takes text-based analysis to a higher level. However, research in other fields, such as neural networks, is propelling the fast advancement of NLP [3]. A study cites the use of Neurosymbolic AI, which blends symbolic reasoning and deep learning, as a promising approach in NLP for knowledge interpretation [4].

Source detection and rumor propagation

A study utilized the TF-IDF to extract features from a dataset of 2000 tweets, comparing the performance of five classifiers and determining that logistic regression is the best classifier for detecting bullying tweets [5], and another study used TF-IDF and LGBM algorithms are employed to identify cyber-attacks in darknet traffic, achieving a high accuracy of 98.97% compared to other algorithms [6]. On the other hand, a study employs two machine learning-based classifiers to analyze a large-scale Twitter dataset, revealing negative sentiment across various themes [7]. Research has identified that the sensitive keywords that could lead to vulnerability when combined with benchmarked cyber keywords using LDA, when analyses of COVID-19 misinformation spread on Twitter through sentiment, emotion, topic, and user attributes, highlights the denial of the pandemic and dissemination of false information [8,9]. NLP techniques have been increasingly developed. Among them, there are processing techniques such as tokenization to divide sentences or paragraphs into smaller ones, stop word removal from removing words that have no critical meaning, stemming used to remove suffixes from meaningless words, lemmatization to return words to their infinitive form (lemma), Part-of-Speech (PoS) tagging to classify each word in a sentence into word type parts, Named Entity Recognition (NER) to recognize and classify named objects, sentiment analysis to analyze emotions in text, topic modeling to find the main themes in a corpus, word embeddings to convert words into vectors [10]. Data processing techniques in NLP are commonly used in text classification, automatic translation, chatbots, and many other fields. The infection graph is then either separated into smaller sections, each containing a single source, or is analyzed as a single portion. The presumed sources with the highest significant value are deemed to be actual. This article examines the first strategy [11].

Research Gap

A study indicates that even cyberbullying can influence people to consider suicide and self-harm, especially if they are younger. This study's primary objective is to automatically identify instances of cyberbullying from chaotic posts. Previous investigations and statistical analyses of cyberbullying have mostly focused on definitions, impacts of cyberbullying, and statistical methodologies. ML and DL techniques help researchers better understand human behavior in factor of computing studies. Recognition of cyberbullying has been seen as an NLP task. According to a study, scientists have used Support Vector Machines (SVM) as a more generic classifier and the Bag-of-Words (BoW) approach as an improved feature extraction technique [12]. Conventional ML approaches are widely used in identifying problematic kinds of human behavior, but automation learning falls behind in predicting cyberbullying more accurately due to a lack of text classification. Hyperparameters critically influence how well machine learning models perform on unseen, out-of-sample data. Systematically comparing the performance of different hyperparameter settings will often go a long way in building confidence about a model's performance [13]. In recent years, there have been notable advancements in the field of cyberbullying detection, with ensemble learning being a key factor. Using a variety of deep learning techniques, including CNN, LSTM, and Conv1DLSTM, has shown to be successful at identifying instances of cyberbullying on social media sites like Facebook and Twitter. The performance of this method is based on two distinct datasets. In terms of accuracy, precision, recall, F1-score, and detection time, the suggested stacked ensemble learning techniques have demonstrated encouraging outcomes, proving their capacity to recognize and categorize derogatory

language on social media. Still, there is a long way to go before creating a more reliable and accurate ensemble learning method for cyberbullying detection. The shortcomings of existing approaches are their restricted generalizability highlighting the necessity of additional research in this area [14]. The suggested method made use of traditional machine learning models with hyperparameter tuning that gain knowledge from uniform training sets. That isn't always the situation with the transmission system's non-stationary behavior. Furthermore, it was assumed that the simulation faults were stationary, which may not be true for real platforms which are subject to aging and physical influences. Utilizing smart techniques with the capacity to continually understand real-time data is necessary to address such difficulties. The defects databases might be updated live using incremental learning [15]. Thus, this research concentrates on identifying the frequent bullying words and the activity of words.

Need for Study

However, previous research has worked on TF-IDF and hyperparameter tuning in NLP with other ML techniques but individually, to detect and classify cyberbullying texts, and these traditional methods have failed to provide accuracy in their techniques. Hence, to accomplish higher accuracy and precision in detecting cyberbullying texts, a modified version of TF-IDF is proposed with hypertuned using Optuna methods, which cannot be found in the previous literature. The research aims to compare the achieved accuracy using the modified methods with the traditional word2vec techniques and hypothesizes to bridge the gap by achieving higher accuracy through modified techniques to classify cyberbullying texts and ensure earlier prediction. Thus, the research focuses on

- To improve cyberbullying detection by text mining sentimental words using NLP libraries and modifying the entropy of term frequency present in the TF-IDF technique.
- To create a hybrid model by integrating hyperparameter tuning of the ML method with MTF-IDF for generating a more accurate prediction of cyberbullying classification detection.

2 Methodology

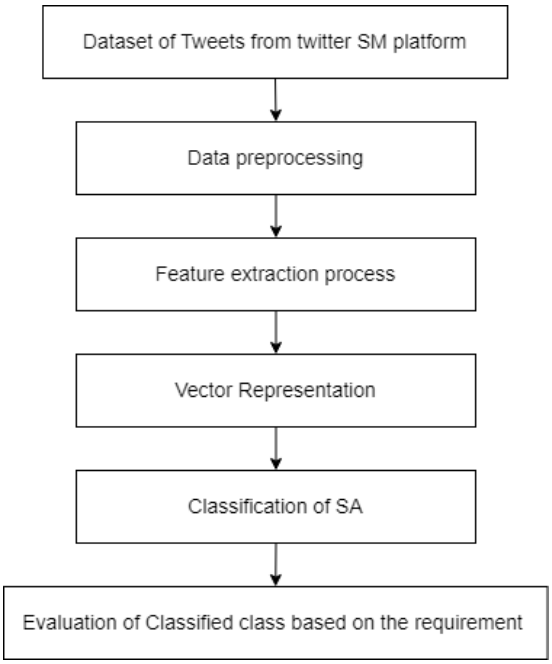

Fig 1. Basic Architecture of ML-Based Method

Feature representation

A study suggests the popular techniques for representing text features fall into two categories: frequency-based embeddings like TF–IDF, Count vector, and Hashing Vectorizer, as well as pre-trained word embeddings such as Word2Vec, Bert, and Glove[16]. The present study primarily employs three feature representation models:

1. The term frequency–inverse document frequency (TF–IDF) assesses a word's value in the tweet dataset and gauges its relevance to the text as a whole. TfidfVectorizer () in Python may produce a TF–IDF matrix by multiplying each word's inverse document frequency metric by its frequency in clean tweets.
2. In Word2Vec, every word is represented as a vector in a vector space that is created based on the whole corpus of tweets. This technique works better for handling semantic connections since words that have identical meanings are going to be closer together in the vector space.

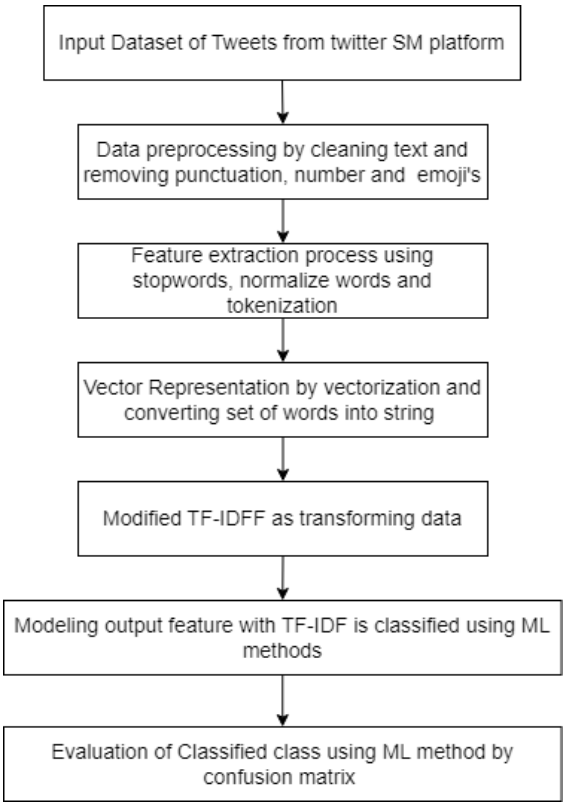

Fig 2. Proposed architecture of hybrid text classification methods

Research Methodology

This research focuses on developing automated cyberbullying detection to detect offensive activities for users on Twitter (social media). It acts as the treat or mislead comments once it gets detected and classified through a hybrid model of DL and ML, such as MTF-IDF with ML techniques. This technique can be applied in the future for a large number of texts that may stream from live tweets. The flow diagram of the proposed hybrid text classification method is shown in Figure 2.

Dataset collection and data preprocessing

The majority of citizens depend on SM as it gets more and more common across all age groups for daily contact. The data collection has to try to expose the cyberbullying issues, and on April 15th, 2020, a warning was declared by UNICEF concerning problem for the growing risk of cyberbullying during the COVID-19 pandemic because of closures of widespread schools that

increase screen time, as well as decreases social interaction with face-to-face. There is an alarm for the issue of cyberbullying through statistics that 36.5% of school students have found to be cyberbullied, and nearly 87% of students in middle and high school students were observed that cyberbullying has impacted the victim has caused degrading performance of academic and even depression to suicidal thoughts. This research has considered 47692 tweets along with the label of cyberbullying classes as the cyberbullying type shown in Figure 3.

tweet_text	cyberbullying_type
In other words #katandandre, your food was crapilicious! #mkr	not_cyberbullying
Why is #aussietv so white? #MKR #theblock #ImACelebrityAU #today #sunrise #studio10 #Neighbours #WonderlandTen #etc	not_cyberbullying
Ian Shit Ain't It The Truth . I Hate You 2 !	other_cyberbullying
@XochitlSuckkks a classy whore? Or more red velvet cupcakes?	not_cyberbullying
@Jason_Gio meh. :P thanks for the heads up, but not too concerned about another angry dude on twitter.	not_cyberbullying
I want to just slap that smug look off Kats face. #annoying #mkr @mykitchenrules	gender
@RudhoeEnglish This is an ISIS account pretending to be a Kurdish account. Like Islam, it is all lies.	not_cyberbullying
@Raja5aab @Quickieleaks Yes, the test of god is that good or bad or indifferent or weird or whatever, it all proves gods existence.	not_cyberbullying
@PressTV Like every Muslim country, they find money for weapons, but they don't have the money for schools.	religion
Itu sekolah ya bukan tempat bully! Ga jauh kaya neraka	not_cyberbullying
Karma. I hope it bites Kat on the butt. She is just nasty. #mkr	not_cyberbullying
@stockputout everything but mostly my priest	not_cyberbullying
Rebecca Black Drops Out of School Due to Bullying:	not_cyberbullying
@Jord_Is_Dead http://t.co/UsQInYW5Gn	not_cyberbullying
The Bully flushes on KD http://twitvid.com/A2TNP	not_cyberbullying
Ughhhh #MKR	not_cyberbullying
RT @Kurdsnews: Turkish state has killed 241 children in last 11 years http://t.co/JlvkE1epws #news ##GoogleÃ‡eviriciTopluluÄŸuKÃ¼rtÃ§eyideEâ€¦	not_cyberbullying
Love that the best response to the hotcakes they managed to film was a non-committal "meh" from some adolescent. #MKR	not_cyberbullying
@yasmimcaci @Bferrarii PAREM DE FAZER BULLYING COMIGO =(UHAHUAH BANDO DE PRETO	not_cyberbullying
@sarinhacoral @Victor_Maggi tadinhu de mim , sofrendo bulling viu MIMI'	not_cyberbullying
@0xabad1dea @kelseytheodore2 twitter is basically the angry letters of our generation.	not_cyberbullying
Best pick up line? Hi, you're cute... ?: I love how people call James Potter is a bully. - mypatronusisyou: http://tumblr.com/xol3xl14zy	not_cyberbullying
Now I gotta walk to classss?! I officially hate the stupid bus system! - _-	not_cyberbullying

Fig 3. Dataset of tweets and its cyberbullying classes

Data preprocessing is initiated by cleaning the tweet text by removing the links, removing punctuations, and removing numbers using replace functions. The emoji is removed through deemojify function and the remaining content of the tweet is considered for the normalization process. Before normalization, the label encoder is utilized for categorizing the cyberbullying type as classes and the count is listed in Table 1.

Table 1. Index and count for various cyberbullying types

Cyberbullying Type	Class Index	Count
Age	0	7992
Ethnicity	1	7961
Gender	2	7973
Not Cyberbulling	3	7945
Other cyberbullying	4	7823
Religion	5	7998

During data normalization, the stopwords function is used to generate a list of words as stop words, whereas the words will not reflect in the clean text. Once stop words are done, the tokenization process is progressed through word-based in which the sentence pattern of each tweet is converted into a set of words. Normalizing words has assisted in modifying the plural words into singular words and vectorizing the words using a count vectorizer with maximum features of 5000. Finally, the MTF-IDF is utilized for the transformation of the list of words into strings and to classify the cyberbullying words quite easily by improving the TF-IDF through weight-based feature words.

Working principle of TF-IDF

The characteristics of a word's weight in the technique of TF-IDF are determined by the frequency in the document set. The feature word has an excellent ability to describe a group whenever it appears in numerous papers in that category. Strong words ought to be assigned greater significance than the feature word. No matter the number of times a feature word is present in a document set since it is prevalent in all categories, it ought to be assigned a lower weight. The frequency of a word is then multiplied by the frequency of the inverse document to determine the feature word's weight. It is the feature word's frequency in the class document, while the inverse document frequency is the feature word's distribution over several categories. The formula for the TF-IDF algorithm consists of two parts namely TF and IDF. The phrase "word frequency," often abbreviated as "TF," refers to the frequency of words in a class. The cross-entropy of feature word likelihood density is shortened to IDF. The

TF algorithm's calculation is shown in the Equation (1).

$$W = C_1 * \log\left(\frac{T}{t_l}\right) + \gamma \tag{1}$$

Where,

Wi = the weight result obtained by the feature word

Ci = the occurrences amount of the feature in the document

T = Total number of files

ti = the of documents number in which the feature word has appeared

c = empirical value typically 0.01.

The process of normalization is given as per Equation (2).

$$W(x_i) = \frac{W_i}{\sqrt{\sum_{j=1}^{t} W_j^2}} \tag{2}$$

In the same manner, shorter documents have fewer feature words overall and fewer appearances of each word than longer documents. Even though it has a lesser weight, it has a very strong categorization capacity for these terms. The normalized form merits are independent of document length, the procedure yields measurement with a "fairer" assessment of the TF value. According to the TF-IDF algorithm's calculation method, the feature word frequency or weight and its inverse document frequency are the two variables used to determine the feature word's weight. Consequently, the feature word's weight might partially indicate the feature word's distribution throughout the classifications and more accurately reflect the feature word's occurrence information in the document set when calculated using the algorithm formula. Nevertheless, the classification accuracy remains low while classifying the model's data using the actual training phase of the aforementioned technique, and the algorithm may continue to be refined further.

However, the frequency of the feature words also affects the weight computation. The IDF calculation outcome remains the same, though because the IDF portion of the method only looks at occurrences and ignores the frequency with which feature words occur in this category. The relationship among feature words and the text number in which they occur has been employed by the TF-IDF algorithm in the present research. However, it overlooks the variation in feature word distribution across classifications, and that is detrimental to improving classification accuracy. Hence, it employs the TF-IDF algorithm and lots of methods to get better. This paper offers an enhanced IDF technique that uses the current discipline classification as a guide to distinguish between feature words that occur in various fields at different frequencies and have distinct meanings. To increase the trust in the TF-IDF algorithm's output, differentiate between the feature word's meanings in various subjects by using the current subject categorization as a point of reference. The formula for IDF is

$$IDF = -\lg\left(1 - \frac{F(m_i)}{F(m_i) + F(O_i)}\right)$$

$$IDF = \lg\left(1 + \frac{F(m_i)}{F(O_i)}\right) \tag{3}$$

$F(m_i)$ = Feature word likelihood (i) being in class m

$F(O_i)$ = Feature word likelihood of (i) existing in further document categories.

The classification that remains after removing class m from the document is referenced by other classes. The two feature words for an algorithm have an identical number of appearances in the document but since the distinct feature words have different distributions, F(mi) and F(Oi) differ, as do their IDF values. The TF-IDF algorithm has been tweaked to enhance its ability to differentiate among various distributions of feature words.

The hyperparameter involved in the optuna technique for the best hybrid model MTF-IDF and the objective function for the optuna is defined as multiclass. The basic parameters are as follows

- Boosting_type: The boosting type is gradient boosting decision tree (gbdt) and it performs as the feature of LightGBM is that it supports a number of different boosting algorithms.
- Learning_rate: The callbacks parameter of the fit method to shrink/adapt the learning rate in training using reset_parameter callback.

- num_leaves: This is the main parameter to control the complexity of the tree model. The reason is that a leaf-wise tree is typically much deeper than a depth-wise tree for a fixed number of leaves. Unconstrained depth can induce over-fitting.
- max_depth: This parameter can be used to limit the tree depth explicitly.
- min_child_weight: Minimum sum of instance weight (hessian) needed in a child (leaf).
- reg_alpha: L1 regularization term on weights.
- reg_lambda: L2 regularization term on weights.

Algorithm of MTF-IDF with hyperparameter

Input: Vectorized words from tweet sentences

Output: Matrix (X) with literacy text classification

Step 1: Initializing TF-IDF matrix 'm' and each word i with j of the matrix.

Step 2: Loop the following process by calculating the frequency $F(m_i)$ of the word I concerning the subjected classification of cyberbullying.

Step 3: Frequency $F(O_i)$ is calculated concerning the word (i) classified with others subjected to cyberbullying type.

Step 4: Frequency $F(tf)$ is calculated concerning the word for any cyberbullying type literature.

Step 5: Updating each value of matrix X formulae using Equation (2)

$$X(i,j) = F(tf) * \log\left(1 + \frac{F(m_i)}{F(O_i)}\right)$$

Step 6: The iteration of a matrix with updated formulae gets terminated, once the loop gets terminated.

Step 7: Optuna is defined with direction, n_trails and various hyperparameter values.

Step 8: The best hyperparameter is identified and executed with the best model of the cybersecurity classification.

The basic text classification currently relies on the hyperparameter tuning of MTF-IDF approach and integration of the suggested MTF-IDF method with the standard text classification procedure. The information is gathered mostly from literary writings. To reduce the tweet content size as much as possible, this study focuses on brief literary data from texts. Using the hyperparameter tuning of the MTF-IDF method described in this study, the feature weight is determined following data preprocessing and normalization. Develop a text classifier for obtaining the classification outcome, and input the feature data that has been decreased in dimension into the classifier.

3 Results and Discussion

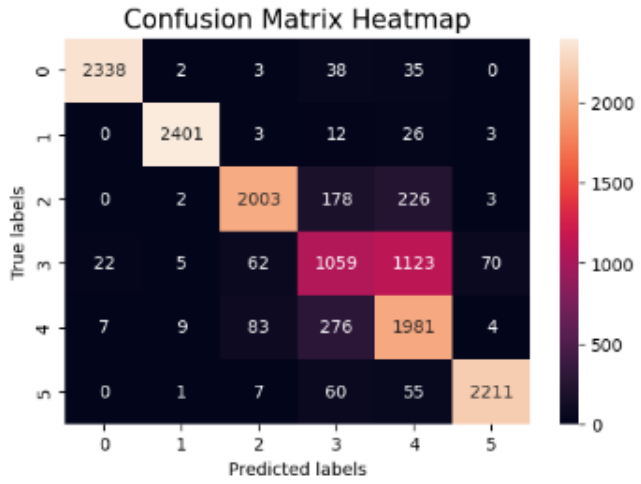

Fig 4. Confusion matrix heatmap for hyperparameter tuned MTF-IDF with LGBM Classifier

The classification method performance is usually assessed by the outcome of the model's prediction in the Light Gradient Boosting Machine (LGBM) classifier, Extreme Gradient Boosting (XGB) classifier, Extra Tree Classifier (ETC), Random Forest

Classifier (RFC), Bagging Classifier, Decision Tree Classifier (DTC), k Neighbor Classifier (KNC), Bernoulli Navies Bayes (BNB).

Hence, the evaluation metrics of MTF-IDF with a given ML model are considered to be best when compared to other ML models. The prediction of detecting cyberbullying in social media is evaluated through confusion matrix metric accuracy and the confusion metric of MTF-IDF with the LGBM classifier. Figure 4 illustrates the confusion matrix heatmap for MTF-IDF with LGBM classifier using hyperparameter tuning in which the True positive has increased due to modification of diagonal values for all classes from 0 to 5.

Table 2 illustrates the ML classifier with MTF-IDF involved with accuracy in which the Tuned LGBM classifier with MTF-IDF has high accuracy with 83.82% followed by 83.63%, 83.32%, and 82.36% from the LGBM classifier with MTF-IDF, XGB classifier with MTF-IDF, ETC with MTF-IDF respectively. The least accuracy is identified in KNC which has very poor prediction in cyberbullying detection over social media. In contrast, Tuned MTF-IDF has high accuracy with LGBM as a high prediction in cyberbullying detection.

Table 2. Shows the metrics of various ML classifiers with MTF-IDF

ML classifier with MTF-IDF	Accuracy(%)	Balanced Accuracy (%)	Time complexity (Sec)
LGBM Classifier	83.63	83.47	6.09
XGB Classifier	83.32	83.15	6.43
ETC	82.36	82.18	147.91
RFC	82.32	82.14	78.52
DTC	79.45	79.25	3.98
KNC	27.22	27.22	0.005
BNB	82.05	81.93	0.018
Hyperparameter-tuned LGBM Classifier	83.82	83.65	7.025

Similarly, the classification method performance is usually assessed by the outcome of the model's prediction in Light Gradient Boosting Machine (LGBM) classifier, Extreme Gradient Boosting (XGB) classifier, Extra Tree Classifier (ETC), Random Forest Classifier (RFC), Bagging Classifier, Decision Tree Classifier (DTC), k Neighbor Classifier (KNC), Bernoulli Navies Bayes (BNB). Hence, the evaluation metrics of word2vec with a given ML model are considered to be best when compared to other ML models. The prediction of detecting cyberbullying in social media is evaluated through confusion matrix metric accuracy of word2vec with LGBM classifier.

Table 3. M etrics of various ML classifiers with Word2vec

ML classifier with Word2vec	Accuracy (%)	Balanced Accuracy (%)	Time complexity (Sec)
LGBM Classifier	79.03	78.86	2.48
RFC	77.56	77.37	32.59
ETC	77.23	77.02	5.09
Bagging classifier	75.64	75.43	39.46
KNC	75.45	75.26	0.34
DTC	70.59	70.38	6.73
BNB	69.61	69.48	0.089
Hyperparameter tuned LGBM Classifier	79.63	79.45	1.622

Table 3 illustrates the ML classifier with Word2vec involved with accuracy in which the tuned LGBM classifier with Word2vec has high accuracy with 79.63% followed by 79.03%, 77.56% and 77.23% from LGBM classifier with Word2vec, RFC classifier with Word2vec, ETC classifier with Word2vec respectively. The lowest accuracy 69.61% is identified in BNB which has very poor prediction in cyberbullying detection over social media. In contrast, Word2vec has high accuracy with tuned LGBM has high prediction in cyberbullying detection. The ML classifier with Word2vec is involved with balanced accuracy in which the tuned LGBM classifier with Word2vec has high accuracy with 79.45% followed by 78.86%, 77.37% and 77.02% from RFC, ETC classifier respectively. The lowest accuracy of 69.48% is identified in BNB which has very poor prediction in cyberbullying detection over social media.

Table 4. Previous studies on cyberbullying

Researcher Name	Study period	Study Participant	Study findings	Model used	Evaluation metric (Accuracy %)
Usman Naseem et al.[17]	February to March 2020	70,000	COVIDSENTI-A, COVIDSENTI-B, and COVIDSENTI-C	Fast Text with RF	82.30
Nikitha GS. et al.[18]	2021	15,307	signify toxic (offensive) and non-toxic (non-offensive)	TF-IDF with Adaboost	82.80
Md. Habeeb Ur Rahman et al.[19]	2022	26,835	Labeled as offensive and Non-offensive	NLP(TF-IDF) and Word2Vec feature extraction with SVM	78.92
Aditya Desai et al.[20]	2021	852	Bullying and Non-Bullying	BERT model	70.89

Table 4 discusses the various research study in detection ofcyberbullying in social media in which different model approaches have been done with less accuracy which influences the understanding pf various NLP with ML and DL models. Hence, this research concentrated on modifying the TF-IDF as text classification with ML classifier for better detection of cyberbullying in social media.

Table 5. Performance accuracy of cyberbullying detection classifier

Cyberbullying detection classifier	Accuracy (%)
MTF-IDF with LGBM classifier	83.62
Hyperparameter-tuned LGBM Classifier with MTF-IDF	83.82
Naive Bayes (NB)[16]	71.25
Support Vector Machine (SVM)[16]	52.70
Adaboost[17]	82.80

Table 5 illustrates the hyperparameter-tuned LGBM classifier with MTF-IDF has better cyberbullying detection with 83.82% which is comparatively higher than MTF-IDF with LGBM classifier and traditional ML algorithms like NB, SVM andAdaboost are 83.62%, 71.25%, 52.7% and 82.80%respectively. The hypertuned LGBM classifier with MTF-IDF get saturated with a learning rate of 0.001. Moreover, the accuracy of the hypertuned model gets decreased after increasing the learning rate by more than 0.001. Hence,the result of TF-IDF based classification from the previous research has been considered for evaluation of proposed MTF-IDF with LGBM classifier in which TF-IDF based SVM[16], TF-IDF based NB[16] and Adaboost[17]. The performance outcome determines that MTF-IDF with LGBM classifier has better accuracy in detecting cyberbullying in social media with 83.62% accuracy.

4 Conclusion

To classify literary texts, this study suggests the tuned MTF-IDF method for the feature words problem that appears in multiple disciplines with varied frequencies as well as meanings across multiple disciplines. The current discipline classification serves as a guide for differentiating traits. The purpose of word meanings in different fields is to increase trust in the output of the TF-IDF algorithm. In comparison, the experiments on detecting cyberbullying by selecting features using text classification by hyperparameter tuned modified entropy of TF-IDF as proposed MTF-IDF improves not only the text classification performance through accuracy as 83.82% which is higher compared to traditional NB and SVM method and also the ability to generalize and reliability of the algorithm. In future work, the text classification is focused on the hybrid DL method in improving cyberbullying detection of Twitter tweets. Furthermore, this research concentrates on emoji detection and focuses on developing an application that will benefit parents and educators in monitoring their children from cyberbullying activities. Hence, cyberbullying AI can be trained using this bullying detection application as a future development.

Acknowledgements

The authors gratefully acknowledge the help and support provided by the teachers and institutions that made this project a success.

References

1) Alamoodi AH, Zaidan BB, Zaidan AA, Albahri OS, Mohammed KI, Malik RQ, et al. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert systems with applications*. 2021;167. Available from: https://doi.org/10.1016/j.eswa.2020.114155.
2) Wang Z, Ho SB, Cambria E. Multi-level fine-scaled sentiment sensing with ambivalence handling. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 2020;28:683–697. Available from: https://doi.org/10.1142/S0218488520500294.
3) Ray P, Chakrabarti A. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Applied Computing and Informatics*. 2022;18:163–178. Available from: https://doi.org/10.1016/j.aci.2019.02.002.
4) Sarker MK, Zhou L, Eberhart A, Hitzler P. Neuro-Symbolic Artificial Intelligence: Current Trends. *arXiv*. 2021;34. Available from: https://arxiv.org/abs/2105.05330.
5) Shah R, Aparajit S, Chopdekar R, Patil R. Machine Learning based Approach for Detection of Cyberbullying Tweets. *International Journal Computer Application*. 2020;175(37):51–56. Available from: https://www.researchgate.net/publication/347736853_Machine_Learning_based_Approach_for_Detection_of_Cyberbullying_Tweets.
6) Rawat R, Mahor V, Chirgaiya S, Shaw RN, Ghosh A. Analysis of Darknet Traffic for Criminal Activities Detection Using TF-IDF and Light Gradient Boosted Machine Learning Algorithm. In: Mekhilef S, Favorskaya M, Pandey RK, Shaw RN, editors. Innovations in Electrical and Electronic Engineering;vol. 756. Singapore. Springer. 2021. Available from: https://link.springer.com/chapter/10.1007/978-981-16-0749-3_53.
7) Pattnaik N, Li S, Nurse JRC. Perspectives of non-expert users on cyber security and privacy: An analysis of online discussions on Twitter. *Computers & Security*. 2023;125. Available from: https://doi.org/10.1016/j.cose.2022.103008.
8) Lanier HD, Diaz MI, Saleh SN, Lehmann CU, Medford RJ. Analyzing COVID-19 disinformation on Twitter using the hashtags #scamdemic and #plandemic: Retrospective study. *PLoS ONE*. 2022;17(6). Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9216575/.
9) Geetha R, Karthika S. Sensitive Keyword Extraction Based on Cyber Keywords and LDA in Twitter to Avoid Regrets. In: International Conference on Computational Intelligence in Data Science;vol. 578. Springer. 2020;p. 59–70. Available from: https://link.springer.com/chapter/10.1007/978-3-030-63467-4_5.
10) Rossetti G, Milli L, Cazabet R. CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*. 2019;4. Available from: https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0165-9.
11) Shelke S, Attar V. Source detection of rumor in social network- A review. *Online Social Networks and Media*. 2019;9:30–42. Available from: https://doi.org/10.1016/j.osnem.2018.12.001.
12) Nguyen D, Liakata M, Dedeo S, Eisenstein J, Mimno D, Tromble R, et al. How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in artificial intelligence*. 2020;3. Available from: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2020.00062/full.
13) Arnold C, Biedebach L, Küpfer A, Neunhoeffer M. The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*. 2024;12(4):841–848. Available from: https://doi.org/10.1017/psrm.2023.61.
14) Muneer A, Alwadain A, Ragab MG, Alqushaibi A. Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. *Information*. 2023;14(8). Available from: https://doi.org/10.3390/info14080467.
15) Bhattacharya D, Nigam MK. Energy efficient fault detection and classification using hyperparameter-tuned machine learning classifiers with sensors. *Measurement: Sensors*. 2023;30. Available from: https://doi.org/10.1016/j.measen.2023.100908.
16) Al-Garadi MA, Hussain MR, Khan N, Murtaza G, Nweke HF, Ali I, et al. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access*. 2019;7:70701–70718. Available from: https://ieeexplore.ieee.org/document/8720155.
17) Naseem U, Razzak I, Khushi M, Eklund PW, Kim J. COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE transactions on computational social systems*. 2021;8(4):1003–1015. Available from: https://ieeexplore.ieee.org/document/9340540.
18) S NG, Shenoyy A, Chaturya K, C LJ, M JS. Detection of Cyberbullying Using NLP and Machine Learning in Social Networks for Bi-Language. *International Journal of Scientific Research & Engineering Trends*. 2024;10(1):153–161. Available from: https://ijsret.com/wp-content/uploads/2024/01/IJSRET_V10_issue1_128.pdf.
19) Rahman MHU, Divya M, Reddy BR, Kumar KS, Vani PR. Cyberbullying Detection using Natural Language Processing. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2022;10(5):5241–5248. Available from: https://www.ijraset.com/best-journal/cyberbullying-detection-using-natural-language-processing.
20) Desai A, Kalaskar S, Kumbhar O, Dhumal R, ITM Web Conf . Cyber Bullying Detection on Social Media using Machine Learning. In: and others, editor. International Conference on Automation, Computing and Communication 2021 (ICACC-2021);vol. 40. EDP Sciences. 2021;p. 1–5. Available from: https://doi.org/10.1051/itmconf/20214003038.