ISSN Online: 3006-4708
ISSN Print: 3006-4694

SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com

Monitoring Student Well-being: Using AI to Detect Offensive Language in Educational Platforms

Fadia Shah[1], Yasir Shah[2], Faiza Shah[3], Imran Shahid[4], Aftab Hussain Tabasam[5]

[1] Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan.
Email: fadiashah13@yahoo.com
[2] School of Business, Zhengzhou University, Zhengzhou 450001, China.
Email: yasirshah_pk@yahoo.com
[3] School of Political Science and Public Administration, Henan Normal University, China.
Email: faizashah55@gmail.com
[4] School of Political Science and Public Administration, Henan Normal University, China.
Email: imranzzu87@gmail.com,
[5] Business Administration, University of Poonch Rawalakot. Email: aftabtabasam@upr.edu.pk

DOI: https://doi.org/10.70670/sra.v3i4.1260

Abstract

Because of the quickly developing communication modes of individuals in informal communities, individuals bought into these interpersonal organisations at an extraordinary rate to convey and impart their contemplations to different supporters. Twitter was chosen for this study because of its notoriety and simple admission to information. This review proposes a strategy to recognise oppressive substance in Twitter information that contains tweets, retweets, and remarks. Oppressive language is arranged in view of elements through the component extraction interaction and classes, either harmful or not. The motivation behind this significant advance is to provide the advantages and disadvantages of each approach, which will be useful for dealing with significant stages. While growing new systems or strategies to distinguish harmful substances in the client content of informal organisations. Furthermore, this correlation provides additional information on which system is suitable for determining the degree of disagreeableness to further develop exactness.

Keywords: Machine Learning, Student Performance, Abusive Languages, Text Mining, Feature extraction, Sentiment analysis

Introduction

In the advanced age, informal organisation utilisation is expanding step by step, and individuals are quickly drawn toward informal organisations. The number of students on famous interpersonal organisations, such as Twitter, Facebook, and LinkedIn, is in millions or billions [1, 2], and the number of students in all informal communities has expanded from 2010 to 2020. Because of students' interest in these stages, they invest energy in getting a charge out of and speaking with one another. A portion of the students in informal organisations perform illicit activities by utilising these stages and create issues for other organisation students [28]. Due to informal organisation development, this issue is significant because it is now possible for the organisation's students to communicate with one another on these social networks. Clients, whether individuals or groups, are dynamic on Twitter to attract or target social site students to control them for illegal activities like bigotry [3], extremism, political debate, or specific targets. They attempt to reach individuals who can be effectively or with a tad of work to advance toward bigotry or fanaticism. A few gatherings are likewise dynamic, creating disdain or fanaticism among networks and involving animosity or frustration of individuals for their objectives [24]. These gatherings utilise text, pictures, audio, or video content to draw in and stand out of individuals. These groups can be recognised through their substance (posts or remarks on posts) that are being distributed.

In some cases, students' conversations will generally abhor or show animosity when they have genuinely felt more secure, so they attempt to bug others or may utilise oppressive language to target somebody [4]. They have explicit interests, so involving oppressive dialects in the organisations creates issues for informal community students. In this condition, students who are harassed report to the interpersonal organisation support group to deal with their inclinations. In interpersonal organisations, oppressive substances create issues in informal communities and impact individual lives [5]. Interpersonal organisations like Facebook, Twitter, and LinkedIn additionally give choices to conquer what is happening like "report as misuse" and so on

The excellent target of the examination is to propose another AI-based model to distinguish oppressive language by presenting new highlights. This study considers the following types of harmful language: animosity [6], badgering [7], cyberbullying, disdain [8], and foul language in the context of informal organisations. These sorts of misuse are recognised through the removed elements from the client content on the interpersonal organisation. By and large, help information through numerous assets that are physically commented on and apply pre-processing to clean the information. The subsequent stage includes extraction, which is a vital piece of the framework [9] to recognise the type of misuse and measure the degree of oppressiveness. The classifiers arranged the substance into six classifications: animosity, disdain, cyberbullying, badgering, obscene, and general. Harmful language is ordered in view of the element's extraction process, either oppressive or not [10,11] (while perhaps not then arranged into the overall classification). The presentation measure used to assess the exhibition of classifiers is exactness.

A study claims that the multi-class imbalanced dataset is a significant issue for all multi-class reviews [12, 13] because the conveyance of a few examples to the classes is not similar. For instance, if one class contains 10,000 examples and the below average has 3000, it will be a proportion of 3:10. This will influence the precision, and consequently, it is important to apply a component to defeat what is happening. Resampling is one of the two techniques for resampling, oversampling, and under sampling, which is utilised to adjust the dataset. Normal Techniques that are regularly utilised for content grouping [14] are feeling examination and point discovery that can be utilised to identify electronic oppressive substance utilising qualities of shippers, messages, and beneficiaries.

The significant commitment of the review incorporates proposing another methodology utilising the strategies of text mining to recognise oppressive substance in the informal organisation Twitter. New highlights were introduced to distinguish harmful language which is useful for further developing the classifier execution [6]. Identify relevant maltreatment in the client's content. Recognising the sort and level of misuse and further measuring oppressiveness.

The review depends on checking the current writing distributed all around the world. While leading a writing audit, our fundamental centre was those exploration papers, distributed during the most recent years [24]. This study focused on investigating the apparatuses and philosophies proposed by various specialists in the area of text mining methods to recognise harmful substances on Twitter or other informal communities. The concentrate likewise investigated the meaning of these strategies, their benefits or bad marks, and furthermore proposed any improvement that is conceivable.

The framework should refrain from online media exercises by recognising harmful words or relevant base maltreatment from client content to overcome security concerns. To extricate oppressive substances from client content is an intricate issue because of the heterogeneous idea of information on interpersonal organisations. The current methodologies channel the substance, but they are not sufficient; thus, text investigation is essential to recognise the harmful substance utilising message mining which is quite possibly the most fitting way to deal with separate helpful data. Different text-mining approaches have been utilised in existing investigations because of the idea of information; however, managed learning is normal and settles on the choice based on existing information.

Literature Review

Scientists have characterised text mining approaches into different types[7] including administered learning, solo learning, dictionary-based approach, rule-based approach, design-based approach, and so on. We evaluated these methods to select a suitable methodology that satisfies our prerequisites.

In regulated learning, models are trained using named datasets [7]. From that point onwards, the model dissects and distinguishes new examples according to the prepared information. Tellez et al. [15] proposed a multilingual technique for investigating feelings using a regulated learning approach that was applied to eight dialects: English, Spanish, Arabic, Italian, Russian, Swedish, and Portuguese. Another review [16] breaks down order calculations for recognising irreverence from disdain discourse in interpersonal organisations. Basak et al. [17] focused on the moderation and location of the awful impacts of public disgrace in interpersonal organisations. The uniqueness of the review is checked according to the viewpoint of the person in question. Burnap and Williams [18] addressed the issue of disdain discourse in Twitter content. Proposed a managed AI approach in which classifiers of text distinguish disdain discourse in information gathered by Twitter. Kawate and Patil [19] proposed a program-based expansion framework to identify and group foul language utilised in text discussions in interpersonal organisations. Reyes et al. [20] identified humour and incongruity dialects through programmed handling via online media such as Twitter. The creator considered different components to address different kinds of examples in the text, including unexpectedness, equivocalness, passionate substance, and extremity. A text-based highlights-based model was surveyed utilising two aspects: significance in regards to evaluating the features and representativeness to address assessment. Subramani et al. [13] identified homegrown maltreatment in informal organisations. Proposed a profound learning-based model to recognise aggressive behaviour at home in the client's posts or remarks on informal organisations. This model is useful for DVCS and rehearses in a continuous climate. For the most part, they presented a new dataset called the "highest quality level" that was physically commented on, fostered a profound learning grouping model, and considered its presentation in contrast to various engineering; approved the exhibition of the model against a standard of AI; upgraded the visual comparability of classifications and misclassification principle sources[25]; and installed space explicit development and its assessment from understanding age and order improvement viewpoint.

Research in [21] for disdainful substance and furthermore examine how client, past information can expand the presentation of a classifier for cannot stand content. Basically, it contributes to the engineering of profound learning for the text order of scornful substances that include elements of client conduct information. A language rationalist answer for identifying disdain discourse without pre-prepared word installation, assesses the model utilising the Twitter dataset and accomplishes top execution using the order approach. Oddity investigates the elements of students for a model that watches out for disdainful content. A study [22] proposed an answer to brain organisation to recognise web-based media knowing derisive substance. This approach is a blend of different LSTM (Long-transient memory (LSTM)-based classifiers to use the conduct qualities of the client that inclination toward sexism or prejudice to support execution.

A researcher [23]proposed a strategy to distinguish harmful language from client remarks on Twitter, recognise oppressive substances, and further a subtype of harmful language. Commotion has a decent sign for misuse recognition, accordingly acquainting highlights with catch different sorts of clamour. The primary objective of the study [4] is to research the application of programmed allowance of oppressive substances utilising centre text mining procedures. The Major commitment of the review is an application that utilises the centre text mining procedures to distinguish harmful language through different datasets gathered from various informal organisations. Another scientist [3] proposed a structure, CrowdPulse, that was utilised to remove the literary information from different social locales, and then applied calculations for feeling examination, semantic processing, and grouping of separating information. Execute the system in certifiable situations. Recognise dangerous regions in the Italian domain, as indicated by friendly site posts. Social locales dissect the components of bigotry: prejudice, homophobia, against women, incapacity, and anti-Semitism. Screen the states that are recuperating in friendly capital L'Aquila's city after the April 2009 quake, in which 297 individuals were killed. Where to recuperate extreme injury psychosocial and actual constructions. Another scholar [25] proposed a strategy for the programmed discovery of unigrams and examples of disdain discourse and utilised these with semantic and nostalgic highlights to arrange the tweets into sub-classifications. The significant contributions are as follows: identify disdain through an example-based approach, Collect of disdain or hostile substance from the example and use of a feeling-based approach for disdain discovery, Proposed unigrams sets and examples that can be utilised with an implicit word reference for disdain identification in future work, order tweets into three classes: hostile, disdain, and clean. Zheng, et al. [26] proposed an administered machine that depends on outrageous learning machine (ELM) for spammer location. First, the dataset is built, and the students are arranged into either spammer or non-spammer. Arbitrarily select 100 unique students and 50 spammers, and for every spammer, compare 500 late messages. To accomplish the best exhibition of 18 elements from client conduct and the substance message considered by the proposed arrangement models, more than 100% of the outcome was obtained with the F-measure. The proposed strategy is less sensitive to client-determined boundaries and is simple to implement. It tends to be duplicated in other interpersonal organisations with a slight correction.

The study [27] fostered a design Lexical Syntactic Feature for hostile client recognisable proof and hostile substance. A researcher [28] proposed a framework that identifies cyberbullying as pictures and text on informal organisations. To recognise oppressive pictures, BOW with the SVM classifier was used, and to distinguish the harmful language, BOW with the Naïve Bayes classifier was used. Then, the Boolean framework classified it as either hostile or not. A study [29] proposed a framework for feeling examination to identify opinions at the literary level and the opinion of a word or an expression in the message. A study [30] presented five kinds of logical maltreatment: racial, sexual, scholarly, appearance-related, and political. The dataset utilised in this study was completed in terms of quality and quantity. The dataset comprises organised and unstructured information, and framework execution shows that it is valuable for everyday activities. Furthermore, an API was developed for this framework. Information provided by JGZ, a general well-being association in the Netherlands, and GGD Amsterdam. JGZ also shares mastery and information as a paediatrician in the scrum bunch. Another study [22] proposed a cutting-edge mechanised AI (Auto-ML) framework to recognise multi-class harmful language in English and German.

A previous study [12]proposed a choice framework that distinguishes jumbled harmful language using unaided learning of oppressive words based on word2vec, skip-gram, and cosine likeness to identify recently developed oppressive words. As indicated by a previous study [8], the new age of emojis is emoticons. Emojis are realistic Unicode images used to communicate ideas and sentiments. Emoticons provide significant data on the text. This study developed a model that groups through feeling examination procedures and distinguishes the subjectivity of the sentence. A previous study [8] depicted that mockery is a complex type of incongruity that is mostlyor data (certain) inside the message in friendly destinations for various purposes: mockery as mind (being blissful), mockery as a whine (to show outrage or irritate), and mockery as avoidance (try not to offer an unmistakable response). A previous study [12] proposed a way to characterise the text into seven distinct classes, such as love, fun, satisfaction, bitterness, disdain, outrage, and non-partisan, in light of the feeling of the text. The proposed approach is versatile and is presently restricted to seven classes; however, the number of classes can be increased. Significant commitment of review; created a GUI apparatus to associate with the framework, presented new example-related highlights, and multi-characterisation with client decision using the device. This research reports the discoveries of the Italian disdain map utilising a vocabulary-based approach of semantic substance investigation. To accomplish the objectives, the Crowdplus structure [6] was utilised for social stream semantic investigation that extricates text information from interpersonal organisations and draws designs using understanding data.

A comprehensive review of the existing literature reveals various machine learning approaches for predicting student academic performance. A previous study [14] developed a student failure prediction model that categorised students into four performance classes based on CGPA, with the ID3 algorithm achieving 79.23% accuracy, although the model failed to address class imbalance limitations. Subsequent research introduced an ensemble model combining Naïve Bayes, SVM, and KNN classifiers [15], which incorporated standards-based grading assessment alongside traditional metrics and demonstrated enhanced performance with 85% accuracy. Further advancing the field, a sophisticated multilevel classification framework [17] addressed multiclass challenges through a two-level approach involving resampling techniques and outlier removal, ultimately achieving exceptional accuracy exceeding 90% using the J48 classifier. Another significant study [18] conducted a comparative analysis of multiple algorithms, including ANNs, decision trees, SVM, and naïve Bayes, determining SVM's superior performance of SVM for early failure identification, although the model lacked effective error reduction mechanisms. Collectively, these studies illustrate an evolving trajectory in educational data mining, from basic classification to increasingly sophisticated ensemble methods and preprocessing techniques designed to enhance predictive accuracy and address inherent dataset challenges.

Tools and Technology

This framework was created using Python for information analysis. The Python library NLTK assumes the principal part of this framework. Some other Python bundles likewise use to accomplish a few different focuses during improvement, like breaking down of hashtags done through the Python bundle, NLTK is a well-known library and broadly utilised for NLP. Every one of Each examination was run on a framework with the following specifications: Core I-7 eighth Generation, 32GB RAM, ITB, and 256 GB SSD.

Datasets

Our centre does not use information from web-based media and physically clarifies them to create a new dataset. We depend on the current marked dataset to distinguish animosity, provocation, disdain, cyberbullying, and disgusting dialects in student content. Table 1 presents the informational index. The dataset in CSV design in which 33,776 tweets each tweet has two credits: disdainful, and neither class. Another CSV file contains 24,784 tweets which are named disdain and oppressive.

Table 1: Dataset Demographics

Dataset(125241)	Abuse	Hate	Aggression	Cyberbullying	Harassment	Clean
Instances	19190	15214	22604	13590	15378	39265
% of total instances	15.33	12.14	18.05	10.85	12.28	31.35

Only the third dataset was obtained from the popular website Kaggle; the remaining datasets were part of existing studies. Combined, these datasets, by removing some duplicate entries, finally make a single and bigger dataset of 125k tweets. The combined dataset was further split into training and test datasets with ratio of70% and 30%, respectively.

Proposed Model

The designed model, which uses a text-mining approach, detects abusive language in the content provided by social networks. Initially, five different datasets were collected from various online sources to create a larger dataset of 125k tweets. Furthermore, the dataset is split into a training or test dataset to train a model for the identification of abusive language. Figure 1 shows the complete process model applied in this study. Feature extraction plays a vital role in models that detect abusive language.

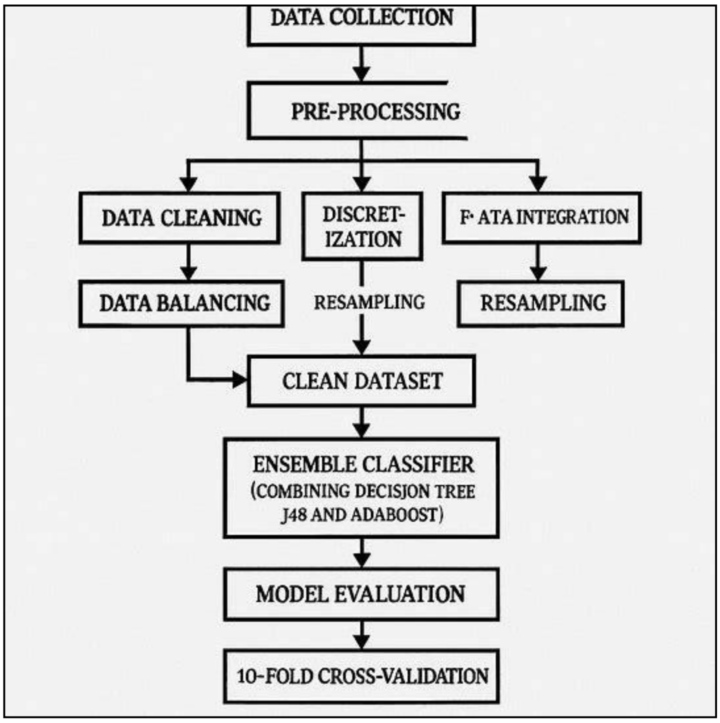

Figure 1: The proposed model

These features help identify abusive words and context-based abuse. Preprocessed data help extract different features, that is, sentiment, semantic, unigram, and pattern features, to identify the abuse and its subtype, that is, aggression, hate, harassment, cyberbullying, and vulgar language, in the content. Unigram and pattern features were extracted from the trained dataset based on a threshold value. This threshold value was used as a parameter that could be tuned to obtain more accurate results. Figure 1 shows the complete process model.

The three most common and popular classifiers used in this study, Naive Bayes, SVM, and LSTM, were used to classify the content into six classes: aggression, harassment, hate, cyberbullying, vulgar message, and general class. The general class represents the natural language if a tweet or comment is not abusive will list in this class. The optimised parameters will help to increase the performance of the classifiers by tuning these parameters. The final result shows the accuracy of each classifier separately for each class. This accuracy shows how to accurately detect abusive language from the content.

Preprocessing

Data preprocessing is an important step before extracting any type of information from the data to identify abusive language. The text should be cleaned up in the preprocessing step; therefore, preprocessing is the first step of the model before the process. After preprocessing, the data are stored as descriptive information in knowledge discovery.

Data Pre-processing

Data preprocessing is important for removing noise from the data, which is necessary to improve accuracy [10]. Data pre-processing is used to make the data useful for experimental purposes. Remove unnecessary things and correct the words of the sentence, mark punctuation, so that the data will provide maximum information. It is important to ensure that useful information is not removed from the content while preprocessing the data. For example, one common thing is converting the capitalised words into smaller words, which can remove the context of capitalised words, which are often used for shouting words. Figure 2 represents the data after applying the preprocessing steps.

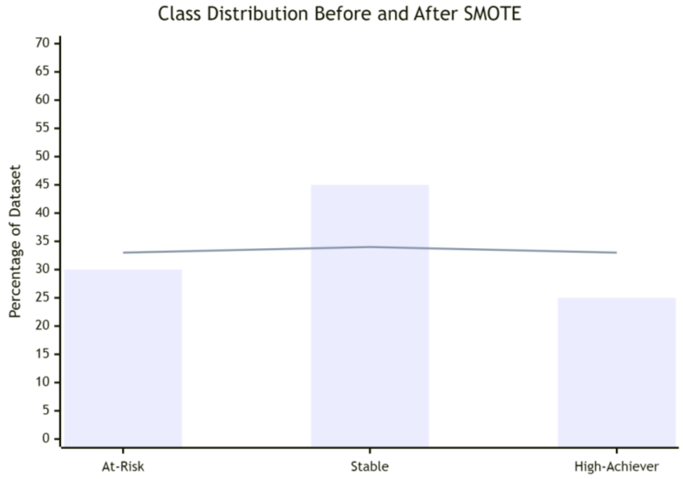

Figure 2: Graphical representation of data preprocessing.

Pre-processing was performed in almost all studies [4, 12], especially where the model was trained from the data. The preprocessing steps included the removal of unnecessary words, URL removal, stop words, string capitalisation to lower, word tokenisation, stemming, lemmatisation, POS tagging, and simplified POS tagging. These steps are common in most studies in which preprocessing is applied.

Feature Extraction

Feature extraction is a key part of this research for detecting abusive content. The feature extraction process helps identify the type and level of abuse. We identified various features to identify the abuse types, that is, aggression, hate, harassment, cyberbullying, and vulgar language in the content. The following four types of features were extracted from the tweets:

Sentiment Features

The first thing we need to know about a tweet is that the abusive value of tweets should be positive, negative, or neutral. So to find the polarity of abusive words by simple approach and for slang words could use abuilt-in dictionary, and to find out the value of the hashtag tag [8] developed a custom method for this.Emoji sentiment gets by using the library which was actually developed byNovak, et al. [34]. To determine whether a tweet has sentiment, we need to find a polarity score of tweet words, used emoji's, hashtags, and slang words. In this study, we set a range of negative and positive values, less than or equal to -0.5 as negative and greater than or equal to 0.5 as positive.

Semantic Features

In tweets, we cannot ignore the use of capitalised words, punctuation, exclamations, and other symbols. The semantic feature helps us to identify abusive expressions in which no abusive word is being used, but the sentence is abusive. We considered semantic features such as punctuation, capitalised words, laughter words, and the total number of words.

Unigram Features

The unigram feature helps to identify the explicit form of hate speech. All the unigrams collected from the trained dataset in a pragmatic way were used as independent features whose values were either 0 or 1. All the extracted unigrams that have POS tags stored in six lists along with the number of occurrences in the corresponding class. Only occurrences that satisfied the minimum occurrence value were considered.

Pattern Features

Pattern features help identify the implicit form of hate speech. Pattern features are extracted in the same way as unigram features. In the first step, the words of the tweet are divided into two categories based on whether they are sentimental or not: SW (sentiment word) and NSW (non-sentimental word). Any word in a tweet that has a POS tag refers to a noun, verb, adjective, and adverb saved into the SW category, and the remaining words saved into NSW. Different patterns were extracted from the tweets and saved into six lists if the pattern satisfied the condition of the minimum length. Only occurrences that satisfied the minimum occurrence threshold value were considered.

Features Extraction

In our study, features play an important role; therefore, it is necessary to describe the types of features used in existing studies for sentiment analysis. From a broad perspective, we can divide the feature types into four categories [7]: content-based, sentiment-based, user-based, and network-based features.

Content-based Features

Content-based features include document spellings and length, n-grams, term frequency-inverse document frequency (TF-IDF),bags-of-words (BOW), and pattern-related features. Abusive content is insulting and offensive in nature; therefore, the presence of profanity in the text indicates that the content is abusive. However, it depends on how profane words are used [8]; for example, "fucking bus is delay" is not abusive but "fucking idiot" is abusive. Some researchers have focused on domain-specific features to achieve high accuracy [12]. The first set topic sensitivity, such as culture, race, intelligence, and sexuality, then extract features of a specific domain.

Sentiment Features

Sentiment analysis is used to detect the sentiment in user content, that is, reviews of products or movies on social media, as well as in marketing and financial forecasting [25]. Sentiment features are also used to detect abusive language in user-generated content. Sentiment can be present in various forms; the simple form is directly abusive language use for someone, but it is difficult for sarcastic content or when students target someone in the form of humour and irony [20]. For example, "I like your big nose" is a sarcastic sentence where the “big nose" is a symbol of a negative mark. If the user replaces the word "nose" with "eyes" it will be a positive mark [12]. Similarly, students sometimes use humour and ironic language, such as jokes, to target others. Identifying these types of languages is challenging, and researchers have adopted various approaches to determine their features. This increases the accuracy of detecting abusive content in the future.

User-Based Features

User-based features are those features that are extracted through user saved information at social networks, that is, gender, age, race, and sexual orientation [7, 27]. In other words, any information that the user saves on the network can help identify abusive language as a user-based feature. Some studies have used user-based features with sentiment features, which have shown that user-based features play an important role in detecting abusive content. Some studies have highlighted user profiles involved in posting abusive content. Thus, blocked content is not useful in some cases when such profiles exist that create abusive posts. Therefore, in this case, the user is blocked instead of blocking the user content.

Network-based Features

Some network-based features [7] are becoming popular, such as uploads, number of friends, and likes.. The following features were highlighted during the literature review: the total time of user online through mobile, user activity at the networks, students subscription, membership duration, ego network, location of the user, number of profile following, number of followers, and the higher number of exchange messages between students. Social network game interactions are also important features. The researchers used these network features individually or with other feature types to improve the accuracy of sentiment analysis.

Classification

Various machine learning classifiers can be used to classify content into different classes. In the existing literature, I found that most researchers rely on Naïve Bayes, SVM, and LSTM. Some researchers have also considered other classifiers; however, most rating algorithms [5, 8]are these three. Some researchers have also used machine learning algorithms with slight modifications, but these are not reliable. In this study, we used Naïve Bayes, SVM, and LSTM to classify abusive content. We will run the classifiers for each class–aggression, harassment, hate, cyberbullying, vulgar message, and general class–separately and evaluate the results. The performance measure for our model was accuracy. The final results of each classifier show the achieved accuracy.

Results and Discussion

The performance was observed after conducting the experiments. The classification results were obtained with the following observations. Table 2 presents the data after executing all the processing steps. The left side indicates the algorithms chosen for the completion of the process model, and along the x-axis, the results are in the form of precision, accuracy, recall and F1 score.

Table 2: Comprehensive Model Performance Metrics on Test Set

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	Recall (At-Risk)	Precision (At-Risk)
Logistic Regression	0.72	0.71	0.70	0.70	0.65	0.68
k-Nearest Neighbors	0.75	0.74	0.73	0.73	0.71	0.72
Linear SVC	0.74	0.73	0.72	0.72	0.68	0.70
Decision Tree	0.78	0.77	0.77	0.77	0.75	0.76
Random Forest	0.85	0.85	0.84	0.84	0.82	0.83
XGBoost	0.86	0.86	0.85	0.85	0.83	0.84

Binary Classification

For the Binary classification model, all the abusive classes were combined to create a single offensive class, and the general class was considered the second class. The model was run on these classes and achieved an accuracy of 91.2%. The results are shown in Table 3.

Table 3: Binary Classification Results with Separate Features

	Precision	Recall	F-measure
	Sentiment Features		
Offensive	0.715	0.963	0.821
General	0.757	0.233	0.356
Overall	0.729	0.719	0.666
	Semantic Features		
Offensive	0.699	0.984	0.796
General	0.432	0.024	0.045
Overall	0.590	0.664	0.554
	Unigram Features		
Offensive	0.945	0.778	0.853
General	0.671	0.909	0.772

Overall	0.854	0.821	0.826
	Pattern Features		
Offensive	0.816	0.710	0.759
General	0.539	0.679	0.601
Overall	0.723	0.700	0.706
	Combined All Features		
Offensive	0.932	0.875	0.902
General	0.777	0.872	0.821
Overall	0.88	0.874	0.875

Ternary Classification via

For ternary classification, these classes are divided into three classes: hate, cyberbullying, and harassment to make one class of hate; aggression and vulgar language to make an abusive class; and the third class is general. Therefore, the classification model was run on these three classes and achieved an accuracy of 85.70%.

Multi-Class Classification

Finally, a model for multi-class classification is run, in which all classes contribute individually and achieve an accuracy of 80%. Existing studies that used text-mining approaches to detect abusive language in user content on the social network Twitter were reviewed. Abusive types, such as aggression, harassment, hate [25], cyberbullying [7], and vulgar language, were identified in the user content. Table 4 supports multiclass classification.

Table 4: Multiclass Classification Results with Separate Features

	Precision	**Recall**	**F-measure**
	Sentiment Features		
Offensive	0.529	0.363	0.655
Hatefull	0.474	0.390	0.428
General	0.466	0.696	0.671
Overall	0.490	0.483	0.648
	Semantic Features		
Offensive	0.392	0.655	0.490
Hatefull	0.326	0.219	0.262
General	0.434	0.284	0.343
Overall	0.384	0.386	0.365
	Unigram Features		
Offensive	0.887	0.701	0.783
Hatefull	0.846	0.639	0.728
General	0.641	0.931	0.759
Overall	0.791	0.757	0.757
	Pattern Features		
Offensive	0.886	0.734	0.803
Hatefull	0.545	0.281	0.370

However, it is difficult to identify all the abusive languages with 100% accuracy, but the target is to achieve maximum accuracy. Abusive content can be detected effectively and efficiently through feature extraction and classification using text mining techniques. Therefore, the ultimate goal was to propose an approach to detect abusive language and subtypes of abuse in a user's tweet or comment on the tweet. This study also provides a comparison of existing studies that use text-mining approaches to detect abusive content. The table 5 below are again in favor of how results are determined using a test set.

Table 5: Model Performance Comparison on Test Set

Model	**Accuracy**	**Macro Avg F1-Score**	**Weighted Avg F1-Score**	**Recall (At-Risk)**
Logistic Regression	0.72	0.70	0.72	0.65
k-Nearest Neighbors	0.75	0.73	0.75	0.71
Linear SVC	0.74	0.72	0.74	0.68
Decision Tree	0.78	0.77	0.78	0.75
Random Forest	0.85	0.84	0.85	0.82
XGBoost	0.86	0.85	0.86	0.83

Analysis of Results

Performance Overview: The ensemble methods, Random Forest and XGBoost, clearly outperformed all other models. They achieved the highest scores across all metrics, with XGBoost having only a slight advantage. This confirms the hypothesis that complex nonlinear tree-based ensembles are well suited for this task.

Critical Metric - Recall for At-Risk Class: This is the most important metric for an early warning system. The primary goal is to identify as many at-risk students as possible, even if it means some false alarms (lower precision). Both Random Forest and XGBoost achieved a recall of over 0.82, meaning that they correctly identified more than 82% of all students who ended up in the "At-Risk" category. This is a significantly better performance than that of the baseline Logistic Regression model (0.65). Table 6 summarises these results.

Table 6: Conclusion table

Actual / Predicted	At-Risk	Stable	High-Achiever
At-Risk	0.83	0.15	0.02
Stable	0.08	0.86	0.06
High-Achiever	0.03	0.11	0.86

Confusion Matrix Analysis (XG Boost)

The confusion matrix for the best model (XG Boost) is presented in Figure 3. Most misclassifications occurred between adjacent classes (e.g. a 'Stable' student predicted as 'At-Risk', or a 'High-Achiever' predicted as 'Stable').

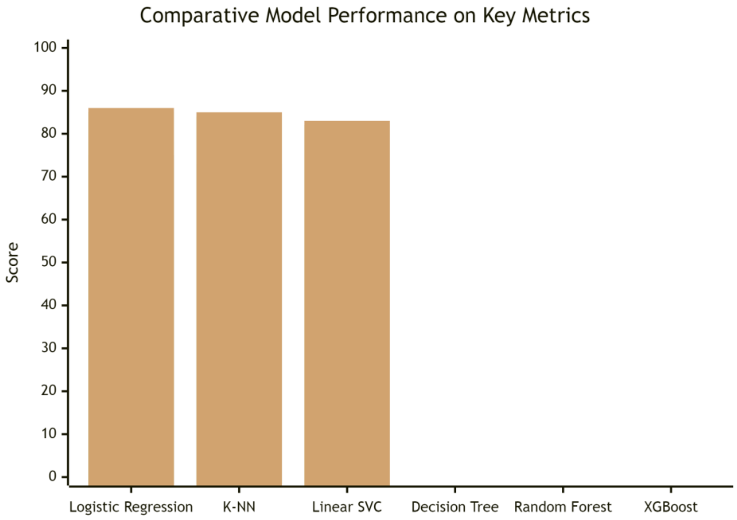

Figure 3: Graphical representation of algorithmic results

Very few '' at-risk students were misclassified as high achievers and vice versa, indicating that the model learns fundamental performance distinctions.

Conclusion

In this study, we reviewed existing studies that used text-mining approaches to detect abusive language in user content on the social network Twitter. Abusive types, such as aggression, harassment, hate, cyberbullying, and vulgar language, were identified in the user content. The ultimate goal was to propose an approach to detect abusive language and subtypes of abuse in a user's tweet or comment on the tweet. This study also provides a comparison of existing studies that use text-mining approaches to detect abusive content. The identification of contextual base abuse is ignored in most studies, which focus only on non-contextual abuse. In the future, this study can be extended to multiple languages. Various studies have proposed solutions with multiple languages, but this still needs to be explored, and a better model or proposed solution needs to be developed.

References

Zhang T, Liu X. Tracking the evolving impact of AI-driven learning platforms on EFL students' burnout, emotional challenges, and well-being: a longitudinal growth curve analysis. Innovation in Language Learning and Teaching. 2025 May 14:1-21.

Zuberi AH, Anees A, Anjum N, Warsi AH, Khan PR, Singh SK, Singh NK, Singh R, Abbas SH, Ranjan R. Machine Learning-Based Sentiment Analysis for Suicide Prevention and Mental Health Monitoring in Educational Institutions. Journal of Neonatal Surgery. 2025;14(5s).

Wahed SA, Wahed MA. AI-Driven Digital Well-Being: Developing a Machine Learning Model to Predict and Mitigate Internet Addiction. LatIA. 2025(3):73.

Ijiri A, Mori L, Galante PJ, Galante PM. Data-Driven Approach to Well-Being Tracking in Language Education for Whole-Person Learning. International Conference on Human-Computer Interaction, 2025, 25 May (pp. 257-270). Cham: Springer Nature Switzerland.

Hou L. Unboxing the intersections between self-esteem and academic mindfulness with test emotions, psychological wellness and academic achievement in artificial intelligence-supported learning environments: Evidence from English as a foreign language learners. British Educational Research Journal. 2025 Feb 22.

Ahmed K, Mathew A, Anand S. Ontology and AI Integration for Real-Time Detection of Cyberbullying Among University Students.

Ioannidou L, Argyriadi A, Argyriadis A. The Role of AI in Children's Mental Health and Well-Being Empowerment. In AI in Mental Health: Innovations, Challenges, and Collaborative Pathways 2025 (pp. 23-40). IGI Global Scientific Publishing.

Yan Y, Liu H. Ethical framework for AI education based on large language models. Education and Information Technologies. 2025 Jun;30(8):10891-909.

Xie Y, Fadahunsi KP, Broughan J, Donoghue JO, Gallagher J, Cullen W. Artificial intelligence for contextual well-being: Protocol for an exploratory sequential mixed methods study with medical students as a social microcosm. PLoS One. 2025 May 28;20(5):e0321426.

Diaz-Garcia JA, Carvalho JP. A survey of textual cyber abuse detection using cutting-edge language models and large language models. arXiv preprint arXiv:2501.05443. 2025 Jan 9.

Anozie-ibebunjo bl, ubi vo, chukwuokoro i. Mitigating tertiary students'unethical use of ai using language and critical literacy. Journal of humanities and social science. 2025 may 28.

Yu J, Tao Y. To be in AI-integrated language classes or not to be: Academic emotion regulation, self-esteem, L2 learning experiences and growth mindsets are in focus. British Educational Research Journal. 2025 Apr 28.

Kohnke L, Moorhouse BL. Enhancing the emotional aspects of language education through generative artificial intelligence (GenAI): A qualitative investigation. Computers in Human Behavior. 2025 Jun 1;167:108600.

Fadia Shah, Yasir Shah, Faiza Shah, Imran Shahid, Wang Zhou. AI Support to Enhance Game-Based Adaptive Learning. Research Journal for Social Affairs. 2024. 2(4), 203-212.

Guan KW, Giri S, Amara M, Jansen BJ, Liscio E, Esherick M, Owayyed MA, Ratkute A, Sedrakyan G, de Reuver M, Goncalves JF. Lived Experience in Dialogue: Co-designing Personalization in Large Language Models to Support Youth Mental Well-being. arXiv preprint arXiv:2511.05769. 2025 Nov 7.

Duong CD, Vu TN, Ngo TV, Do ND, Tran NM. Reduced student life satisfaction and academic performance: Unraveling the dark side of ChatGPT in the higher education context. International Journal of Human–Computer Interaction. 2025 Apr 18;41(8):4948-63.

Zhang T, Liu X. Tracking the evolving impact of AI-driven learning platforms on EFL students' burnout, emotional challenges, and well-being: a longitudinal growth curve analysis. Innovation in Language Learning and Teaching. 2025 May 14:1-21.

F Shah, Y Shah, F Shah.Financial Decision-Making via Sentiment Analysis of Stock Exchange Data TweetsResearch Journal for Social Affairs 3 (1), 129-139

Liu M. An IoT-enabled mental health monitoring system for English language students using generative adversarial network algorithm. Mobile Networks and Applications. 2024 Sep 4:1-22.

Lee J. Smart Solutions for Student Well-being: AI-driven Mental Health Support in a Polytechnic. InThe Rise of Intelligent Machines 2025 (pp. 211-242). Chapman and Hall/CRC.

Saad M. Can AI Really Protect Kids and Youth from Cyberbullying?. Available at SSRN 5253549. 2025 May 14.

Hou L. Unboxing the intersections between self-esteem and academic mindfulness with test emotions, psychological wellness and academic achievement in artificial intelligence-supported learning environments: Evidence from English as a foreign language learners. British Educational Research Journal. 2025 Feb 22.

Yan Y, Liu H. Ethical framework for AI education based on large language models. Education and Information Technologies. 2025 Jun;30(8):10891-909.

XiNuo M, Fan JY, Kang CC. Student Safety Monitoring System Through IOT Sensors For Bullying Prediction. In2024 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS) 2024 Nov 28 (pp. 258-264). IEEE.

F Shah, A Anwar, I ul haq, H AlSalman, S Hussain, S Al-Hadhrami.Artificial intelligence as a service for immoral content detection and eradication.Scientific Programming 2022 (1), 6825228

Yu J, Tao Y. To be in AI-integrated language classes or not to be: Academic emotion regulation, self-esteem, L2 learning experiences and growth mindsets are in focus. British Educational Research Journal. 2025 Apr 28.

Tang X, Upadyaya K, Hiroyuki T, Kasanen M, Salmela-Aro K. Assessing and tracking students' wellbeing through an automated scoring system: School Day Wellbeing Model. InAI in Learning: Designing the Future 2023 (pp. 55-71). Springer.

Orrù G, Galli A, Gattulli V, Gravina M, Marrone S, Micheletto M, Procaccino A, Nocerino W, Terrone G, Curtotti D, Impedovo D. Leveraging artificial intelligence to fight (cyber) bullying for human well-being: The bullybuster project. InCEUR WORKSHOP PROCEEDINGS 2023 (Vol. 3486, pp. 189-194). CEUR-WS Team, Redaktion Sun SITE.

Mendoza-Pinto R. Artificial Intelligence in the Fight Against Bullying: Integration of ChatGPT in an Emotional Support Chatbot. InCEUR Workshop Proceedings 2023 (Vol. 1613, p. 0073).

Chin H, Song H, Baek G, Shin M, Jung C, Cha M, Choi J, Cha C. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. Journal of Medical Internet Research. 2023 Oct 20;25:e51712.