International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I9P104

Volume 73 Issue 9, 32-47, September 2025
© 2025 Seventh Sense Research Group®

Original Article

Cyberbullying Detection, Categorization and Severity Classification in Networking Platforms for Teenagers and Young Adults

Sakshi Khanapure[1], Shilpa Deshpande[2], Brishti Basu[3], Arundhati Korlahalli[4], Anandita Rathod[5]

[1,2,3,4,5]*Computer Engineering Department, Cummins College of Engineering for Women, Pune, Maharashtra, India.*

[2]*Corresponding Author : shilpa.deshpande@cumminscollege.in*

Received: 25 February 2025 Revised: 30 August 2025 Accepted: 03 September 2025 Published: 30 September 2025

Abstract - *The emergence of various social networking platforms has made it effortless for individuals to connect globally and exchange their hobbies and interests. Especially among teenagers and young adults, having a social media presence and participating in multiplayer games has become a way of life. Cyberbullying is a prevalent crime that misuses the feature of staying anonymous online to bully and threaten people through digital platforms. Cyberbullying detection is hence the need of the hour. This paper offers a system to enable Cyberbullying Detection, Categorization, and Severity classification (Cb-DCS) of social media comments consisting of text along with emojis. The research focuses on detecting cyberbullying, categorizing it according to type, and classifying it according to severity with Machine Learning complemented by Deep Learning strategies. More specifically, the proposed work makes use of algorithms that include Bidirectional Long Short-Term Memory (BiLSTM), Multinomial Naïve Bayes (MNB), Bidirectional Encoder Representations from Transformers, and Support Vector Machine (SVM) for detecting cyberbullying. The categorization of cyberbullying comments is explored using techniques that include MNB, SVM, Random Forest (RF), and Convolutional Neural Networks. The severity identification of the comments is carried out using SVM, RF and MNB. The Cb-DCS system uses emoji embedding to extract the sentiment of the emojis. Experimental results show that the performances of MNB with Global Vectors (GloVe) for representing words, RF, and SVM are superior to the other corresponding techniques concerning accuracy and F1-score for the tasks of cyberbullying detection, categorization, and severity classification, respectively.*

Keywords - *Cyberbullying Categorization, Emoji, Machine Learning, Severity, Online Platforms.*

1. Introduction

The problem of cyberbullying is a complex one [1]. It can have adverse consequences for anybody who goes through this experience. But especially for teenagers who spend a large part of their time on the internet and can be easily influenced, getting bullied online can lead to serious mental health issues like depression [2], anxiety and self-harm. Moreover, cyberbullying creates a malevolent and toxic environment [3] that prevents youngsters from using the online network and social media to express themselves and connect with peers. It is crucial to realize that cyberbullying is not just a problem that affects individuals but also has a wider impact on society as a whole. By allowing cyberbullying to go unchecked, there is a risk of perpetuating a culture of cruelty and intolerance that may adversely affect mental health, social cohesion, and overall well-being [4]. Cyberbullying can create a negative online environment that promotes hate, intolerance, and negativity [5]. It can discourage users from participating in online communities and limit freedom of expression. Therefore, finding a solution for cyberbullying is essential to promote a positive online community and ensure the welfare of social media platform users.

1.1. Research Gap

The relevance of developing mechanisms to detect cyberbullying is evident. There have been attempts in the literature to provide solutions to the problem of cyberbullying. However, despite the existing work, there are still gaps in the research in this area, which are discussed below. Conventional rule-based approaches [1, 2] detect cyberbullying only on the basis of the occurrence of specific keywords. These approaches are not adequate for handling internet slang, which is changing every day. These approaches may also result in false positives, declaring non-harmful comments as cyberbullying. Addressing cyberbullying merely in the context of online textual content has also been explored earlier. However, consideration of only textual comments cannot be adequate, as they have been proliferating on social platforms to include visual or emotional aspects in emojis [2]. Many of the previous approaches [4] rely solely on determining if it represents a bullying or a non-bullying comment. But, such a binary identification of cyberbullying does not provide any insights into the specific type of bullying and its severity, which is needed for further analysis of its impact. Cyberbullying detection pertaining to a specific social networking platform [3] may not assure its effectiveness where various social media platforms are proliferating in practice.

1.2. Need for Solution and Contributions

Approaches need to consider the flexibility of processing to encompass a variety of internet slang and emojis along with the textual comments, to handle the problem of cyberbullying [2]. For the purpose of comprehensive detection, the approaches also need to take into account comments from different online platforms. Moreover, these approaches also need to reveal the specific type of bullying, such as hate speech or harassment and its gravity to enable necessary preventive measures in future.

To address the aforementioned issues, authors have developed a novel system to detect cyberbullying, categorize according to type and categorize according to severity. The proposed Cb-DCS system stands for Cyberbullying Detection, Categorization and Severity classification. Machine Learning (ML) and Deep Learning (DL) algorithms possess flexibility and adaptiveness by design and therefore, can be more helpful to detect bullying content effectively. Consequently, the proposed Cb-DCS system uses ML and DL algorithms to carry out the detection, categorization and severity classification tasks of cyberbullying.

The more specific contributions that this research presents are described below.

1) Create a carefully researched dataset: Develop a dataset by extensively researching and gathering data from various social online platforms, including comments and online chats.
2) Detection of cyberbullying comments that include text and emojis, along with analyzing text, the research considers the role of emojis. Incorporating the meaning of emojis enhances the accuracy and reliability of cyberbullying detection.
3) Optimizing Cyberbullying Detection Solutions by combining various ML and DL algorithms with different feature extraction and sampling methods.
4) The detected cyberbullying comments are categorized into five main types: religious, racist, sexist, homophobic, and hate.
5) Classifying the comment detected as cyberbullying based on its severity, such as low, medium or high.
6) Comparative analysis of various ML and DL algorithms based on their underlying accuracies and F1-score, using feature extraction techniques and sampling techniques for Cyberbullying detection, severity classification and categorization.

2. Literature Review

With digital transformation, social media platforms are increasingly being adopted by the young generation. These platforms enable enhanced online interactions. At the same time, they also pose the risks of cyberbullying [5]. Cyberbullying entails online harassment, carrying concerning ramifications. It manifests in various forms, predominantly as text-based content across numerous social platforms [6]. The scrutiny surrounding cyberbullying victimization, especially among young individuals, has grown more intense according to authors in [7]. Research [8] demonstrates that exposure to various forms of cyberbullying contributes to a rise in suicidal thoughts among adolescents.

The earlier approaches, which suggest solutions to the problem of cyberbullying, mainly fall into two types, viz., lexicon-based and rule-based [2, 4] methods. Lexicon-based methods [4] recognize cyberbullying with the help of lists of words and the occurrence of words in the lists. Rule-based methods [5] rely on predefined rules with which text contents are compared to detect bullying. Neither of these traditional methods can handle a language's dynamic online content and nuances. In recent years, researchers and practitioners have directed their efforts towards developing ML [9, 10] and DL [11, 12] based strategies to tackle the issue of cyberbullying identification in social networking platforms. These approaches aim to automatically identify instances of cyberbullying, enabling timely interventions and support for victims.

Authors in [9] make use of different ML classifiers and deduce the Logistic Regression (LR) model as the most effective one in identifying cyberbullying. An overview of predicting cyberbullying based on the contents created by users and the details about the users is given in [10]. Although an approach proposed in [11] considers more than one platform, it focuses only on the cyberbullying detection part, and further analyzing the severity of bullying is not addressed by it. An approach is proposed by the authors in [13] for detecting cyberbullying using a Neural Network (NN) and Support Vector Machine (SVM). Cyberbullying detection using a Convolutional Neural Network (CNN) at the character level is recommended by authors in [14]. Authors in [15] suggested a transformer-based model for identifying cyberbullying. The work in [16] explores the detection of bullying for YouTube comments. The approaches [12-16] focus only on the textual data gathered from social networks for the identification of cyberbullying. An approach is proposed in [17] for identifying the comments as not-bullying or bullying using SVM. Although the accuracy of cyberbullying detection gets improved by including user-related features, the proposed work focuses on the comments in the context of a specific application only.

Addressing the issue pertaining to cyberbullying detection has been attempted with respect to various social media platforms, which are popular among teenagers and young adults, such as Twitter [9, 18-23], Instagram [21, 24-26] and Reddit [23]. Approaches [18, 20, 23] are confined to textual data only for cyberbullying detection, and they do not take into account the use of emojis in comments for the analysis. A hybrid approach based on ML and DL is proposed in [21]. However, the scope of the work includes textual comments only for detecting cyberbullying. Approaches [9, 19, 21] focus only on identifying whether it is bullying or not, and they do not take into consideration further analyzing the type of bullying comments. The solutions to the problem of cyberbullying suggested by [22, 24-26] do not include the classification of comments based on the severity levels.

Work in [27] focuses on detecting cyberbullying in India. Hence, for the cyberbullying detection on text, the work considers the combination of English and Hindi languages, which are commonly used in communication on social media. An approach is proposed by the authors in [28] to identify the toxicity of comments shared on social networks. Authors have tackled the detection of cyberbullying in the context of online gaming platforms [29, 30]. Approaches that consider the combination of Natural Language Processing (NLP) along with ML are proposed in [31, 32] for detecting the type of cyberbullying. These approaches consider the comments from the Twitter platform. However, they do not take into account the analysis of the severity of the bullying comments.

Researchers have also highlighted the challenges associated with cyberbullying detection, such as the dynamic nature of online content, the evolution of bullying tactics, and the cultural and contextual nuances [33, 34] that influence the interpretation of messages.

In summary, the above review of existing work underscores the importance of addressing the issue of cyberbullying in today's world. The majority of the approaches focus on the detection part only, which means detecting whether it is cyberbullying or not. However, identifying the specific type of cyberbullying and assessing its severity level are also significant aspects that need to be addressed to handle the varied nature of online harassment. Often on social media, comments are accompanied by emojis. However, most of the reviewed approaches do not consider emojis, and they are bound to the text comments only for detecting cyberbullying. Most of the work only considered comments from a specific social media platform to detect cyberbullying. However, consideration of comments involving multiple social platforms is essential in comprehensively testing the effectiveness of any of the approaches.

The proposed Cb-DCS system intends to fill the gaps discussed above in the existing work. In addition to detecting cyberbullying, the Cb-DCS system categorizes it based on its type and classifies it according to severity using ML and DL techniques. The Cb-DCS system takes into consideration both the text and the emojis in social media comments to include the associated sentiments for effectively detecting cyberbullying. For the purpose of thorough experimentation in the context of the proposed system, a dataset containing the text comments along with the emojis is created by referring to multiple social media platforms.

2.1. Novelty of the Proposed Cb-DCS System

The qualitative comparison of the Cb-DCS system with the other approaches addressing cyberbullying is shown in Table 1. Table 1 indicates that, along with the detection of cyberbullying, categorization of bullying comments into the specific type, identifying the severity of the bullying, consideration of emojis along with the text in the comments, and a customized dataset of comments prepared from various social platforms are the distinguishing features of the Cb-DCS system, in comparison to the other approaches.

3. Methodology

3.1. Dataset Creation

In this section, the various steps followed to create the dataset that is used during the experimentation are elaborated.

3.1.1. Data Collection

Creating a dataset for cyberbullying detection, categorization, and severity classification is essential. It lays the foundation for developing effective models to address this important issue. In recent years, networking applications like Instagram, Reddit, and Twitter have become prevalent sources of bullying and harassment on the web. Comments and other text-based content are collected from the aforementioned platforms to prepare a dataset for the proposed Cb-DCS system.

Table 1. Qualitative comparison of approaches addressing cyberbullying

Reference	Use of Emojis in Comments	Supported Functionality – Cyberbullying Detection, Categorization, Severity Classification	Data from Single or Multiple Online Platforms
Muneer and Fati 2020 [9]	Not considered	- Detection in the form of a bullying or a non-bullying comment	Single platform - Twitter
Iwendi et al. 2023 [12]	Not considered	- Detection in the form of a normal comment or an insult in a comment	Single platform - Kaggle
Hani et al. 2019 [13]	Not considered	- Detection in the form of a bullying or a non-bullying comment	Single platform - Kaggle

Saifullah et al. 2024 [15]	Not considered	- Detection in the form of a bullying or a non-bullying comment	Multiple platforms like Twitter, Facebook
Alsubait and Alfage 2021 [16]	Not considered	- Detection in the form of a bullying or a non-bullying comment	Single platform - YouTube
Proposed system: Cb-DCS	Emojis are considered along with the text in comments	- Detection of a bullying or a non-bullying comment. - Categorization of bullying comments into the specific type - Classify the bullying comment based on its severity.	Creation of new data along with compiling and refining comments from multiple platforms – Twitter, Instagram, Reddit

From the aforementioned platforms. The data is carefully selected and refined to confirm its relevance and appropriateness for effectively training and assessing the system. In this context, a new dataset is created, comprising 5050 comments consisting of text and emojis.

3.1.2. Data Annotation

Currently, data is manually annotated, labelling each comment as either "bully" or "not bully," and categorizing them into types of bullying, such as racism, sexism, religious discrimination, homophobia, and hate speech.

Additionally, the severity level of each comment is assessed, classifying it as high, medium, or low. To ensure data quality, this work employs techniques for inter-annotator agreement [35] and implements data validation processes. These steps significantly enhance the effectiveness of the proposed Cb-DCS system. Labelling the comments manually as described above requires expertise in the domain. Hence, the number of comments is confined to 1350 for labelling and further experimentation.

3.1.3. Data Analysis

Figure 1 illustrates the five main categories and the corresponding subcategories that are used for comprehensive data collection. These categories include Racist, Religious, Sexist, Hate, and Homophobic comments. The categorization process is based on meticulous research, which involves exploring various papers [8, 36-38] related to psychological aspects of cyberbullying. The aim has been to create a comprehensive framework encompassing most areas where offensive and harmful comments occur.

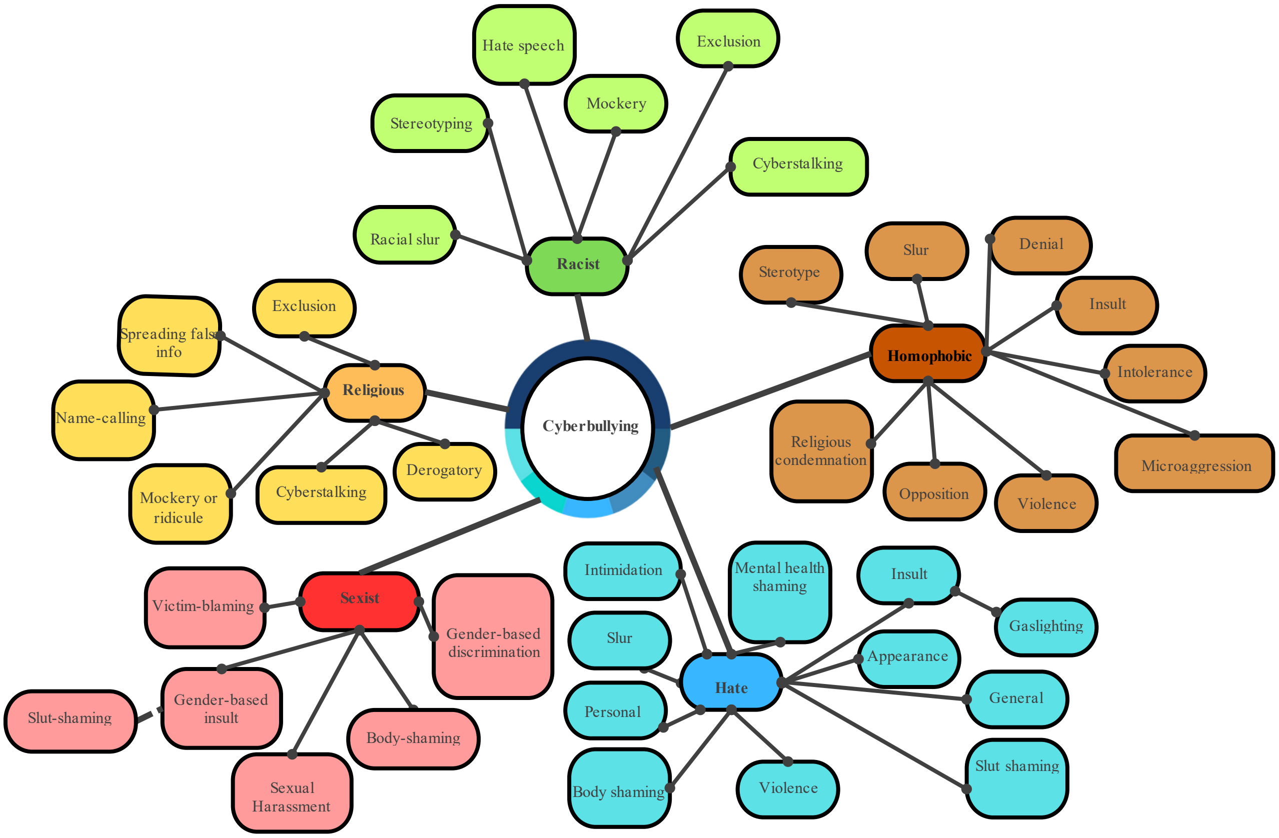

Fig. 1 Cyberbullying data categories and subcategories

Within the Racist category, the analysis further examines various subcategories, including racist slurs, stereotyping, hate speech, mockery, exclusion, and cyberstalking. By carefully scrutinizing these subcategories, it aims to capture the diverse nature of racist comments and understand the various ways in which racial animosity can manifest online. The Religious category again has various subtypes. A common one is "Name-Calling," which involves subjecting individuals to offensive labels due to their religious beliefs. The category "Exclusion" ostracizes individuals from specific religions. "Spreading false information" is used to disgrace or show a particular religion in a bad light, while "mockery or ridicule" can target religious practices and revered figures. Also included in this is "cyberstalking", which is a serious issue that can involve tracking and badgering individuals based on their faith, leading to anguish among victims. These practices collectively represent the various facets of religion-based prejudice in society.

The data used for the category of Sexist comments covers a spectrum of gender based insults. It also incorporates cases of sexual harassment where disrespectful and provocative comments are aimed at someone due to their gender. Furthermore, it includes victim-blaming, which shifts the blame to the victim for the abuse that they experience. Another subcategory is ‘gender based discrimination’- a practice perpetuating skewed gender stereotypes. The category of Hate is extensive and focuses on the core elements, covering instances of slut-shaming, a deeply insulting practice aimed at demeaning and stigmatizing individuals based on their sexuality or sexual behaviour. Another aspect addressed is mental health shaming, which highlights the importance of sensitivity and empathy towards individuals who are dealing with mental health challenges. Further, this category also includes situations of intimidation, which try to create fear and misery, along with remarks that aim to demean and belittle the person. Finally, the Homophobic comments category highlights the comments that convey denial, insults, offensive statements, and violent behaviours that target people because of their sexual identity. The main intention is to show the harmful impact of homophobia and thus try to promote an inclusive and accepting digital space. Figure 2 illustrates the word clouds for the Racist and Hate categories of the data.

(a)

(b)

Fig. 2 Word clouds for the data of sample categories (a) Racist, and (b) Hate.

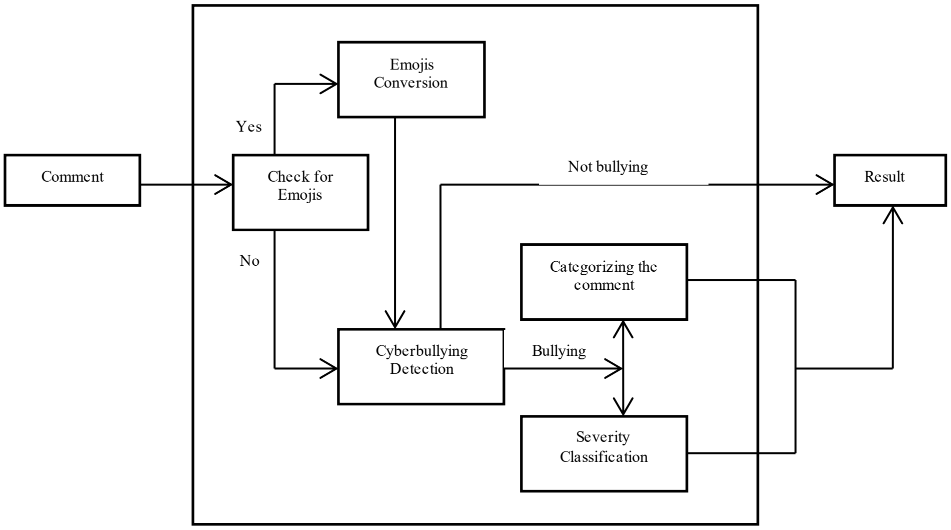

Fig. 3 Functional overview of the Cb-DCS system

3.2. Functional Overview of Cb-DCS System

Figure 3 portrays the functional overview of the proposed Cb-DCS system. As depicted in the figure, the Cb-DCS system takes a comment as input and determines whether the message contains cyberbullying content. Additionally, it provides information about the category and severity level of the comment. Accordingly, the Cb-DCS system determines whether it is bullying content. If it is bullying content, then the result also gives the category and severity level of the bullying comment. The Cb-DCS system comprises four essential modules, which are described below.

3.2.1. Emoji Conversion

In this module, the system examines comments for the presence of emojis. If emojis are detected, they are processed by replacing them with their corresponding meanings. The emoji2vec model [39] is trained on a dataset of one million tweets to identify the word most commonly associated with each emoji. For example, the "😂" emoji is replaced with the word "haha". A literal translation is used for newer emojis that are not present in the training corpus. For instance, "🤡" becomes "clown_face".

3.2.2. Cyberbullying Detection

Within the dataset, some comments contain cyberbullying content, while others do not. Through annotations, the model used in the Cb-DCS system discerns which comments are mean or hurtful (cyberbullying) and which ones are friendly. During training, the model strives to find patterns and indicators that distinguish cyberbullying from non-cyberbullying content. It learns to recognize specific words, phrases, or combinations of words often associated with cyberbullying behaviour. After training, the model can take a new message as input and predict whether it constitutes cyberbullying. The cyberbullying detection is done using the technique with the best performance. Four different techniques, viz. Bidirectional Long Short-Term Memory (BiLSTM), Multinomial Naïve Bayes (MNB), Bidirectional Encoder Representations from Transformers (BERT), and SVM are considered in this work. The details of cyberbullying detection are described in the ensuing subsection.

3.2.3. Categorization of Comment

At this stage, the model has learned patterns associated with bullying comments. However, these patterns can be associated with various categories of cyberbullying. Bullying comments in the dataset are further annotated based on their respective categories. The model then learns different words associated with each category. For example, the "sexist" category may include words such as "whore" and "slut." The categorization of comments is done using the technique with , showing the best performance. Four different techniques, viz. MNB, SVM, Random Forest (RF), and CNN are considered for the categorization of comments. The details of the categorization of comments are described in the forthcoming subsection.

3.2.4. Severity Classification

To classify the severity of cyberbullying, the system evaluates the impact of the comment on the victim. For instance, a comment containing a death threat is classified as high severity. A comment like "The public should unite if someone is harassed by these individuals. 🔫🗡️" contains a death threat and is categorized as high severity. The severity classification is carried out using the technique with the best performance. Three different techniques, viz. SVM, RF, and MNB are considered for this work. The details of severity classification are described in the approaching subsection.

3.3. Data Cleaning and Pre-processing

Data pre-processing contributes vitally towards preparing the dataset for machine learning models, particularly when dealing with cyberbullying comments. Duplicate comments are identified and removed to ensure data uniqueness, while missing values within the comments are dropped to maintain data completeness. All capitalized text within the comments is converted to lowercase to standardize the data, promoting consistency in subsequent analyses. Privacy concerns are addressed by removing usernames and mentions of other users from the comments. Abbreviated expressions are expanded to their full forms to facilitate accurate understanding and interpretation. Errors with spellings in comments are corrected and will result in accurate and reliable data.

The comments are then converted into a list of words, making further analysis and processing possible. Emojis in the comments are interpreted to their associated sentiments, thus providing further meaning to the text. Afterwards, the list of words is concatenated back into single strings, which results in structuring the data for efficient analysis. The special characters and punctuation within the comments are eliminated, bringing the data into line with a uniform and standardized format. The full forms of commonly utilized acronyms are included to ensure accuracy and clarity in the underlying meaning. URLs or links in the comments are removed in order to eliminate any irrelevant information that may cause bias in the analysis.

Accent marks or diacritics are eliminated from the comments, which further improves data consistency and standardization. Extra spaces between words are normalized, making the data easier to handle and aiding in future processing. Lemmatization is utilized to convert each word in the comments to its root form, enhancing uniformity and consistency within the dataset. The above-mentioned data preprocessing ensures cleanliness and makes the data suitable for further analysis. Thus, it improves the reliability in addressing cyberbullying.

3.4. Emoji Pre-Processing

To determine the sentiment associated with emojis, an approach is adopted that leverages the Word2Vec model [40, 41] with embeddings. The model mentioned represents the words into vectors, which take into consideration the semantic relationship between the words. Pre-processing involves processing a dataset that includes both text and emojis. Individual lines in the dataset are divided into separate words. These words collectively form a list of sentences, which will serve as the source to train the Word2Vec algorithm.

The Word2Vec model has been trained by means of the prepared list of sentences. During training, the model adjusts its parameters to create meaningful vector representations for words and emojis based on their context within the sentences. Specific hyperparameters, such as the hidden layer size, window size, minimum count, and negative samples, are set to guide the training process and the quality of the resulting embeddings. By using Word2Vec embeddings, the approach aims to map emojis, like "😂," to their corresponding sentiments, such as "haha," considering the semantic correlations acquired from the training data.

A prediction class is defined to determine the sentiment associated with emojis. This class loads the trained Word2Vec model from the binary file. The getPrediction() method, within the Prediction class, takes an emoji as input and returns the most similar words from the model's vocabulary. By setting the emoji_only parameter to True, the method filters out non-emoji words, ensuring that only the sentiment associated with the emoji is extracted.

To facilitate further analysis and predictions, the Prediction class also provides the get_vector_embedding() method. This method retrieves the word embedding for a given word within the model's vocabulary, allowing for sentiment analysis of emojis within the dataset. By utilizing Word2Vec embeddings, the methods mentioned above enable the conversion of emojis into their associated sentiments. The resulting model and methods provide a means to analyze and interpret the sentiments conveyed by emojis, enhancing the preprocessing pipeline for effectively addressing emoji-related sentiments within the dataset.

3.5. Feature Extraction (FE)

In the domain of cyberbullying, feature extraction plays a vital part in identifying and detecting instances of online harassment, abuse, or harmful behaviour. The goal is to convert comments from text or online interactions into numerical data to extract relevant information, enabling the building of effective cyberbullying detection models. This process is accomplished through feature extraction techniques, as shown in Figure 4. These techniques are discussed in the following paragraphs. Word-to-integer mapping, also known as word encoding or word indexing, is an activity of representing words in natural language text as corresponding integer values. This conversion is essential in various NLP functions, since ML techniques typically operate on quantitative figures.

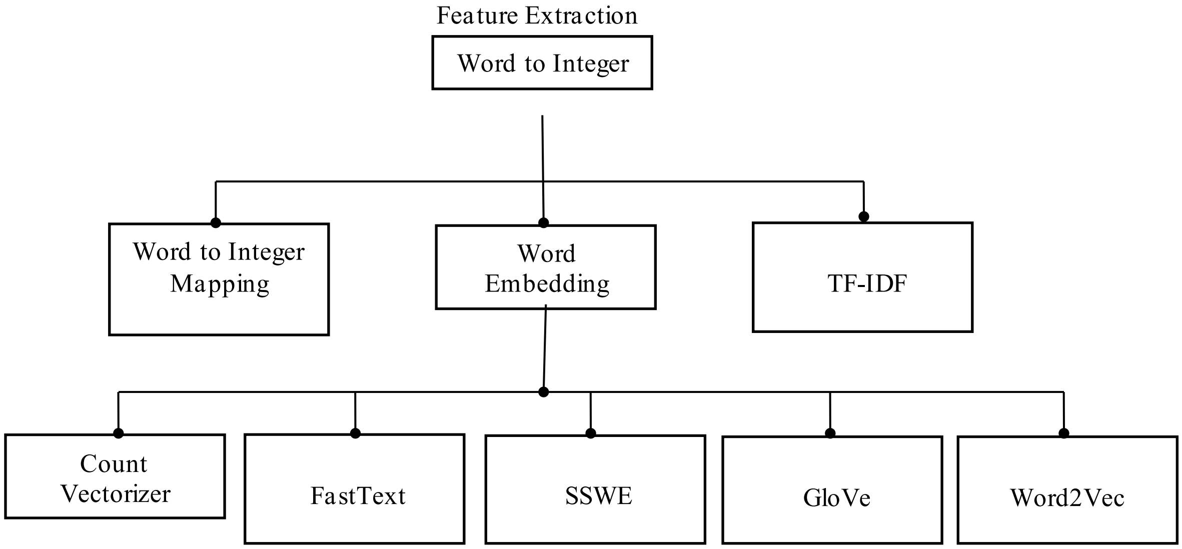

Fig. 4 Feature extraction

Term Frequency-Inverse Document Frequency (TF-IDF) [9] represents an arithmetical value employed in NLP to assess a word's significance within the text in relation to a corpus. Word embeddings, which include Word2Vec, CountVectorizer (CV), Global Vectors (GloVe) [41], and FastText [40, 41], are essential in understanding the context of a particular word's use. Effective text data analysis is rendered possible by these methods, which offer dense vector representations that capture syntactic and semantic properties of words. GloVe is unique because of its capacity to capture sub-linear relationships between word vectors. This adds more practical meaning to the representations. Rather than using individual words, relationships between a pair of words are considered, giving information about the semantic perspective in which the words are used. Hence, GloVe is a great choice for deciphering complicated language patterns.

Alternatively, the CountVectorizer is also an essential technique in the analysis. In contrast to the word embeddings, which capture semantic knowledge, CountVectorizer represents words by how frequently they occur in the text. It generates a sparse vector representation in which each dimension corresponds to a distinct word and its occurrence count. This can be a useful approach for detecting specific patterns and keywords linked to acts of cyberbullying. Post validating multiple approaches, one can confirm that GloVe is more appropriate for the cyberbullying detection models compared to Word2Vec, FastText, Sentiment Specific Word Embedding (SSWE), CountVectorizer and TF-IDF.

While Word2Vec and FastText focus on capturing contextual relationships, they may not provide the necessary semantic insights for identifying harmful language patterns. SSWE's sentiment-focused embedding may not fully address the multifaceted nature of cyberbullying detection. While useful for term importance, TF-IDF lacks the comprehensive semantic context offered by GloVe and the CountVectorizer. Thus, taking into account the robust representations, GloVe and CountVectorizer can be chosen. It, in turn, enables effective identification of cyberbullying.

3.6. Sampling Methods (SM)

The performance of algorithms is impacted when there are imbalanced classes. Sampling methods play a vital role in balancing these classes. There are different sampling techniques as follows:

1) Within the realm of cyberbullying, oversampling involves amplifying the instances found in the less frequent class to level the playing field of class distribution. This technique can be executed by replicating existing instances or artificially creating new samples belonging to the minority class.
2) In no-sampling, one does not make any changes to the original dataset. The existing distribution of the instances stays as it is, and no modifications are made to neutralize the imbalanced classes.
3) In undersampling, the issue of class imbalance is tackled by decreasing the instances of the greater size class to tally with the count of the smaller size class. To achieve this, instances from the dominant category are selected randomly and eliminated.
4) The Synthetic Minority Oversampling Technique (SMOTE) represents a method that creates interpolated instances between existing minority class data points to create synthetic samples. By introducing artificially generated instances, this method seeks to deal with class inequity by increasing the representation of the minority class.

3.7. Cyberbullying Detection Algorithms

Cyberbullying detection is the ability to distinguish a comment as either cyberbullying or non-cyberbullying. The task at hand is a binary, black and white classification. Various algorithms for the same have been explored, as shown below.

3.7.1. BiLSTM

BiLSTM indicates a variation of Recurrent Neural Networks (RNN) that facilitates bidirectional information flow within the network. This means that it can process data in two ways, back and forth, enabling it to catch related information from previous and next elements in a sequence. RNNs are a popular choice for NLP tasks because they are capable of processing sequential data. This helps in taking into consideration the context of a word or sentence, as that can depend on both the preceding and succeeding words. BiLSTM does not necessarily require feature extraction, but the Word2Vec word embedding technique has been used, as it is known to improve performance. Another advantage of using Word2Vec is that it helps reduce dimensionality, which can increase accuracy. A pre-processed dataset was employed to train this model. Random oversampling has been done.

3.7.2. MNB

The MNB algorithm uses Bayes' theorem as a base for classification. During training, the algorithm estimates the earlier probability of each class and the conditional probability of each feature specified for every class. MNB has one of the key benefits of being computationally efficient and works well with relatively small training datasets.

Various techniques are explored for feature extraction. First is TF-IDF, which is used to assign weights to the features. It works well with the MNB as it represents textual data in a manner that the MNB can utilize with ease.

The second technique is a word embedding technique, FastText. FastText describes words in the form of a collection of character n-grams, which allows it to capture morphological and syntactic information in addition to semantic information. But it has to be scaled in order to be given as input to the MNB model.

The third technique is also a word embedding technique known as GloVe. It utilizes the co-occurrence matrix involving words within a collection. The resultant word embeddings capture the statistical relationships among the words. For better representation of data, oversampling is explored with TF-IDF. However, since class imbalance for the detection task is not very significant, sampling does not affect the effectiveness of the cyberbullying detection task in particular.

3.7.3. BERT

In cyberbullying detection, BERT employs a multi-step process that involves preprocessing, embedding, and fine-tuning. The text data is preprocessed and split into individual words called tokens. BERT then generates embeddings or numerical representations of these tokens to capture their contextual meaning. These numerical representations are then used to generate fixed-length vectors that are provided to the fine-tuned BERT model. Training of the model has been done on a labelled dataset of cyberbullying texts using backpropagation and stochastic gradient descent to reduce the variation between predicted and true labels. In order to balance the classes, the technique of oversampling is implemented. During training, the weights assigned to the pre-trained BERT model were updated, along with the weights of the classification layer, to suit the particular task of cyberbullying detection.

3.7.4. SVM

SVM is a good choice for the detection of cyberbullying due to its accuracy in classification jobs. This algorithm operates by mapping data onto a high-dimensional space. It also maximizes the margin between the closest points of two different classes. Datasets containing outliers, sarcasm and slang are common in the context of this work, but SVM ignores instances that do not fit the general trend of the data, leading to better accuracy. The dataset has been balanced using oversampling and converted to vectors using TF-IDF. After this, SVM is used for classification. This is a good combination as TF-IDF vectors are usually high-dimensional and sparse, and SVM performs well in these conditions.

3.8. Cyberbullying Categorization Algorithms

Categorization in the context of cyberbullying involves grouping different instances of online harmful behaviours into distinct categories of Religious, Hate, Homophobic, Racist and Sexist based on their attributes and intent. This process aids in recognizing different types of cyberbullying, like harassment, threats, impersonation, and hate speech, allowing for focused analysis, prevention strategies, and effective interventions. By identifying and labelling specific types of cyberbullying, individuals, platforms, and authorities can better address the issue and tailor appropriate responses to combat the diverse nature of online harm. The different types of algorithms explored for categorization are given as follows.

3.8.1. MNB

Due to the success of MNB in the task of detection, the same is applied for categorization. Considering the effectiveness of GloVe with MNB in detection, word embeddings are initially generated using GloVe. Undersampling is initially used to balance the dataset. While it equalizes data in all five classes, the overall reduction in samples affects accuracy. Subsequently, the feature extraction technique of TF-IDF has been explored with MNB using oversampling to address the issue of class imbalance and improve classifier performance.

3.8.2. SVM

SVM is a classification algorithm that operates by determining the hyperplane that maximally segregates the classes. SVM works well when the number of features is large relative to the number of samples. This is often the case with textual data; hence, SVM is suited for NLP tasks like categorizing cyberbullying. In the considered data as well, the count of features is greater than the count of samples.

Oversampling is implemented to balance the classes. SVM works better with oversampling rather than undersampling. The reason is that, while dealing with imbalanced data, SVM can be sensitive to the minority class. This means that undersampling the majority class can cause important information to be lost. This can cause a reduction in accuracy as the SVM model may struggle to find a good boundary between the classes. For feature extraction, TF-IDF is used.

3.8.3. RF

In an RF algorithm, many decision trees have been constructed based on various subsets of input features, and their outputs are combined to make a final prediction. The decision trees are built with recursive splitting of the input data into smaller subsets depending on the input features, until a stopping criterion is met. In conjunction with RF, SMOTE is a technique used to balance data and improve model performance. When applied to RF, it can aid in reducing the possibility of overfitting and improving the stability of the algorithm to learn. The pipeline module provides a Pipeline class that allows chaining together multiple processing steps in a machine learning workflow. CountVectorizer is one such step that transforms text into a numerical representation, counts the occurrences of every word (or n-gram) within the text, giving a document-term matrix where every row signifies a document and every column signifies a word (or n-gram) within the vocabulary. The RF model with SMOTE shows the best accuracy and is a good choice due to its ensemble learning technique, which consolidates many decision trees for making predictions, improving model performance and stability.

3.8.4. CNN

CNN signifies a type of neural network broadly employed in tasks involving images and signals, but it can also be used in NLP tasks like text classification. In this context, a CNN can be trained on a large dataset of text, with examples from various categories (e.g., "spam" vs. "not spam," "cyberbullying" vs. "not cyberbullying"). The network learns to automatically discover relevant features from the supplied text to classify new messages into these categories.

The typical architecture of a CNN for text classification includes an initial layer that maps words in the input text to high-dimensional vectors, followed by layers that apply convolutional as well as pooling operations for extracting features from the text at different levels. To create the final classification, these features are subsequently run using one or many completely connected layers. After training, the model can process incoming messages through the network and classify them. The final layer indicates the probability that each message falls into one of the various categories. Text sequences are padded to a predetermined length in this procedure to guarantee consistent input size, which is frequently required when working with text data. Furthermore, numerical values derived from category variables are turned into one-hot encoded vectors for further processing. In order to properly preprocess the text data, the model uses Keras' Tokenizer [42] for transforming the raw textual data into quantitative sequences, which is an essential step for feeding the data into the neural network. By assigning each word a unique index, this transformation produces a numerical representation that the model can process and understand efficiently.

3.9. Severity Classification Algorithms

In the current context, severity refers to the extent of harm that will be caused, ranging from low to medium and high. In the case of cyberbullying incidents, it evaluates the consequences and seriousness based on various parameters like vulnerability of the victim, explicit content, contextual elements and persistence. With this understanding, various appropriate measures can be taken, and support can be provided to those impacted by diverse levels of online aggression and abuse. Prioritization of cases can be done by educators, individuals and platforms as the level of harm can be determined based on the severity. The various kinds of algorithms that have been examined for severity classification are described below.

3.9.1. SVM

SVM with TF-IDF often provides a strong baseline; however, its effectiveness can be limited by class imbalance. To address this point, oversampling has been performed to balance the classes, which in this case are the medium severity classes.

The combination of SVM with the SMOTE sampling technique has proven to be effective. It is a suitable algorithm because it can handle non-linearly separable datasets, which is important for accurately classifying instances of cyberbullying severity. Additionally, SVM is a robust algorithm capable of handling noisy and outlier data, which is important for cyberbullying severity classification, as the text data may contain misspellings, slang, or sarcasm.

The SMOTE sampling method thus balances the dataset by generating synthetic examples of the minority class. CountVectorizer, as an FE technique, converts text data into numerical vectors based on word counts, enabling SVM to operate on text inputs effectively. Overall, SVM with SMOTE and CountVectorizer is a powerful combination for the classification of cyberbullying severity.

3.9.2. RF

The RF approach combines the effects of the number of decision trees to produce a single outcome. Since it can handle both classification and regression problems, its versatility and ease of use have driven its adoption. CountVectorizer is used to extract the features, and the data has been oversampled before building the decision trees.

3.9.3. MNB with TF-IDF

MNB operates by assuming feature conditional independence specified in the class label, which often does not hold in datasets like cyberbullying severity analysis. TF-IDF adjusts word frequencies based on their importance across documents, thus prioritizing informative terms for classification tasks; hence, it has been chosen. Oversampling can be used to mitigate the common issue of imbalanced data, as in the case of severity classification, the severe cases may be less in number.

3.9.4. MNB with GloVe

In the case of detection and categorization, this approach has been proven as a good choice, and thus, this approach has been chosen for severity analysis as well. However, upon using this technique, it is realized that accuracy is not as expected. Even though GloVe embeddings have proven to capture semantic relationships within words, they may not represent the nuanced aspects of severity in cyberbullying comments.

4. Results and Discussion

A prototype is developed for the Cb-DCS system using Python for experimentation and evaluation. This prototype is executed using Google Colaboratory and a Jupyter notebook environment. The prototype implementation makes use of the dataset, which has been created as described in Section 3.1, for training and testing with various social media comments. This dataset has been split into 80 percent and 20 percent in that order, for the purpose of training and testing.

4.1. Evaluation Metrics

To validate the effectiveness of various algorithms used for cyberbullying detection, categorization and severity classification, accuracy and F1-score are employed as the metrics. These metrics give knowledge about the performances of algorithms and are used to select the best performing algorithms in each of the cyberbullying detection, categorization and severity classification parts of the proposed Cb-DCS system.

Accuracy is the measure of correctly predicting the classes of comments by comparing them to the ground truth labels. Cyberbullying detection indicates the degree of correctly predicting whether the comments are bullying or non-bullying comments. Similarly, in the context of categorization, accuracy implies the measure of correctly predicting the type of bullying comments. In severity classification, accuracy signifies the degree of correctly predicting the severity level, such as high, low, or medium, of the comment.

Precision indicates the proportion of the number of rightly predicted bullying comments with respect to the total count of predicted bullying comments. In cyberbullying categorization, precision signifies the proportion of the count of correctly predicted comments in the specific category, such as homophobic, racist or hate, with respect to the total number of predicted comments in that specific category.

Recall signifies the proportion of the count of rightly predicted bullying comments with respect to the total count of actual bullying comments. Similarly, in cyberbullying categorization, recall implies the proportion of the count of rightly predicted comments for the specific category with respect to the total count of actual bullying comments belonging to that specific category. In the context of severity classification, recall signifies the proportion of the count of rightly predicted comments for the specific severity level with respect to the total count of actual bullying comments corresponding to that specific severity level.

F1-score depicts a value obtained by proportionately blending Precision and Recall. Bullying comments getting rightly predicted as the bullying ones and the non-bullying comments getting rightly predicted as the non-bullying ones are equally significant. Similarly, in the categorization and severity classification of bullying comments, rightly predicting the classes of comments is significant. Hence, this work considers the F1-score as a metric instead of considering merely Precision or Recall.

4.2. Results and Analysis

The performances of various algorithms used in cyberbullying detection, categorization, and severity classification are compared using accuracy and F1-score.

4.2.1. Assessment of Accuracy and F1-Score in Cyberbullying Detection

For cyberbullying detection, the algorithms described in Section 3.7 are explored with various sampling and feature extraction techniques to identify the algorithm with the optimized scores concerning accuracy and F1-score. The results from Table 2 depict that the MNB algorithm, when applied with no sampling and GloVe as a feature extraction technique, gives the maximum accuracy among the other algorithms. This means that when the MNB algorithm with no sampling and the GloVe technique is employed in detecting cyberbullying on the set of comments, it gives the lowest error in classifying the comments, and the rate of rightly predicting the bullying or the non-bullying comments will be the highest. Figures 5 and 6 graphically demonstrate the experimental results for accuracy and F1-score obtained in cyberbullying detection, respectively. The results from Table 2, Figures 5 and 6 show that the MNB algorithm with no-sampling and GloVe as a feature extraction method possesses the topmost values of accuracy as well as F1-score in comparison with the other algorithms. The MNB algorithm is more effective for a higher number of features. Also, the GloVe technique offers the benefit of capturing the semantic relationship among the words instead of individual words, which is optimally utilized by MNB. Hence, the combination of MNB and GloVe gives better results in cyberbullying detection than other algorithms.

Table 2. Comparative results of cyberbullying detection

Algorithm	SM	FE	Accuracy (%)	F1-score (%)
BiLSTM	Over-sampling	Word2-Vec	73	72
MNB	Over-sampling	TF-IDF	90	90
MNB	No-sampling	FastText	65	64
MNB	No-sampling	GloVe	92	92
BERT	Over-sampling	BERT	88	88
SVM	Over-sampling	TF-IDF	88	87

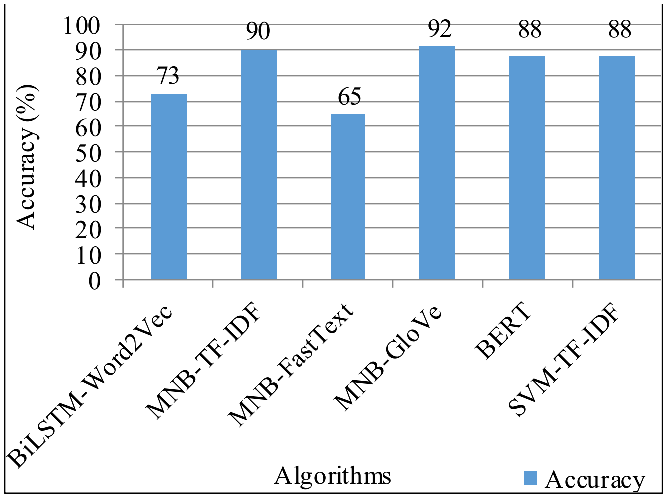

Fig. 5 Accuracy in cyberbullying detection

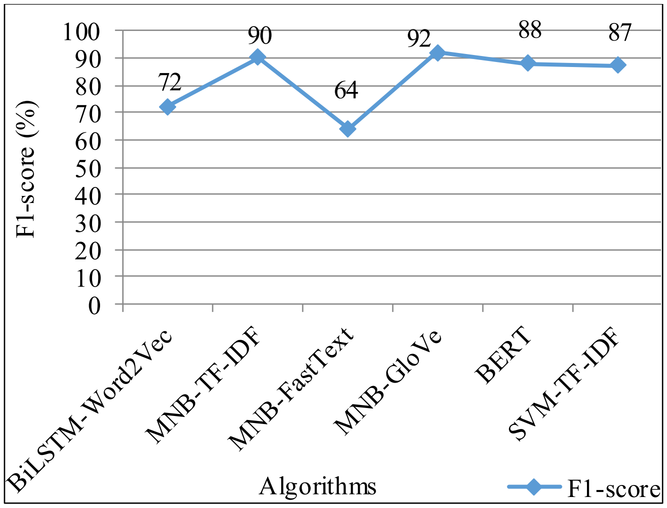

Fig. 6 F1-score in cyberbullying detection

Thus, because the MNB algorithm with no sampling and the GloVe technique demonstrate the maximum accuracy and the F1-score, it is the most suitable technique for cyberbullying detection in the Cb-DCS system.

4.2.2. Assessment of Accuracy and F1-Score for Bullying Comments Categorization

In categorization of cyberbullying comments, the algorithms discussed in Section 3.8 are applied with different sampling and feature extraction techniques to determine the algorithm with the optimized scores concerning accuracy and F1-score. The values of the metric recorded inside Table 3 illustrate that the RF algorithm with SMOTE sampling and CountVectorizer as a feature extraction technique has the maximum accuracy among the other algorithms. This implies that when the RF algorithm with SMOTE sampling and CountVectorizer technique is applied for cyberbullying categorization on the set of bullying comments, it gives the lowest error in classifying the comments in various categories, and the rate of correctly predicting the specific category of comments will be the highest.

Table 3. Comparative results of categorization of comments

Algorithm	**SM**	**FE**	**Accuracy (%)**	**F1-score (%)**
MNB	Under-sampling	GloVe	69	68
MNB	Over-sampling	TF-IDF	70	70
SVM	Over-sampling	TF-IDF	83	82
RF	SMOTE	CV	90	90
CNN	No-sampling	Tokenizer	55	55

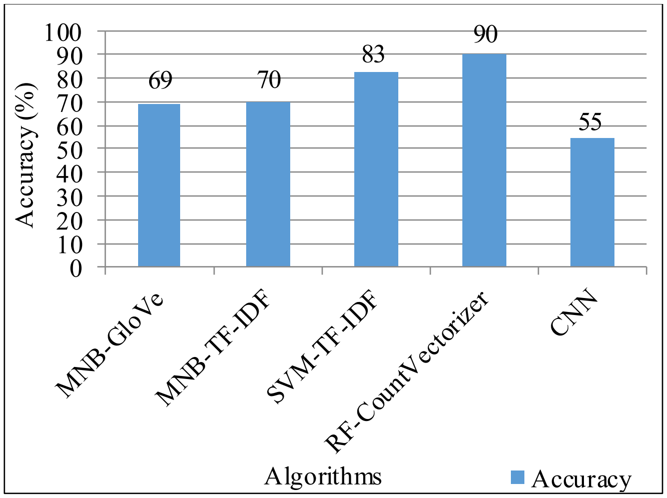

Fig. 7 Accuracy in categorization of comments

Figures 7 and 8 graphically depict the experimental results for accuracy and F1-score obtained in cyberbullying categorization, respectively. The results from Table 3, Figures 7 and 8, demonstrate that the RF algorithm with SMOTE sampling and CountVectorizer as a feature extraction technique has the highest values of accuracy as well as F1-score in comparison with other algorithms.

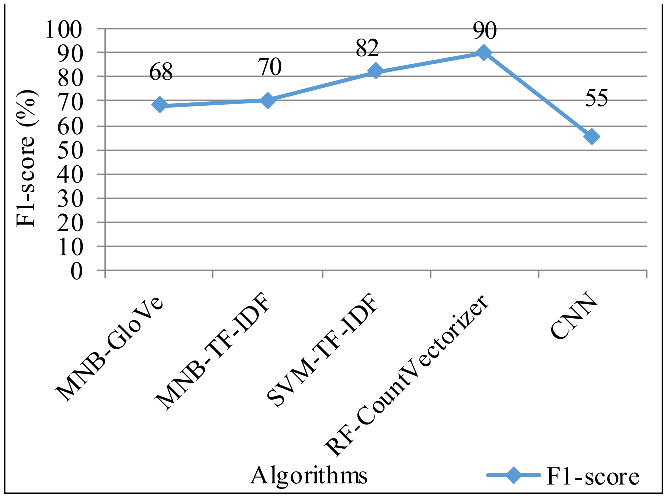

Fig. 8 F1-score in categorization of comments

SMOTE sampling helps in raising the count of samples belonging to the minority classes, which represent different cyberbullying categories. RF model makes use of an ensemble strategy to consolidate many decision trees with balanced data on cyberbullying to predict the category of bullying. This, in turn, results in the improvement of the performance of the categorization algorithm. Moreover, the CountVectorizer method becomes helpful in detecting the specific words corresponding to the various categories of cyberbullying. Therefore, the combination of RF, SMOTE and CV gives better results in categorizing the bullying as compared to other algorithms.

Thus, as the RF algorithm with SMOTE sampling and CountVectorizer technique shows the maximum accuracy and the F1-score, it is the best-suited technique for cyberbullying categorization in the Cb-DCS system.

4.2.3. Assessment of Accuracy and F1-Score in Severity Classification

For the severity classification of cyberbullying comments, the algorithms discussed as part of Section 3.9 are applied with various sampling and feature extraction techniques to identify the algorithm with the optimized performance with regard to accuracy and F1-score.

The values of the metric recorded inside Table 4 illustrate that SVM with SMOTE sampling and CountVectorizer as a feature extraction technique, gives the maximum accuracy among the other algorithms. This indicates that when SVM with SMOTE sampling and CountVectorizer technique is applied for severity classification on the set of bullying comments, it gives the lowest error in classifying the comments in distinct severity levels, and the rate of correctly predicting the specific severity level of comments will be the highest.

Table 4. Comparative results of severity classification

Algorithm	SM	FE	Accu-racy (%)	F1-score (%)
SVM	Over-sampling	TF-IDF	73	72
SVM	SMOTE	CV	87	87
RF	Over-sampling	CV	48	44
MNB	Over-sampling	TF-IDF	42	42
MNB	No-sampling	GloVe	50	44

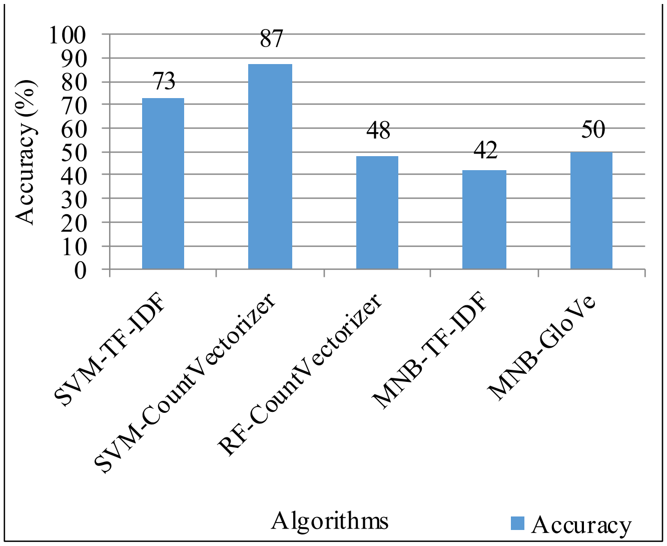

Fig. 9 Accuracy in severity classification

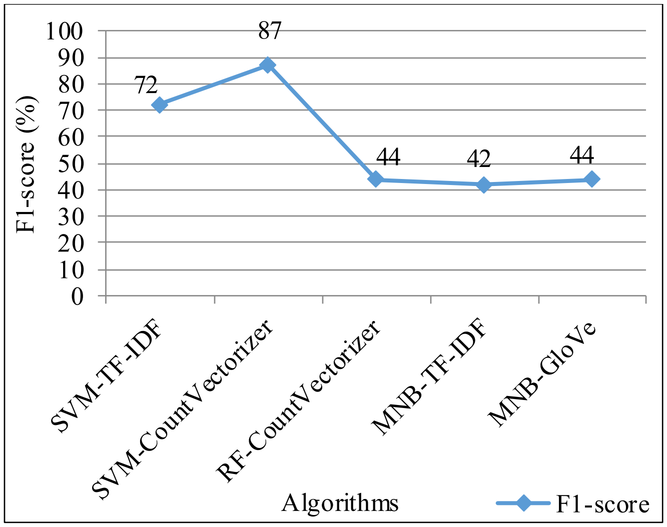

Fig. 10 F1-score in severity classification

Figure 9 and Figure 10 graphically illustrate the experimental results of accuracy, along with the F1-score obtained in the severity classification of cyberbullying comments, respectively. The results in Table 4, Figures 9 and 10, depict that the SVM with SMOTE sampling and CountVectorizer as a feature extraction technique has the highest values of accuracy as well as F1-score in comparison with other algorithms. SMOTE sampling is used to balance the data belonging to different classes, which represent the severity levels pertaining to the bullying comments. This enables the SVM model to learn and classify into the different severity levels of bullying in an optimal manner. This, in turn, contributes to enhancing the performance of the severity classification algorithm. Moreover, the CountVectorizer method turns out to be very useful in identifying the frequency of specific words, implying the various severity levels of cyberbullying. Hence, the combination of SVM, SMOTE, and CV gives better results in the severity classification of bullying comments than other algorithms.

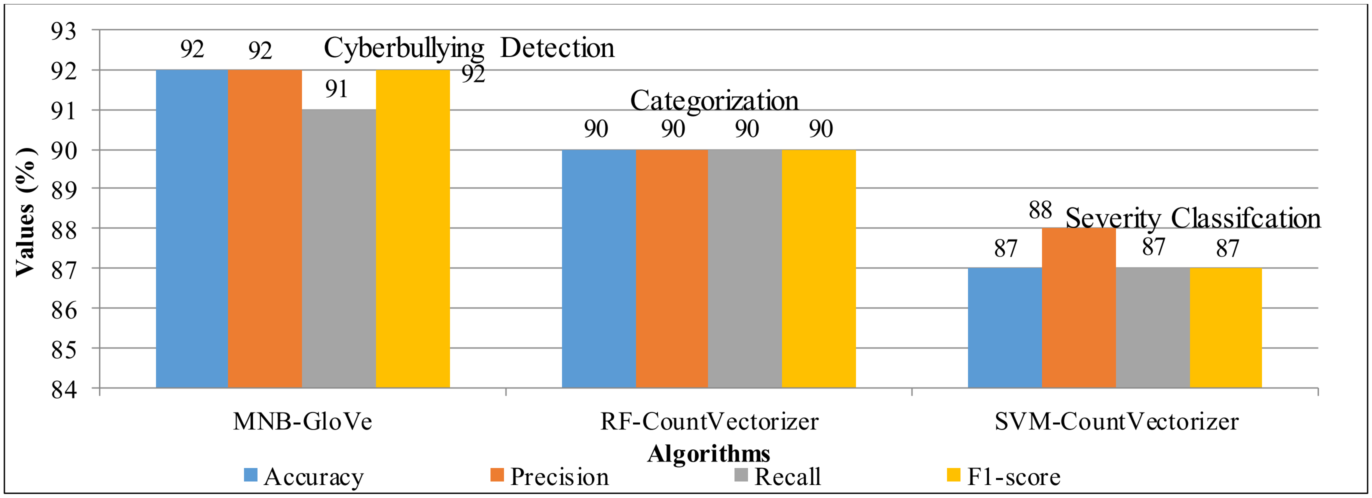

Fig. 11 Summary of optimal results in the Cb-DCS system

Thus, as the SVM with SMOTE sampling and CountVectorizer technique demonstrates the maximum accuracy and the F1-score, it is the best-suited technique for severity classification of cyberbullying comments in the Cb-DCS system. Figure 11 illustrates the summary of optimal results with regard to accuracy, precision, and recall along with F1-score pertaining to the above demonstrated best suited techniques in the Cb-DCS system for cyberbullying detection, categorization and severity classification.

4.2.4. Comparison of Results with Earlier Approaches

The performance of the proposed Cb-DCS system is compared with the other [9, 12, 13, 15, 16] cyberbullying detection approaches. Table 5 illustrates the results of this comparative analysis. Table 5 indicates that each of the approaches provides the detection of cyberbullying using distinct techniques. The sources of data for these approaches are specified in Table 1.

Table 5. Comparison of results with earlier approaches

Reference	Technique	Accuracy (%)	F1-score (%)
Muneer and Fati 2020 [9]	LR with TF-IDF and Word2Vec	90.57	92.8
Iwendi et al. 2023 [12]	*BLSTM	82.18	88
Hani et al. 2019 [13]	NN with TF-IDF, Sentiment analysis and n-gram	91.76	91.9
Saifullah et al. 2024 [15]	BanglaBERT	88.04	87.85
Alsubait and Alfage 2021 [16]	LR with CountVectorizer	---	78.6
Proposed system: Cb-DCS	MNB with GloVe	92	92

**BiLSTM is referred to as BLSTM in [12]*

The values of the metrics in Table 5 depict that the Cb-DCS system, which includes the MNB with the GloVe technique, outperforms the other approaches of detecting cyberbullying with regard to accuracy and F1-score. Moreover, unlike these other approaches, as described in the earlier subsections, the Cb-DCS system categorizes the bullying comments into specific types, such as religious or hate and classifies the bullying comments based on the severity, like low, medium or high.

5. Conclusion

A novel system to enable Cyberbullying Detection, Categorization, and Severity identification (Cb-DCS) for comments from networking platforms has been presented within this paper. The Cb-DCS system enables detecting the bullying comments, categorizing them according to the type, such as religious, racist or hate and classifying them as per the severity levels, like low, medium or high. The Cb-DCS system focuses on the text as well as on the emojis in the social media comments, which makes it a comprehensive one. For the purpose of experimentation, a dataset of comments consisting of text and emojis is created. As emojis can significantly alter the meaning of a sentence, in this work, the emoji2vec model is used to convert emojis to their corresponding sentiments. In each of the stages of detection, categorization and severity classification of cyberbullying comments, comparative analysis of various algorithms is made with different sampling and feature extraction techniques to identify the algorithm with the optimized performance with regard to accuracy and F1-score.

Experimental results have shown that for cyberbullying detection, the performance of the MNB algorithm with GloVe as a feature extraction technique is the highest among the other algorithms with regard to accuracy and F1-score. Whereas considering cyberbullying categorization, the RF algorithm with SMOTE sampling and CountVectorizer as a feature extraction technique has illustrated the optimal performance in comparison with other algorithms. The results have further demonstrated that for the severity classification of cyberbullying, the performance of SVM with SMOTE sampling and CountVectorizer as a feature extraction technique is superior to the other algorithms.

Thus, with the optimal algorithms included at each stage, the proposed Cb-DCS system enables the efficient detection, categorization and severity classification of cyberbullying comments across social networks such as WhatsApp, Facebook or Instagram. This implies that the Cb-DCS system can be effectively used to facilitate a positive online community and ensure the safety of social media platform users, especially teenagers and young adults. The number of comments can be increased in future for further experimentation. Also, images and videos, apart from comments, can be considered in future for the detection of cyberbullying.

References

[1] Samaneh Nadali et al., "A Review of Cyberbullying Detection: An Overview," *2013 13th International Conference on Intelligent Systems Design and Applications*, Salangor, Malaysia, pp. 325-330, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[2] Lulwah M. Al-Harigy et al., "Building Towards Automated Cyberbullying Detection: A Comparative Analysis," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] H. Rosa et al., “Automatic Cyberbullying Detection: A Systematic Review,” *Computers in Human Behavior*, vol. 93, pp. 333-345, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Seunghyun Kim et al., “A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1-34, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Peiling Yi, and Arkaitz Zubiaga, “Session-Based Cyberbullying Detection in Social Media: A Survey,” *Online Social Networks and Media*, vol. 36, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Saloni Mahesh Kargutkar, and Vidya Chitre, “A Study of Cyberbullying Detection using Machine Learning Techniques,” *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 734-739, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Vimala Balakrishna, Shahzaib Khan, and Hamid R. Arabnia, “Improving Cyberbullying Detection using Twitter Users’ Psychological Features and Machine Learning,” *Computers & Security*, vol. 90, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Syed Mahbub, Eric Pardede, and A.S.M. Kayes, “Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and its Impact,” *Security and Communication Networks*, vol. 2021, no. 1, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Amgad Muneer, and Suliman Mohamed Fati, “A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter,” *Future Internet*, vol. 12, no. 11, pp. 1-20, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Daniyar Sultan et al., “A Review of Machine Learning Techniques in Cyberbullying Detection,” *Computers, Materials & Continua*, vol. 74, no. 3, pp. 5625-5640, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sweta Agrawal, and Amit Awekar, “Deep Learning for Detecting Cyberbullying across Multiple Social Media Platforms,” *European Conference on Information Retrieval*, Grenoble, France, pp. 141-153, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Celestine Iwend et al., “Cyberbullying Detection Solutions Based on Deep Learning Architectures,” *Multimedia Systems*, vol. 29, no. 3, pp. 1839-1852, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] John Hani et al., “Social Media Cyberbullying Detection using Machine Learning,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 703-707, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Nijia Lu et al., “Cyberbullying Detection in Social Media Text based on Character-Level Convolutional Neural Network with Shortcuts,” *Concurrency and Computation: Practice and Experience*, vol. 32, no. 23, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Khalid Saifullah et al., “Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models Approach,” *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 11, no. 1, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Alsubait T., and Alfageh D., “Comparison of Machine Learning Techniques for Cyberbullying Detection on Youtube Arabic Comments,” *International Journal of Computer Science & Network Security*, vol. 21, no. 1, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Maral Dadvar et al., “Improving Cyberbullying Detection with User Context,” *European Conference on Information Retrieval*, Moscow, Russia, pp. 693-696, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[18] Andrea Perera, and Pumudu Fernando, “Accurate Cyberbullying Detection and Prevention on Social Media,” *Procedia Computer Science*, vol. 181, pp. 605-611, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Suliman Mohamed Fati et al., “Cyberbullying Detection on Twitter using Deep Learning-based Attention Mechanisms and Continuous Bag of Words Feature Extraction,” *Mathematics,* vol. 11, no. 16, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Arwa A. Jamjoom et al., “RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection with GloVe Features,” *IEEE Access*, vol. 12, pp. 58950-58959, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] T. Nitya Harshitha et al., “ProTect: A Hybrid Deep Learning Model for Proactive Detection of Cyberbullying on Social Media,” *Frontiers in Artificial Intelligence*, vol. 7, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Mohammed Al-Hashedi et al., “Cyberbullying Detection based on Emotion,” *IEEE Access*, vol. 11, pp. 53907-53918, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Akshita Aggarwal, Kavita Maurya, and Anshima Chaudhary, “Comparative Study for Predicting the Severity of Cyberbullying across Multiple Social Media Platforms,” *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 871-877, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Lu Cheng et al., “Xbully: Cyberbullying Detection within a Multi-modal Context,” *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Melbourne VIC Australia, pp. 339-347, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ammar Almomani et al., “Image Cyberbullying Detection and Recognition using Transfer Deep Machine Learning,” *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 14-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Mengfan Yao et al., “Cyberbullying Detection on Instagram with Optimal Online Feature Selection,” *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, Spain, pp. 401-408, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[27] Krishanu Maity et al., “Emoji, Sentiment and Emotion Aided Cyberbullying Detection in Hinglish,” *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2411-2420, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[28] Michael Aquino et al., “Toxic Comment Detection: Analyzing the Combination of Text and Emojis,” *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, Denver, CO, USA, pp. 661-662, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Hanh Hong-Phuc Vo, Hieu Trung Tran, and Son T. Luu, “Automatically Detecting Cyberbullying Comments on Online Game Forums,” *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Hanoi, Vietnam, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Alwin T. Aind, Akashdeep Ramnaney, and Divyashikha Sethia , “Q-Bully: A Reinforcement Learning based Cyberbullying Detection Framework,” *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[31] Fawzya Ramadan Sayed, Eman Hassan Elnashar, and Fatma A. Omara, “Cyberbullying Detection in Social Media using Natural Language Processing,” *Scientific African*, vol. 28, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] P. Vivekananth, and Navneet Sharma, “Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework,” *Indian Journal of Science and Technology*, vol. 18, no. 5, pp. 380-389, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[33] Mohammed Ali Al-Garadi et al., “Predicting Cyberbullying on Social Media in the Big Data Era using Machine Learning Algorithms: Review of Literature and Open Challenges,” *IEEE Access*, vol. 7, pp. 70701-70718, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[34] Despoina Chatzakou et al., “Detecting Cyberbullying and Cyberaggression in Social Media,” *ACM Transactions on the Web (TWEB)*, vol. 13, no. 3, pp. 1-51, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[35] Ron Artstein, “Inter-Annotator Agreement,” *Handbook of Linguistic Annotation*, pp. 297-313, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[36] Eloy López-Meneses et al., “Socioeconomic Effects in Cyberbullying: Global Research Trends in the Educational Context,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 12, pp. 1-29, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[37] Ángel Denche-Zamorano et al., “Science Mapping: A Bibliometric Analysis on Cyberbullying and the Psychological Dimensions of the Self,” *International Journal of Environmental Research and Public Health*, vol. 20, no. 1, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[38] Arpita Chakraborty, Yue Zhang, and Arti Ramesh, “Understanding Types of Cyberbullying in an Anonymous Messaging Application,” *Companion Proceedings of the Web Conference*, pp. 1001-1005, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[39] Ben Eisner et al., “Emoji2vec: Learning Emoji Representations from Their Description,” *arXiv Preprint*, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[40] Akshita Aggarwal et al., “Did You Really Mean What You Said?: Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings,” *arXiv Preprint*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[41] Md. Tarek Hasan et al., “A Review on Deep-Learning-Based Cyberbullying Detection,” *Future Internet*, vol. 15, no. 5, pp. 1-47, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[42] Monirah Abdullah Al-Ajlan et al., “Deep Learning Algorithm for Cyberbullying Detection,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, 2018. [CrossRef] [Google Scholar] [Publisher Link]