IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Automatic Detection Of Cyberbullying Using Machine Learning

Prof. Manali Patil[1], Sawmya Pandey[2], Samruddhi Yadav[3], Shreya Deshmukh[4], Bhavana Jagdale[5]

Assistant Professor, Department of Computer Engineering[1]

Students, Department of Computer Engineering[2,3,4,5]

Alard College of Engineering and Management, Pune, India

Savitribai Phule Pune University, Pune, Maharashtra, India

***Abstract*:** This paper provides a survey of the existing literature and research carried out in the area of using different models, methodologies, and frameworks on Cyber bulling detection. It provides insights on the overall process for cyberbullying detection and most importantly overviews the methodology. Though data collection and feature engineering process has been elaborated, yet most of the emphasis is on feature selection algorithms and then using various algorithms for prediction of cyber bullying behaviours. Finally, the issues and challenges have been highlighted as well, which present new research directions for researchers to explore.

***Index Terms*** - Cyberbullying, deep learning, machine learning, NLP, BERT Sentiment analysis, Hate speech detection

## I. INTRODUCTION

ICT has significantly impacted our lives, leading to an increase in cyber-hate crime, particularly on social media networks. Research is being conducted to combat this issue, with calls for providers to filter comments before publication. Manual text classification is time consuming and human-influenced, making machine learning (ML) approaches like classical ML, ensemble, and deep learning useful for hate speech detection. Natural language processing (NLP) has also shown superior outcomes in ML methods. To address this issue, it is crucial to review literature and keep professionals and researchers updated on current developments in this research area.

Abusive messages in social media are a complex phenomenon with a broad range of overlapping modes and goals. Cyberbullying and hate speech are typical examples of abusive languages that researchers have put more interest in the past few decades due to their negative impacts in our societies. Several research have been conducted to automatically detect these undesirable messages among other messages in social media. The automatic detection of hate speech using machine learning approaches is relatively new, and there are very limited review papers on techniques for automatic hate speech detection.

## II. RELATED WORK

The recent and related survey papers available on review of hate speech detection methods during this research work were few. The following were the available traditional literature review related to automatic detection of hate speech using MLA.

ML algorithms have contributed immensely in hate speech detection and SM content analysis generally. Offensive comments such as HS and cyberbullying are the most researched areas in NLP in the past few decades. ML algorithms have been of great help in this direction in terms of SM data analysis for the identification and classification of offensive comments. The advances in ML algorithms researches have made significant impacts in many fields of endeavour which led to some important tools and models for analysing a large amount of data in real-world problems like SMNs content analysis. In this survey conducted by, the authors presented a brief review on eight hate speech detection techniques and approaches

## III. MOTIVATION

The cases of hate speeches have become rampant due to the SM adoption by a large population. Researches have shown that hate speeches can influence political discourse and can change the narrative negatively[3]. It is of great importance to police the SMNs to allow democracy to take it natural cause without undue influence through hate speech spread. It is also obvious that countries where their democracy is still at the infant stages are more vulnerable in the face of hate speeches than those with matured democracy. Therefore, developing a hate speech detection system can help in keeping countries in mutual coexistence. Committing cyber hate requires just a smartphone, internet connection and a person with a corrupt mind[3][4]. The hate speech post can be escalated to every nooks and cranny in a matter of seconds.

Therefore, developing an effective hate speech detection on SM is of great significance. There is nothing the targeted person or group can do to stop the spread of this offensive post. To a reasonable extent now, SM is an integral part of our daily lives. It is necessary to fight the systematic racism rooted in almost all societies around the globe. This study is also a timely contribution in reducing hate speech on social media.

## IV. METHODOLOGY

The methodology used for this work is explained as follows. The following databases were mainly used to get the required articles for this review work: IEEE Explore, ACM, ScienceDirect, and Scopus. Key terms or phrases used in the search retrieval include hate speech detection, offensive comments, aggressive comments, cyberbullying, profanity and toxic comments on SM. The filter tools available in each database were used to filter the articles[5]. For instance, the subject was restricted to computer science, engineering, and mathematics. In this case, only the most relevant were downloaded after all filter tools have been employed. The second phase involves going through the abstract of each article to apply the inclusion or exclusion criteria.

## v. DATASET

The dataset used in this research was crawled and collected by various research paper's authors which is very popular among adolescents and has increasingly been used for cyberbullying studies. Only the English corpus was used for this research. The following table shows the class distribution of the datasets used for cyberbullying detection. There were other publicly available cyberbullying-related datasets available.

Dataset table:

| Topic Name | Used Dataset |
|---|---|
| 1. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges [6] | The search engines and academic databases used for the retrieval of relevant papers were as follows: Scopus, Clarivate Analytics' Web of Science, DBLP Computer Science Bibliography, ACM Digital Library, ScienceDirect, SpringerLink, and IEEE Xplore |
| 2. Cyberbullying Detection on Social Networks Using Machine Learning Approaches [3] | It contains collected Facebook comments from different posts and the twitter comments dataset from kaggle.com |
| 3. Ensemble Learning With Tournament Selected Glowworm Swarm Optimization Algorithm for Cyberbullying Detection on Social Media [2] | Twitter dataset |
| 4. Advance in Machine Learning Algorithms for Hate Speech Detection in Social Media [4] | The first fundamental problem is the availability of hate speech dataset across different regions of the world. |
| 5. Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches [5] | • It was an open dataset with significant size, with more recent corpus collection period.<br>• More variation of cyberbullying topics were covered in the corpus, such as curses, defamation, defense, insult, sexual, and threats. |
| 6. Identification and characterization of cyberbullying dynamics in an online social network [8] | My space dataset<br>clustering coefficient<br>post sentiment |
| 7. Cyber bullying detection using machine learning [1] | texts or images, natural language or images are non-numeric data. |

## VI. ALGORITHM AND CLASSIFIERS

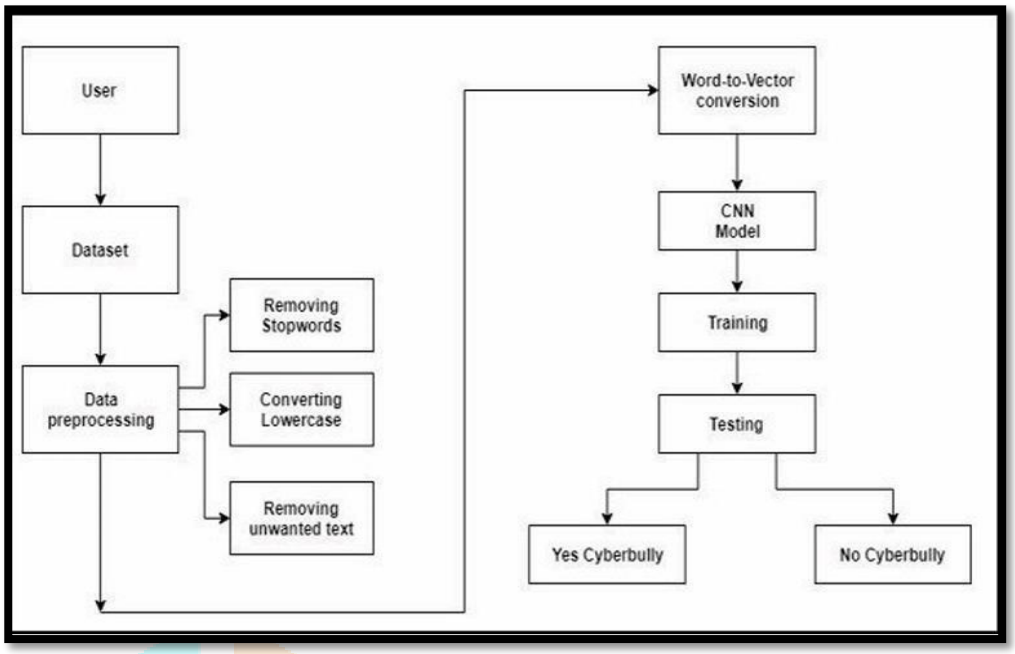


Fig a) Digram for algorithm and classifiers.

### CLASSICAL MACHINE LEARNING:

### A. DATA PREPROCESSING:

Data Processing Pipeline for Natural Language Toolkit.

Tokenization: Splits raw text into meaningful words using Regex Tokenizer.

Stemming: Converts a word into a root word or stem using Porter Stemmer.

Stop word Removal: Removes words that don't add meaning to a sentence.

### B. FEATURE EXTRACTION:

The following project studies three Feature extraction methods Bag of Words, TF-IDF and Word2Vec.

- Bag of Words model: The BOW that is bag of words model is a simple method of extracting features from documents that uses occurrence of words within a document.
- TF-IDF Model: The value of TF-IDF increases with the increase in frequency of a word in same document and decreases with decrease in frequency of documents that have the word in the corpus.
- Word2Vec: It is used to represent word in vector form. There are two methods for the construction of the word embeddings:
- Common Bag of Words Model (CBOW): Common Bag of Words model takes as input of multiple words and predicts the word based on the context.
- Skip Gram Model: The skip gram model is just the reverse of CBOW model in which multiple context words are predicted using a single input word.

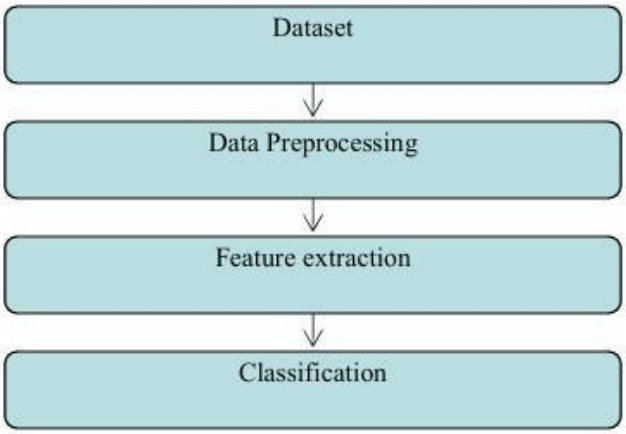


Fig b) Process for machine learning

FEATURE SETS : Texts are unstructured data, but ML algorithms use mathematical modeling. To convert them into structured feature spaces, unstructured data must be converted into a vector space. Noise, such as unnecessary numbers and common words, must be removed. Vectorization methods can then be used to convert the dataset into a vector space.

Detailed Features in Table

| Features | Description |
|---|---|
| N-gram | Used unigram, bigram and trigram as binary features. |
| Count | Tokenized the comments and count the occurrence of each token in it. This way we created a sparse matrix of NxV. |
| TF-IDF score | Used to calculate the importance of words in documents based on how frequently they are used. |
| Occurrence of pronouns | This is additional feature which helps in detecting cyber-aggressive comments based on pronounce “You”. |
| Skip-grams | Adds a long-distance word as a feature. Used to detect co-occurrences of some words like “You idiot”. |

Some of the approaches and methods which can be used for used for the hate speech detection are:

**CLASSICAL MACHINE LEARNING**: This approach is also called shallow method. This method relies on manually or automatically coded dataset that can be used for training purposes. This labelled dataset is used to train the learning algorithms to produce a model which can be used for detecting and classifying text as hate speech or non-hate[4]. Examples include support vector machines (SVM), Naive Bayes (NB), Logistic Regress (LR), Decision Trees (DT), K-Nearest neighbour (KNN), etc.

**DEEP LEARNING APPROACH**: CNN and LSTM networks are two most popular architectures: CNN is an effective feature extractor, whereas LSTM suitable for modeling orderly sequence learning problems. We also observe the use of LSTM or GRU with pretrained Word2Vec, GloVe, and fastText embeddings to

fed into a CNN with max-pooling to produce input vectors for a neural network[5]. CNN extracts word or character combinations, e.g., phrases, n-grams, and LSTM learn a long-range word or character dependencies in texts. ConvLSTM is a robust architecture to capture long-term dependencies between features extracted by CNN and found more effective than structures solely based on CNN or LSTM in tasks like Named Entity Recognition (NER).

**BERT MODEL**: The main activity of a BERT model is to generate word and sentence embeddings (inbuilt pooling) for input to classifiers. BERT has proved to give state-of-the-art results for many NLP related tasks and is used in Google search engines since 2018. As defined in, BERT is a technique of pretraining language representations, meaning that it is trained on a general-purpose "language understanding" model on a large text corpus (like Wikipedia), and then used for various downstream NLP tasks. Pre-trained embedding can either be contextual or context-free, and contextual embedding can further be categorized as unidirectional or bidirectional[5].

**NLP**: NLP is Natural Language Processing (NLP) which is a fascinating and rapidly evolving field that intersects computer science, artificial intelligence, and linguistics. NLP focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language in a way that is both meaningful and useful[3]. With the increasing volume of text data generated every day, from social media posts to research articles, NLP has become an essential tool for extracting valuable insights and automating various tasks.

## VII. CONCLUSION AND FUTURE WORK

Although hate speech as a social issue has long been studied in the humanities and arts, it is still relatively new in the field of computing. To keep researchers informed, it is necessary to continuously update them on developments or advancements. In order to identify hate speech on social media, we examined methods from deep learning, ensemble, and conventional machine learning. According to this paper, there is more research being done on the use of classical machine learning for hate speech identification.

This study aimed to investigate the features that might be created from text and provide insight into the methodological procedures for adopting textual features, sentiment, and emotional features, even if the textual feature is still widely utilized and mostly used to classify cyberbullying.

Cultural differences, pandemics or natural disasters, data sparsity, imbalance dataset issues, and dataset availability concerns are some of the outstanding challenges in hate speech identification that this article also identified.

There are certain restrictions on the current investigation. Even though the cyberbullying corpus includes input from several roles within cyberbullying episodes, our approach is restricted to binary text categorization and does not assist us in identifying the poster of the cyberbullying post.

Last but not least, the research on detecting cyberbullying should be expanded to include multilingual environments and examine context from additional metadata, including memes, photos, and videos. The researchers hardly ever used features from other media, like picture, video, time, and network embeddings.

## VIII. REFERENCES:

[1] Varun Jain, Vishant Kumar, Dinesh Kumar Vishwakarma and Vivek Pal, "Detection of Cyberbullying on Social-Media Using Machine Learning" IEEE 2021.

[2] Saloni Mahesh Kargutka and Prof. Vidya Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques" IEEE 2020 .

[3] Vikas S Chavan and Shylaja S S, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network" IEEE 2015.

[4] Nanlir Sallau Mullah and Wan Mohd Nazeem Wan Zainon,"Advance in Machine Learning Algorithms for Hate Speech Detection in Social Media" IEEE 2021.

[5] TEOH HWAI TENG AND KASTURI DEWI VARATHAN, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches" IEEE 2023.

[6] J. K. Peterson and J. Densley, "Is social media a gang? Toward a selection, facilitation, or enhancement explanation of cyber violence, Aggression Violent Behave", 2016.

[7] BBC. (2012). "Huge Rise in Social Media" [Online]. Available: http://www.bbc.com/news/uk-20851797

[8] P. A. Watters and N. Phair, "Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA), in Cyberspace Safety and Security. Berlin, Germany: Springer", 2012, pp. 6676.

[9] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions", IEEE Commun. Surveys Tuts., vol. 16, no. 4, pp. 20192036, 4th Quart., 2014.

[10] N. M. Shekokar and K. B. Kansara, "Security against cyber attack in social network", in Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES), 2016, pp. 15.