**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Comparison of Machine Learning and Deep Learning Techniques for Identifying Hate Speech on Various Social Media Platforms Using Diverse Data Sets

[1*]Rakesh Bharati, [2]Dr. Jyoti Bharti, [3]Dr. Vasudev Dehalwar

[1*]Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal 462003, India.
[2]Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal 462003, India.
[3]Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal 462003, India.

**Email Id:** * goswami.rakesh@gmail.com, jyoti_2202@yahoo.com, vasudevd@manit.ac.in.

**ARTICLE INFO**

Received: 22 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

**ABSTRACT**

This study examine task is to evaluate a variety of machine learning techniques and methodologies for the purpose of detecting the appearance of hate speech on social media (SM). This research was to investigate the essential components of hate speech classification using Machine Learning (ML) and Deep Learning (DL) techniques. Additionally, it explored the numerous obstacles that are experienced by different models. Generally speaking, the challenge of predicting hate speech is described as a task that involves categorizing text. It focused on five key areas such as feature extraction, dimensionality reduction, classifier development and selection, data exploration and collection, and model evaluation. Over time, the efficacy and efficiency of machine learning algorithms used to identify hate speech have significantly improved. There has been an influx of new performance measurements and datasets into the literature. A precise, thorough, and current state-of-the-art is required to educate researchers about new developments in automated hate speech identification. The findings of this study add up to three things. To begin, readers should be informed about the crucial procedures involved in hate speech identification utilizing machine learning algorithms. Second, the flaws and strengths of each technique are appraised to help researchers solve the algorithm Choice conundrum. Finally, significant research gaps and unsolved problems were discovered.

**Keywords:** Hate Speech, Deep Convolution Neural Network, Random Forest, Naïve Bayes, Social Network, Machine Learning.

## INTRODUCTION

These days, we expect the widespread availability of inexpensive Internet to potentially attract a large population to online social networking (OSN) sites, and every age group and every class of people are now active on the internet. According to data for January 2024, the number of active Internet users globally has drastically passed 5.04 billion, which is 62.3 percent of the global population [1]. OSN web sites drew a total of 3.8 billion unique visits, claiming eight out of every ten Internet users.

Because of the user-friendly nature of social networking websites (OSN), their platform has become a global communication platform (like Facebook, WhatsApp, WeChat, X, Instagram, etc.). [1]. Platform prevalence: Due to the availability of the wide range of content types, frequent daily updates (in audio, video, and image formats), and high production values of shows help to propagate these platforms in Figure 1. They cover politics, cinema, technology, science, music, space, wildlife, and so on. The web supports OSNs, which have an impact on all aspects of life. People are using online social networking sites (OSNs) with the functionality of fulfilling the specific needs of individuals [2], [3].

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

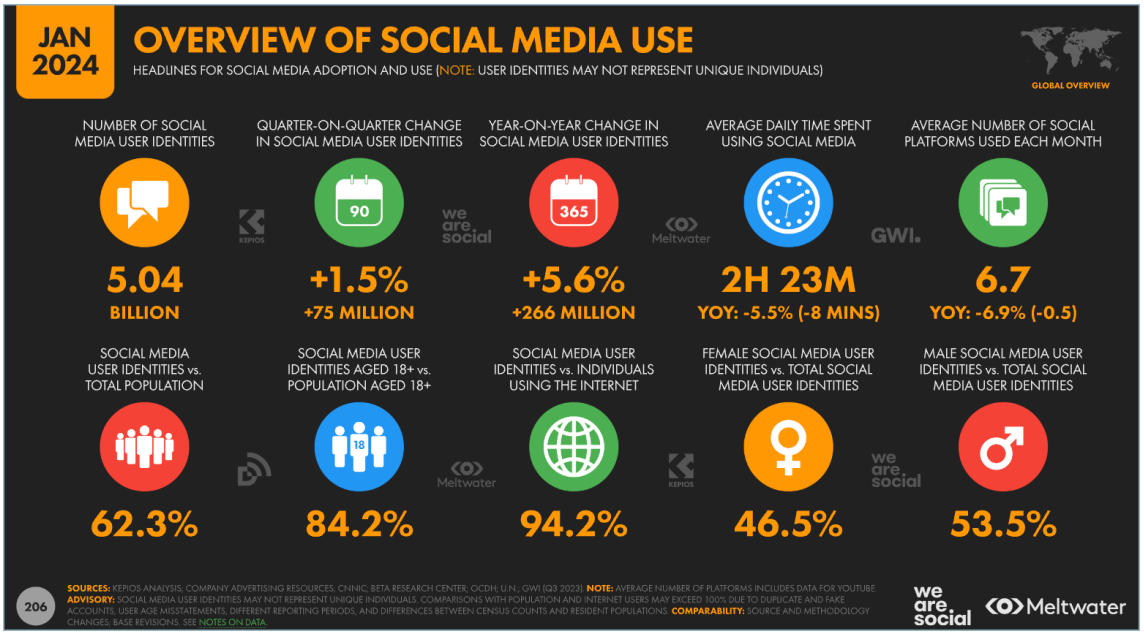


**Figure 1: Use of Social Media.**

Twitter is an online social networking service (OSN) that allows users to post microblog-style messages with a character restriction of 280. Finally, the various social networking web sites, including Twitter, provide an interface for making friend requests. In order to receive alerts, you must follow individuals such as friends, acquaintances, or your favorite celebrities. A user following your account must read the comments, which also serves as authentication [4]. In general, people use Twitter to stay updated about what's happening in the world and to get updates from people they know, both professionally and personally. Twitter's lack of constraints allows people to freely post any content, including critical remarks and hateful messages [5], [7]. Facebook and Twitter recently had a significant influx of hate speech-related messages concerning the ShaheenBagh demonstration in Delhi, India's capital. The demonstration began on December 11, 2019, opposing the “National Population Register” (NPR), the “Citizenship Amendment Act” (CAA), and the “National Register of Citizens” (NRC). During the COVID-19 epidemic, two hate speech-related tweets started to become viral on Facebook and Twitter (with the hashtags #HateSpeech). The issue highlighted above has attracted the attention of researchers in recent years, resulting in the creation of models that rely on ML and DL approaches [8] and [14]. Several tweets related to HS are still present on Twitter and shared across the platform, indicating that current models fail to meet the necessary requirements. Therefore, we decided to create a model that could efficiently capture a significant number of postings related to HS. Sentiment analysis, inquiry answering, document classification, phrase classification [15], spam filtering, and other text-related problems have all been effectively addressed by researchers using the CNN model [18] [20]. A “deep convolutional neural network” (DCNN) is used in numerous studies to tackle the problem of identifying hate speech. To extract a remark's semantic core, DCNN uses convolution operations on tweets. Additionally, some researchers examined the Convolutional-LLSTM (C-LSTM) and “Long Short-Term Memory” (LSTM) models and concluded that the DCNN method was the best option. This document offers a comprehensive overview of the model's key contributions, based on DCNN and C-LSTM. Using only tweets as input for prediction, the suggested DCNN model aims to lessen the need for human feature extraction. The suggested method achieved better accuracy and training time than competing models when presented with an imbalanced dataset.

## I. MOTIVATION AND RELATED WORKS

### A. MOTIVATION

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

A perceptible rise in the frequency of hate speech events has been brought about by the extensive usage of social media platforms and the global availability of internet access. Studies suggest that hate speech might impact political and corporate discussions and alter people's narratives [12], [13]. Social media networks (SMNs) need to implement regulations to safeguard democracy and businesses from the harmful effects of hate speech distribution. Young democracies are more susceptible to hate speech than established democracies, a trend that is also observable in the present global landscape. Hate speech detection technologies can subsidize to the maintenance of peace among nations.

All it takes to engage in cyber hatred is a smartphone, an internet connection, and a disturbed state of mind. A message promoting hate speech can rapidly spread across the entire internet in mere seconds. In order to spread and broadcast hate speech on social media networks, a specific physical place doesn't seem necessary. As a result, it is imperative that social media platforms have a system for monitoring hate speech. The designated receiver or group possesses a limited ability to impede the dissemination of this harmful communication [14].

Social media has now become a commonplace aspect of our lives to a certain degree [15]. Racism is a pervasive issue that impacts virtually all societies worldwide and necessitates immediate attention. One significant and urgent concern is the prevalence of hate speech on social media. This study provides an insight into various models employed by different authorities to promptly address this problem.

## RELATED WORKS:

Social media abuse is a dynamic and complex phenomenon, characterized by a variety of strategies and aims that often overlap with each other [17]. In recent decades, scholars have increasingly prioritized studying types of abusive language, such as cyberbullying and hate speech, due to their significant impact on our communities. Multiple experiments have been conducted to automatically detect these specific conversations amidst the numerous other messages on social media.

The term "HS" lacks any formal definition [21]. A "harmful statement" (HS) is a remark that causes damage to another person. However, few scholars have also used the phrase "hate speech" to describe HS in their research [23] [28]. Researchers used DL techniques with more traditional machine learning approaches, such as supervised machine learning, to identify hate speech on SM. However, the latter method was more commonly utilized. In the following sections, we will analyze recent research that has employed these two approaches. Our goal is to assess each strategy's benefits and drawbacks and provide a succinct breakdown of the particular situations when one technique is better than the other.

### (a) ML Methodology for Hate Speech Detection

The fields of AI and ML have had a significant impact on the identification of HS and the thorough examination of SM data. Studies on natural language processing have mostly examined HS and cyberbullying (CB) in recent decades [21]. In SM data, ML methods have shown to be quite helpful in identifying and classifying remarks that are inappropriate. Research on machine learning algorithms has resulted in the creation of important tools and models for addressing real-world problems, particularly in the realm of content analysis for social media networks [23].

This survey by [20] examined eight different strategies and techniques for hate speech identification. One of the eight approaches is the template-based approach; the others are the TF-IDF, dictionaries, N-grams, sentiment analysis (SA), bag-of-words (BW), and part-of-speech (PS) techniques.

Hate speech posters commonly target their victims based on attributes such as religious, racial, ethnic, political, physical appearance, poor lifestyle, and marital status. The exponential rise in the volume of data produced by social media networks led to the developing of the term "big data" [15].Out of the global population of 7.7 billion, the following estimated number of individuals are involved in online activities:

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

For this study, Warner and Hirschberg combed through Yahoo! and the American Jewish Congress's websites for data [26]. For the purpose of calculating the F1 score, recall, accuracy, and precision, the SVMlight classifier was employed [29]. Results of the ideal result was 0.68, 0.60, 0.64%, and 95%. To extract characteristics and classify tweets, Kwok and Wang used a Naive Bayes classifier using the Bag-of-Words method [25]. An optimal accuracy rate of 76% was achieved by the model using a 10-fold cross-validation configuration. The researchers had failed to classify HS-related tweets using the Bag of Words approach. They claimed it may be even better if the feature set included bi-gram and tweet emotion scores to increase accuracy.

Burnap and Williams [27] collected 450,000 tweets for their research. The n-gram (1–5) word characteristics that were taken from tweets were classified using the supervised model. The accuracy of classification of a selected ensemble classifier, support vector machines (SVM), along with Bayesian logistic regression was assessed using the data. It obtained the highest accuracy, recall, and F1-scores all rounded to the closest whole number by using the Voted Ensemble Classifier. A data set with labels of 16,000 cases provided by Waseem and Hovy [23] was employed to identify hate speech. The unigram, bigram, trigram, and quadgram characteristics of the tweets were assessed. Utilizing a ten-fold cross-validation setup, the logistic regression (LR) classifier obtained F1-scores of 73.89 percent, 73.66 percent, 73.62 percent, and 73.47 percent for each one member of the four feature sets.

Davidson et al. proposed an automated HS detection technique. After a thorough analysis, the data gathered from various sources was divided into three groups: offensive, hate speech (HS), and neither. Weighted uni-, bi-, tri-, and quad-gram features were extracted from the labeled dataset using the tf-idf approach. Each model's efficiency in shrinking the size of a 5-fold cross-validation dataset were assessed: random forest models, decision trees, linear SVMs, and naive Bayes.

In conclusion, utilized logistic regression with L2 regularization to classify the tweets, and their F1-score of 0.90 was in line with previous study results [9, 27]. Forty percent of HS patients had their classifications incorrect. Gao and Huang proposed a paradigm for HS detection. The logistic regression and LSTM models outperformed the baseline model (based on Char) by 3 and 4 percentage stages, each according to the disclosure.

### (b) An Approach to HS Detection Based on DL

Djuric et al. [28] examined the users' remarks using a model they created to identify HS. To create a low-dimensional representation of the comments, the CBOW and paragraph2vec algorithms were employed. Based on these features, the comments were classified as either clean or hostile. The classifier known as paragraph2vec demonstrated the highest level of success, attaining an “Area Under the Curve” (AUC) value of 0.80, which serves as an effective performance metric. Using logistic regression and a CNN model, Park and Fung [30] sorted tweets into non-hate and hate categories.

The researchers discovered that a combination of conventional ML classifiers and DL models yielded superior results. Zhang et al. [14] proposed a network architecture that integrates a gated recurrent unit (GRU) with a CNN to identify HS. Their model outperformed the other six when tested on seven datasets that were publicly accessible. The average F1-score increased by 1–14%. In simple terms, the model can understand and remember the information's semantic meaning and sequential nature. Using a combination of English and Hindi tweets, Kamble and Joshi [31] were able to identify cases of hate speech (HS). They used a large dataset with a variety of codecs mixed together to build the model.

The experimental findings demonstrated that the created code-mix embedding worked compared to the pre-trained word embedding. The experiment was carried out using a variety of supervised machine learning models, such as SVM, Random Forest, CNN-1D, LSTM, and Bi-LSTM. A number of indicators were used to assess the models' performance. These models outperformed all others in terms of accuracy, recall, and F1 score, surpassing 83.34. The highest F1 score was 80.85.

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

Researchers utilized conventional ML and DL methods to tackle the issues related to hate speech on Twitter. One-hot encoding was used to encode the features that were retrieved using the Bag-of-Words, n-gram, Tf-idf, and tf-idf techniques. Prior machine learning models for anticipating HS tweets required extensive feature building and a thorough understanding of the topic, making them time-consuming and demanding. The omission of one-hot encoded tweets from the perceptron model leads to a significant misclassification rate. This work uses a CNN-based model to get around these restrictions and improve prediction accuracy.

## II. METHODOLOGY

The techniques utilized in this investigation are delineated as follows: We primarily obtained the required publications for this review from databases such as IEEE Explore, ACM, Science Direct, Scopus, and UniversitiSains Malaysia. In doing the review, the researchers restricted their search for publications to a certain time frame of ten (10) years, specifically from 2010 to 2020. The researchers implement terms or phrases such as "hate speech detection," "offensive comments," and "aggressive comments" in their search.

The filtering options of each database were utilized to refine the findings. The subjects encompassed computer science, engineering, and mathematics. For instance. Filtering technologies were employed to guarantee the download of only the most pertinent files. At this juncture, we carefully examine each abstract and employ our criteria for determining what to include and exclude. Following the completion of the inclusion test, the articles were arranged by the date of publication. The paper must have directly addressed themes related to offensive statements made on social media mediums, such as HS, cyberbullying, aggressive remarks, poisonous remarks, and so forth. This is the main need for inclusion. Two elements of each paper the title and the abstract were used for this purpose.

## III. THE THEORY OF HS AND HS MODELLING

### THE THEORY OF HATE SPEECH

A group or a person may be the target of hate speech if they are attacked or exposed to prejudices because of protected traits or sensitive information [5]. Religious and ethnic connections, nationality, marital status, health, race, color, disability, sexual orientation, descent, gender, and other distinguishing characteristics are all protected characteristics. Every law-abiding citizen in the globe has realized that HS is an everyday reality and that it is everyone's shared enemy. Stopping people from engaging in this risky and unlawful activity is a top priority! Many of the hate speech communications on social media (SM) are made up of text messages. [32] As a result, hate speech is often accompanied by images and noises. [32] Text classification is the best way to tackle this problem from a computer standpoint. There is no universally accepted definition of HS, and no one phrase has been agreed upon [33].

It has been found that a more defined definition of HS simplifies the annotators' tasks and increases their agreement rate [34]. In some nations, it may be difficult to discern between hate speech and appropriate speech. Consequently, it is now more difficult to define HS as a phrase which is widely recognized. As an example, the United States' First Amendment lacks a clear distinction between HS and non-HS. Since hate crimes include any statement that incites criminal activity, The question of what qualifies as hate speech has been rekindled as a result of the global Black Lives Matter (BLM) movement. Following George Floyd's death, the BLM movement emerged.

Apart from HS, there are also other online behaviors that need to be explained, such as cyberbullying. Cyberbullying is described as "repeated aggressive conduct using SM in an attempt to purposely and regularly endanger or harm people who are powerless to defend oneself" and is a kind of cyber harassment [36, 37]. Online abuse can take many forms, including HS, as well as cyberbullying. Cyberbullying is a form of harassment.

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

It's hate speech when a victim's vulnerable or protected feature is the target of attack. Unlike cyberbullying, hate speech has implications for the entire community or society, not just one individual [18]. Both human beings and computer systems have struggled to grasp the full scope of what constitutes hate speech [40].

## HS MODELLING

For the purpose of this study, we are trying to figure out how many people are tweeting about some one on certain issue. There are two categories into whom tweets may be placed in this binary categorization issue, i.e. HS or NHS (NHS). The problem statement can be expressed mathematically as follows:If there are many tweets arriving, let's say { t1; t2; t3; tn} each of which may fall into one of the following categories:{ c1; c2; c3; ; cn.} There will be two classes, c1 and c2, where c1 represents tweets about hate speech and c2 represent tweets about everything else (NHS). To test the DCNN deep neural model, existing machine learning-based classifiers were utilized as a baseline. Tf-idf approaches were used to extract the characteristics needed for baseline models. However, with the aid of many filters in the intermediate layers, the suggested DCNN model used the convolution process to extract the tweet's primary characteristics from the text itself.

## V. HS CLASSIFICATION

Recent decades have seen extensive investigation into text classification, which has found practical applications such as the identification of hate speech. There is a growing interest among academics in developing applications that use text categorization algorithms, especially due to recent advancements in NLP and text mining. Figure 2 illustrates that categorizing hate speech using machine learning generally involves five parts: activities such as feature extraction, dimensionality reduction, classifier selection, collecting data and exploration, and evaluations.

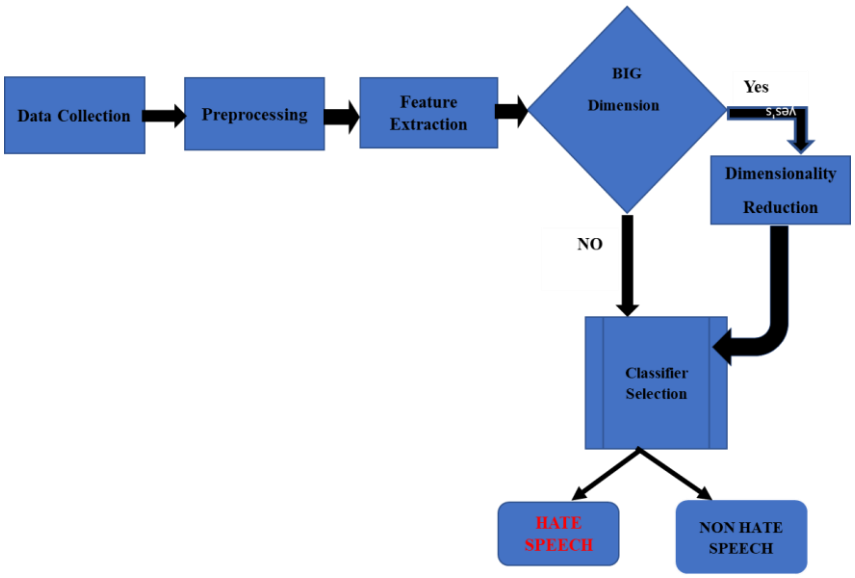


**Figure 2: Illustrates that categorizing hate speech using ML.**

## COLLECTION AND EXPLORATION OF DATA

At this stage, the researcher will make a decision in collaboration with colleagues on the methodology and location for data collection, which will be used to train the selected machine learning algorithm. A researcher can be lucky enough to find an existing dataset, or they might be unlucky enough to have to start from scratch while creating a new dataset. Two important factors should be considered when deciding whether to use an existing dataset or produce a new one: availability and relevance [42].

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

There is a possibility that the dataset is not accessible at all or is completely outdated. Given this situation, we must decide between generating a fresh dataset or modifying an already existing one. Generally, the procedure of generating a fresh dataset is a laborious and costly undertaking, although it is highly valuable and justifies the time and financial resources allocated to it. The importance of the dataset's relevance is crucial when choosing the dataset for constructing any predictive modeling model.

It must create certain criteria depending on the type of the issue we are attempting to address before labeling a dataset. The study can easily modify the dataset if it aligns with the initial research aim, as illustrated in [43] and [46]. Nevertheless, in the event that an antiquated and valuable dataset becomes inaccessible, it will be imperative to generate a fresh dataset.

## EXTRACTION OF A FEATURE

People often perceive texts as unstructured data due to their lack of organization or structure. A structured feature space must be created from the unstructured text input as all ML techniques inevitably include mathematical modeling within their algorithms [10]. Noise in the dataset, such as frequently used terms, non-English phrases, and irrelevant statistics, must be eliminated. Utilizing vectorization methods, the cleaned dataset may be converted into a vector space.

## DIMENSIONALITY REDUCTION

This is particularly true in the field of SM, where, as we move into the big data age, the volumes of data created is growing by the second. It is also true that it is becoming more and more difficult to find a meaningful trend in this enormous data collection because of the abundance of less important data [47, 48]. There are far more of these meaningless data points [49] than there are useful data points. The resulting data is known as high-dimensional data and is usually sparse and unevenly distributed over the search space. The term "curse of dimensionality" [50] describes how the excessive complexity of data makes it harder to identify patterns in the current big data age. To maximize the classifier's effectiveness, much of the worthless data must be eliminated or reduced to the bare minimum before using this dataset to train a model.

The solution to this difficulty is found in a technique known as dimensionality reduction. Every machine learning specialist works to eliminate any noise from the data and any characteristics that don't add to the learning value of the model. This effort may lead to further problems like overtying and data leaks. Insufficient data points lead to overtting, which in turn causes the classier to learn inadequately. The classier performs badly when presented with uncertain facts. This occurs when certain data from the training and testing datasets are found to be similar to one another when dividing the little amount of data available for cross-validation. This will give a very high accuracy, but the classifier will perform poorly when it is given a new dataset. Finding the crucial dimension of the relevant data set is the first step towards resolving this issue. The minimum characteristics needed to train a classifier and the capacity to predict with a high enough degree of accuracy are two important aspects of a data set [47, 48]. Generally speaking, the critical dimension prevents researchers from over tagging by preventing them from lowering the features in the features space. When using the dimensionality reduction approach, the classifier should be capable to gather enough from the decreased features to do the Classication or clustering operation as efficiently as possible.

## HS CLASSIFIER SELECTION

The HS issue is often modeled as a text classification tasks. To address the issue of hate speech classification, a number of classifiers are provided. Choosing the optimal classifier for the solution is one of the most crucial steps in the hate speech detection procedure. Consequently, in order to guide algorithm selection, a thorough conceptual understanding of each hate speech classifier must be acquired. Generally speaking, there are three types of ML: the DL technique, the ensemble approach, and the classical method [51]. The most important component of this study that we are concerned with

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

is the progress that has been made so far in these strategies. Table 1 shows a comparison of some relevant strategies that have been used in recent years.

**Table 1: Comparison of related techniques for hate speech dection.**

| Author | Classifier | Noval Contribution | Feature Extraction Technique | Performance Metrics |
|---|---|---|---|---|
| [52] | NV ,LG,RF,LG,DT,SVM,DL | Improvemenet of Lamophobia detection | Word Embedding | Accuracy,Precision ,Recall and F1 Score |
| [53] | DL | HS in Text | Embedding | Accuracy ,Recall Precision,F1-Score |
| [54] | Ensemble Method | Multi tier Meta Learning Method | Character n-gram and word n-gram method | Recall ,Precision,F1-Score |
| [55] | SVM ,NB,RF,DT | To Detect Arbic context based HS | BoW and TF-IDF | Accuracy, Recall, Precision and G-Mean |
| [56] | NB,LR,SVM,KNN,DT,RF | Address Code switch | TF-IDF | Confusion Matrix |
| [51] | LR and LSTM | Multi Lingual Analysis aspect of HS | BoW | F1-Score |
| [57] | RF | Improved RF for HS Detection | Count Vectors | F1-Score , Precision and Recall |
| [58] | Lexicon,RNN | The Building of Arbic Data Set | N-Gram, Embedding | F1-Score,Recall, Precision,AUROC |
| [59] | SVM , NB,RF | Emotional Analysis | N-Gram | Precision and Recall |
| [3] | RF , SVM and J48graft | Combination of 3 diffrerent dataset which gives wider coverage | Unigrams | Precision , Recall and F1-Score |
| [60] | n-gram word | Identifying Cyber Hate | BoW | Precision , Recall and F1 |

## CLASSICAL ML

The general technique is a popular name for this tactic. A dataset that has been manually or mechanically coded and can be utilized for training is the foundation of this method. It is suitable when the size of the dataset is modest. By training the learning algorithms on a classified dataset, this approach produces a model that can identify and categorize text as either HS or non-HS depending on its content. SVM, naïve bayes (NB), logistic regression (LR), decision trees (DT), random forests, and K-nearest neighbor (KNN) are a few supervised ML techniques that ML approaches that are most commonly employed for identifying hate speech. There are the most uses for SVM. Sorting social media

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

data into categories for HS and non-HS is the work assigned to researchers. Prior to the random forest, logistic regression is ranked third. In this instance, algorithms like regression, NB, and passive aggressive are also quite successful.

## ENSEMBLE APPROACH

The ensemble technique is a methodology that involves intelligently combining multiple weak models to create a powerful model. To clarify, the collective efficiency of several classifiers is consistently superior to that of the top-performing individual classifier [61]. In order to overcome the shortcomings of individual weak machine learning algorithms and enhance their individual capabilities, the ensemble technique was developed [62]. Without a doubt, each model has its own distinct limitations, which means that no model can be perfect. Ensemble approaches have limitations even if their goal is to combine the advantages of several models to outperform any one model alone [63]. It is feasible to lower their variability and significantly improve their learning capacity by incorporating several ML methods [64, 65]. Ensemble tactics such as random forest, bagging strategy, and boosting approach can be employed to enhance performance. Table 2 demonstrates that each of these solutions possesses distinct advantages and disadvantages in addressing hate speech tasks.

**Table 2: Advantages and Disadvantages of Ensemble Approach**

| Ensemble Technique | Pros | Cons |
|---|---|---|
| **Random Forest** | • Random forest surpasses most non-linear classifiers. This method is also quite resilient, as it relies on many decision trees to get its result.<br>• The random forest classifier avoids overfitting by averaging all predictions, eliminating biases and thereby overfitting.<br>• Missing values are no problem for random forests. So they can either use median values for continuous variables or compute the proximity-weighted average of the missing values.<br>• In this method, you may easily select the most influential features for your classifier. | • This algorithm is slower than others since it uses many decision trees to predict. Whenever a random forest classifier makes a prediction, every tree in the forest must vote on the same. This can take a long time.<br>• Random forest classifiers are sluggish and thus inappropriate for real-time predictions.<br>• Unlike a decision tree, where you can make a selection by following the tree's path, the model is difficult to interpret. But a random forest has numerous decision trees, thus that's not conceivable. |
| **Bagging** | • Several studies have demonstrated that it can reduce variation in a classification task.<br>• Using N learners of the same size on the same algorithm to deal with variance provides an environment, which may be used to deal with variance. | • In the case of bias or underfitting, this is not a desirable thing.<br>• The values with the highest and lowest results, which might make a substantial difference as well as having an average outcome, are sometimes disregarded. |
| **Boosting** | • Reduces the variance of the categorization as well as the bias of the classification. | • Calculation power is high. Noise is sensitive to the progression of time. |

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

| | | |
|---|---|---|
| | • can produce more trustworthy categorization results when applied.<br>• Every step of the way through the results, a record of net errors is preserved.<br>• Using this method, the weighing of the bigger sample accuracy and the smaller sampling accuracy is performed, and the cumulative performance is then calculated.<br>• When dealing with data sets that have bias or under-fitting, this can be extremely beneficial. | • Although the faults in the predecessors must be rectified by each classification method, the algorithm is susceptible to outliers.<br>• It's almost impossible to scale up at this point.<br>• It is capable of disregarding overfitting in the data set.<br>• The classification becomes more complex as a result of this.<br>• Time and calculation can be too expensive in some cases. |

## DEEP LEARNING METHOD

Standard machine learning algorithms are incapable of conducting effective analyses of certain text datasets due to their extreme size and lack of linear separability. When nonlinear data cannot be linearly separated, they are simply nonlinear data which is hard to depict on a hyperplane because of their nonlinearity. The DL method was developed to solve the previously described problem of forecasting significant trends in linearly non-separable data. [65] An extension of ANNs, the DL algorithm is a kind of ML technology. Its objective is to closely resemble the human brain. The complexity of the subject matter is the primary factor determining the profundity of the analysis. For instance, the utilization of an increased number of concealed layers is customary in image processing activities. The CNN and the Recurrent Neural Network (RNN) have garnered significant attention from researchers due to their superior ability to capture emotion semantics and phrase semantics in comparison to other models. CNNs have shown to be quite successful in textual content analysis particularly when it relates to capturing the syntax and semantics of words involved in the final construction of a sentence [68].

Various types of deep learning algorithms have been employed to accurately predict instances of HS on SM platforms. For Task 6 of the SemEval-2019 competition, [69] employed CNN along with two types of RNN, specifically Long Short-Term Memory (LSTM) and GRUs, to address the challenge. Identifying and classifying abusive language on social media is task six. The LSTM-CNN and CNN-LSTM models, two methods proposed by earlier researchers, were also used in this study's trials, and both produced promising outcomes. A thorough analysis revealed that BiLSTM-CNN produced a higher F1-score. Using three deep neural networks (DNN) for analysis, recent research looked at hate speech detection [71]. Several DNN variations, including CNNs, LSTMs, and FastText, as well as their combinations, were used in this investigation [71]. The study's performance was significantly improved, outperforming the state-of-the-art by around 18 points. The distinction between ML and DL is evident. As the learning graph shows, ML can operate with a smaller dataset, but DL requires a huge dataset to reach high learning levels. The red dotted line represents the DL algorithms' learning curve. The curve keeps growing along the performance-axis (vertical axis) due to the increase of data, and this expansion is indicative of the algorithm's effectiveness. Stated differently, DL performs better the more data there is ensemble approaches have their own set of pros and cons, which are outlined in Table 2. The blue line, on the other hand, represents classical machine learning, which suggests that the algorithm will probably stop learning even if the amount of data increases and that it will not converge until it hits a saturation point. Instead, than focusing on artificial intelligence, previous study on automated hate speech recognition has mostly employed traditional machine learning methods to identify various forms of hate speech on social media. Every second, the amount of data created on social media reaches tremendous parameters due to its exponential growth [72]. When the size of the data set exceeds a

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

certain threshold, it becomes necessary to utilize deep learning to address the problem at hand. DL for HS identification is a relatively new field with limited published research. Table 3 presents a comparison of DL algorithms for hate speech detection with alternative methods.

**Table 3: Comparative analysis of dl methods for detecting HS.**

| Author | Futures Extraction method | Deep Learning Algorithm | Evaluation metric | Aim of Study |
|---|---|---|---|---|
| [73] | word embedding | CNN | std deviations = 0.84 | To solve discriminatory problem |
| [17] | character ngram and CBOW | CNN and RNN | Pr=0.81, Rc=0.78, A=83, Fl=0.79, AUC =0.89 | To identify hate speech in Arabic Tweets |
| [74] | CBOW and Continuous skip-gram | CNN,LSTM,CNN+GRU | Fl=93.35 | To improve the performance |
| [71] | Char ngrams TFIDF BOWV | CNN and LSTM | pr = 0.93, Rc = 0.93, Fl = 0.93 | To classify a tweet as racist sexist or neither |
| [43] | NA | Deep LSTM | A= 90.82, Pr = 83.82, Rc =84.23 | Detection and explanation of |

## HATESPEECH DETECTION PERFORMANCE METRICS PARAMETER

Assessment markers generally approach performance assessment, an investigation topic encompassing numerous fields. Performance measurement criteria are logical-mathematical constructs derived from the calculation of the disparity between real and expected outcomes in a certain context [75]. Standard metrics for performance assessment for HS detection models encompass precision, recall, and F1-score, all of which are based on classical statistics. The irregular composition of the hate speech dataset renders it the most commonly utilized. The optimum choice when working with a balanced dataset is always accuracy. Comprehensive descriptions of precision, recall, and accuracy are given in [15], [65], and [76]. The F1-score assessment metric works effectively in imbalanced data sets. Assume that our algorithm has been trained to distinguish between hate post and non-hate post in tweets. For example, we have a batch of 40 tweets that contains 10 tweets that are considered hate speech (1) and 30 tweets that are not(0). 12 tweets were correctly identified as hate speech by the model. There were a total of 12 tweets identified, 8 of which were hate speech (true positives), and 4 of which were not (false positive). The model inaccurately classified four tweets (false negatives) that contained hate speech, while 26 tweets were suitably classified as non-hate speech (true negatives). This example can be illustrated by using confusion matrix in Table 4.

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

**Table 4: Confusion Matrix.**

| Actual \ Predicted | Predicted: NO | Predicted: YES | Total |
|---|---|---|---|
| Actual: NO (30) | TN = 26 | FP = 4 | 30 |
| Actual: YES (10) | FN = 4 | TP = 8 | 10 |
| Total (N=40) | 30 | 10 | 40 |

## PRECISION

Precision can be defined as the percentage of accurately predicted positive instances. Precision was used by the researchers in the following studies to evaluate the performance of their models.
This can be given and formulated by the equation as:

$$Pr = \frac{TP}{TP+FP} \qquad (1)$$

For the sake of this study, the letter Pr stands for precision.

The proportion of properly recognized positive classifications that the model correctly detected is known as precision [77]. For instance, in the aforementioned scenario, the percentage of true positives that were accurately detected is 8. Thus, 8/12 (truepositives / all positives) = 0.67 is the model precision. The abbreviation for true positive is TP. According to the preceding situation, TP is 8. The program successfully identified eight out of 10 tweets as hate speech.

The abbreviation for false positive is FP. This includes tweets that were labeled as offensive even if they weren't meant to be. In this particular example, four tweets were disregarded because they were categorized as HS when, in fact, they weren't.

## RECALL

The Recall (Rc) statistic measures the percentage of successfully recognized positive instances. [55], [57], [78], and [79] used recall as a method of evaluating their findings. Mathematically This can be reperesented as:

$$Rc = \frac{TP}{TP+FN} \qquad (2)$$

Rc is an abbreviation for Recall in this paper. The percentage of correctly determined actual positives might be referred to as a true positive. In this case, recall is computed as 8/10 (true positives / all positives) = 0.8. This suggests that 80 percent of the hate tweets were accurately identified by the program.

The abbreviation FN stands for false negative for the purposes of this research. This refers to HS tweets that did not include hate speech words and were thus not identified as such by the algorithm. Although these were hating tweets in the conventional sense, the model classified them as non-hate tweets. Just two tweets in the previous instance were mistakenly classified as non-HS when they were really HS.

## F-SCORE

The model is evaluated using a weighted average of accuracy and recall, and this is also known as the F-beta score. The weights between recall and precision are controlled by the beta parameter, the default value is 1 so the most common F-beta is F1 score. When a dataset is unbalanced, this evaluation measure is typically used to determine its quality. This can be formulated and given by the equation:

$$F = 2 * \frac{Pr*Rc}{Pr+Rc} \qquad (3)$$

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

When the distribution of classes is not uniform, the model's performance is evaluated using the F-measure, often known as the F1-score. Because an unbalanced class distribution is often seen in real-world text classification issues, the F1-score is a more suitable measure to employ when evaluating a model [51].

According to the previous example, F = 2*(0.67*0.8)/(0.67 + 0.8) = 1.072/1.47 = 0.72. For the sake of simplicity, the F1-measure of the model is 72 percent.

## ACCURACY

The ratio of accurate forecasts to total observations might be used to determine a prediction's accuracy.

It describes an effective and accurate model as one that is trained on a dataset that is nearly homogeneous, meaning that the values of FP and FN for the two-class classification issue are nearly equal. Accuracy is not the best choice when working with diverse and unbalanced datasets (i.e., if two classes are not proportionately well), thus alternative metrics and assessment measures, such the F1-score, should be used and properly reviewed. The correctness of the results was examined in the following studies: [45], [52], and [80].

Accuracy (A) can be expressed mathematically in the following way:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

## VI. WHAT THE STATE-OF-THE-ART HAS TO OFFER AND WHAT IT CANNOT DO

Table 5 summarizes the paper under consideration and identifies its strengths and weaknesses (contribution and limits).

Table 5 clearly identifies the subsequent research gaps, which require further rigorous and comprehensive inquiry. Numeric symbols, special characters, specific indications, and distinctive vocalizations that may convey hate speech messages were overlooked in all studies studied for this project and must be incorporated in the prediction of hate speech. A comprehensive coding guide benchmark is typically essential to assist annotators in effectively addressing this type of issue. Addressing contextual hate speech requires more comprehensive and effective research.

**Table 5: Relative works' contributions and drawbacks.**

| RELATED WORK | DATASET SPIRCE | CONTRIBUTIONS | DRAWBACKS |
|---|---|---|---|
| [52] | Twitter | • Annotation guidelines for the new dataset were developed by professionals.<br>• In order to assist annotators, a clear description of Islamophobia was provided.<br>• There was a high degree of inter-coder agreement, with an 89.9% accuracy rate | • • The data gathered was restricted to people in the UK who follow the most important politicians; this reduces the distribution.<br>• Restrictions in data collection led to a lack of heterogeneity.<br>• Only focus on Islam phobia; other hate-related aspects were not explored.<br>• The context of a word is irrelevant.<br>• Prior to the pre-processing, all numeric symbols were removed from the data. |

Research Article

| | | | |
|---|---|---|---|
| [53] | Twitter | • There was good diverse coverage of tweets<br>• The majority of hatred characteristics were taken into consideration | • In order to properly annotate a text, it is necessary to follow a proper guideline.<br>• Health status, marital status, and transgender status were not taken into account.<br>• As part of pre-processing, special characters and numeric symbols were eliminated. |
| [54] | Twitter | • Annotators who were specialists in South African politics were taught before tabulating the dataset, resulting in comprehensive coverage of a heterogeneous community. | • There are no additional languages in South Africa save for code-mixed included in the dataset.<br>• Pre-processing eliminated all numerical symbols.<br>• It was a bad idea to utilise annotators with an even number (i.e. 2) because the confused post may have one annotator express hate and the other non-hate. As a result, this can be an issue. |
| [45] | Face book | • Excellent illustrations of annotators' instructions.<br>• Extensive coverage of despise variables<br>• A thorough investigation into cyber-hate in languages other than English.. | • The published dataset was used, and as a result, the dataset inherited any flaws that were linked with it.<br>• The code-mixed date post was not taken into consideration.<br>• Numbers that some may see as having useful connotations were not taken into consideration.<br>• Gender identity and marital status were not taken into consideration.<br>• Only Spanish-language texts were examined. |
| [81] | Twitter | • Excellent cross-validation of up to ten variables.<br>• Excellent illustrations of annotators' instructions. | • Texts that were code-mixed were eliminated, which may have resulted in the loss of critical information.<br>• Numeric symbols, photos, and emojis were excluded from consideration.<br>• Data posts that were code-mixed were not taken into consideration. |
| [82] | Face book | • Good coverage of variables that people dislike.<br>• Cohen's kappak value was produced in order to test the agreement between codes. | • The transgender and married statuses were not taken into account.<br>• Numeric symbols that some may signifies beneficial significance were eliminated. |

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

| | | | |
|---|---|---|---|
| [9] | Twitter | • Comprehensive data on lexicon creation was gathered.<br>• Context and not simply bad words were taken into account.<br>• The annotators were guided by clear definitions and explanations. | • Pre-processing deleted numerical symbols, and there was no thorough guide to assist annotators. |
| [83] | Twitter | • Language switching amongst speakers of different languages was also considered.<br>• In the investigations, both multi-lingual and multi-dimensional issues were taken into consideration. | • Special characters and numeric symbols were not taken into account.<br>• The question of context in texts was not dealt with. |

## VII. UNSOLVED PROBLEMS IN DETECTING HATE SPEECH

Using standard ML techniques to identify hate speech on social media sites presents a number of unexpected challenges. These challenges may take many different forms, such as the length of time needed to formulate and define the issue, the length of time needed to gather and annotate the data, cultural differences, and other associated factors.

### HATE SPEECH DETECTION CHALLENGE AND DATASET

The initial and most significant obstacle is the absence of suitable hate speech data sets in a well-defined and homogeneous format across various platforms worldwide. This holds true everywhere in the globe. A large dataset is required for social media network research [60]. Additionally, [84] has emphasized how vital it is to broaden the concept of the campaign to prohibit hate speech to include both Western and non-Western parts of the world's population. The identification and classification of hate speech campaigns are greatly influenced by culture, religion, society, gender, and tradition. Table 6 depicts the availability of datasets in various parts of the world, categorized by region.

**Table 6: Comparison of the geographic distribution of the cyber-hate dataset and its availability.**

| Reference | Domain | SM Source | Availability | Dataset Source | Origin (Country) |
|---|---|---|---|---|---|
| [43] | General | Twitter | Available | Adopted [9] | USA |
| [52] | Specific (Politics) | Twitter | Available | New | UK |
| [44] | General | Twitter | Available | Adopted [9] | USA |
| [53] | General | Twitter | Not Available | New | Jorden |
| [54] | General | Twitter | Not Available | New | South Africa |
| [45] | General | | Available | Adopted [82] | Taiwan |

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

| | | | | | |
|---|---|---|---|---|---|
| [85] | General | Facebook/survey | Not Available | New | Germany |
| [81] | General | Twitter | Not Available | New | Spain |
| [46] | General | Twitter | Available | Adopted [10] | Pakistan |
| [3] | General | Twitter | Available | New | Japan |
| [86] | General | Twitter | Not Available | New | Portugal |
| [59] | General | Twitter | Not Available | New | India |
| [82] | General | Facebook | Available | New | Taiwan |
| [9] | General | Twitter | Available | New | USA |

### B.THE PROBLEM OF DATA SPARISM

The additional issue is the data's sparsity within the dataset. For example, Twitter limits each post to 140 characters [87]. In this scenario, the information included in a solitary tweet may not be adequate to draw broad conclusions about a specific post. This is a frequent problem that can be found in any short message text mining project**.**

### C.CHALLENGE OF UNBALANCED DATASET

As is inherent in the majority of real-world issues, the dataset's unequal class distribution is a common occurrence in hate speech identification [51]. A typical (non-hateful) post is much more common than an uncommon (cruel) one in the vast majority of cases [88]. Because the algorithm will learn more from information collected by the majority class (those that do not include hate speech) than from data representing the minority class (those that do), this will result in biased learning.

### D.DIFFERENT CULTURAL OPPORTUNITIES

Depending on the culture and tradition, cultural variances can have a direct impact on the definition of hate speech, or what constitutes HS. For instance, what is deemed normal communication in the US may be seen as HS in Nigeria. The way that speech is classified as offensive or non-offensive depends largely on the culture and traditions of the community. In order for social media companies to address the HS issue on their platforms holistically, experts recommended that non-Western parts of the globe be investigated for research pertaining to HS [13].

### E. PANDEMIC OR NATURAL DISASTER

Victims of pandemics or normal disasters often fall into stereotypes. The COVID19 pandemic is a good illustration of this, as Chinese people have been stereotyped in many parts of the world as a result of the outbreak. Figure 2 depicts a typical stereotypical tweet by former President Trump, as well as a response.

Figure 3 illustrates Trump's description of the COVID-19 pandemic as a Chinese-developed virus. Many people objected to such accountability measures. According to the most recent data available, Trump is the sixth most followed individual on the Twitter social media platform, with more than 87 million followers.

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

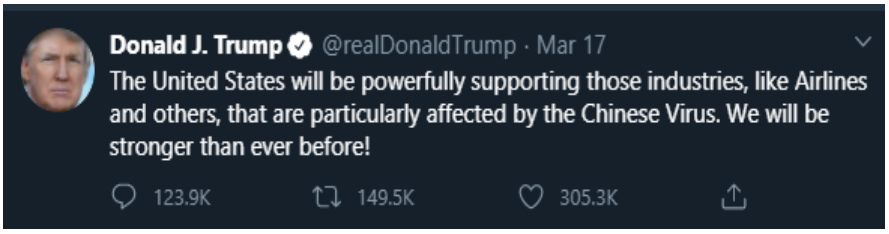


**Figure 3: Trump's description of the COVID-19 pandemic.**

With over 87 million followers, the volume of retweets, likes, dislikes, thumbnails, and comments is substantial, and the impact and influence can be profoundly detrimental to all Chinese individuals worldwide. The Centers for Disease Control and Prevention (CDC) has issued a stringent caution against designating diseases by geographical areas, asserting that this practice stigmatizes individuals. Each disease or calamity possesses an own nomenclature, complicating the process of identification.

## VIII. LIMITATIONS OF THE STUDY

The primary limitation of the study is the lack of experimentation on any ML or DL model using a specific data set. The research and innovative work of other researchers, on the other hand, was subjected to critical evaluation. In order to synthesize the work of other researchers, we created the conclusion that is presented in the following portion of the paper.

## IX. CONCLUSION

This article discusses the recent and latest development made in involuntary hate speech detection on SMP. Although hate speech research and innovation are well-established in the arts, humanities, and social sciences, they are relatively new to the technological and political fields, particularly on the social media platform. Therefore, it is essential to keep researchers updated on new developments. To detect HS content across various SM platforms, researchers employ a variety of methodologies, including conventional machine learning techniques, ensemble learning approaches for combining multiple weak models, and deep learning strategies. The research revealed a higher utilization of conventional ML compared to ensemble and DL methods. This research study highlights the need for further investigation into the application of ensemble and deep learning methodologies. The study also examines the benefits and drawbacks of each approach so that researchers may choose the best one. Moreover, some unresolved issues in hate speech identification were highlighted, including cultural differences, data scarcity, and dataset accessibility concerns. Encouraging the use of machine learning for high-risk detection on social media is crucial. This review targets beginners in hate speech classification in social sciences. It outlines the necessary steps for text categorization using machine learning and discusses the challenges in the field.

## REFERENCES:

[1] H. Watanabe et al.: “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions” IEEE Access Volume 6,2018

[2] F. M. Plaza-del-Arco et al.:” MTL Approach to HS Detection Leveraging SA” IEEE Access Volume 9,2021

[3] M.Luo,K.Wang,Z.Cai,A.Liu,Y.Li,andC.F.Cheang,“Using imbalanced triangle synthetic data for machine learning anomalydetection,”*Comput.,Mater.Continua*,vol.58,no.1,pp. 15–26,2019.

[4] M. S. Albarrak, M. Elnahass, S. Papagiannidis, and A. Salama, “The effect of Twitter dissemination on cost of equity: A big data approach,” Int. J. Inf. Manage., vol. 50, pp. 1–16, Feb. 2020.

[5] Cai, H. Xu, J. Wan, B. Zhou, and X. Xie, “An attention-based friend recommendation model in

social network," Comput., Mater. Continua, vol. 65, no. 3, pp. 2475–2488, 2020.

[6] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," IEEE Access, vol. 6, pp. 13825–13835, 2018.

[7] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, Sep. 2018.

[8] A. Guterres, "United nations strategy and plan of action on hate speech," United Nations, New York, NY, USA, Tech. Rep., 2019.

[9] Q. Li et al., A Survey on Text Classification: From Shallow to Deep Learning, vol. 37, no. 4. New York, NY, USA: Cornell Univ. Library, 2020.

[10] Q. Al-Maatouk, M. S. Othman, A. Aldraiweesh, U. Alturki,W. M. Al-Rahmi, and A. A. Aljeraiwi, "Task-technology fit and technology acceptance model application to structure and evaluate the adoption of social media in academia," IEEE Access, vol. 8, pp. 78427–78440, 2020.

[11] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, andD. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, pp. 1–68, 2019.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proc. 11th Int. Conf. Web Soc. Media (ICWSM), 2017, pp. 512-515

[13] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. NAACL Student Res. Workshop, 2016, pp. 88–93.

[14] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An appli- cation of machine classification and statistical modeling for policy and decision making," Policy Internet, vol. 7, no. 2, pp. 223–242, Jun. 2015.

[15] S. S. Bodrunova, A. Litvinenko, I. Blekanov, and D. Nepiyushchikh, "Con- structive aggression? Multiple roles of aggressive content in political dis- course on Russian YouTube," Media Commun., vol. 9, no. 1, pp. 181–194, Feb. 2021.

[16] F. Tulkens, "The hate factor in political speech. Where do responsibilities lie?" Polish Ministry Admin. Digitization Council Eur., Warsaw, Poland, Tech. Rep., 2013.R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," Comput. Hum. Behav., vol. 29, no. 1, pp. 26–32, Jan. 2013.

[17] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke,Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "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.

[18] M. Stegman and M. Loftin, "An essential role for down payment assis- tance in closing America's racial homeownership and wealth gaps the price of the homeownership gap," Urban Inst., Washington, DC, USA, Tech. Rep., 2021.

[19] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere," Appl. Sci., vol. 10, no. 23, pp. 1–16, 2020.

[20] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: A survey on multilingual corpus," in Proc. Comput. Sci. Inf. Technol. (CS IT), Feb. 2019, pp. 83–100.

[21] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1-10.

[22] A. Alrehili, "Automatic hate speech detection on social media: A brief survey," in Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA), Nov. 2019, pp. 1–6.

[23] A. Rodriguez, C. Argueta, and Y.-L. Chen, "Automatic detection of hate speech on Facebook using sentiment and emotion analysis," in Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC), Feb. 2019, pp. 169–174.

[24] G. Weir, K. Owoeye, A. Oberacker, and H. Alshahrani, "Cloud-based textual analysis as a basis for document classification," in Proc. Int. Conf. High Perform. Comput. Simul. (HPCS), Jul. 2018, pp. 629–633.

[25] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in Proc. 9th Int. Conf. Web Soc. Media (ICWSM), 2015, pp. 61–70, 2015.

[26] T. Granskogen and J. A. Gulla, "Fake news detection: Network data from social media used to predict fakes," in Proc. CEUR Workshop, vol. 2041, no. 1, 2017, pp. 59–66.

[27] L. Tamburino, G. Bravo, Y. Clough, and K. A. Nicholas, "From population to production: 50 years of scientific literature on how to feed the world," Global Food Secur., vol. 24, Mar. 2020, Art. no. 100346.

[28] V. S. Raleigh, "Trends in world population: How will the millenium compare with the past," Hum. Reprod. Update, vol. 5, no. 5, pp. 500–505, 1999.

[29] S. Paul, J. I. Joy, S. Sarker, A.-A.-H. Shakib, S. Ahmed, and A. K. Das, "Fake news detection in social media using blockchain," in Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC), Jun. 2019, pp. 1–5.

[30] World Data Lab. Population.io by World Data Lab. Accessed: Jan. 16, 2020. [Online]. Available:https://population.io/?utm_source=google&utm_medium=search&utm_campaign=population&campaignid=1695828135&adgroupid=64502612525&adid=329422103483&gclid=Cj0KCQiAjfvwBRCkARIsAIqSWlN28TwzgVkTSJkTgfnwPfK7fh96 cxYU3iglDqWphMuGFdiwTd-04dcaAgofEALw_wcB

[31] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li, and W. Cheng, "SocInf: Membership inference attacks on social media health data with machine learning," IEEE Trans. Comput. Social Syst., vol. 6, no. 5, pp. 907–921, Oct. 2019.

[32] J. van Dijck, "Governing digital societies: Private platforms, public val- ues," Comput. Law Secur. Rev., vol. 36, Apr. 2019, Art. no. 105377.

[33] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in Proc. SIGIR 42nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2019, pp. 45–53

[34] C. Ring, "Hate speech IN social media: An exploration of the problem and its proposed solutions" 2013

[35] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PLoS ONE, vol. 14, no. 8, pp. 1–16, 2019.

[36] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki,"Measuring the reliability of hate speech annotations: The case of the European refugee crisis," 2017.

[37] P. Smith, J. Mahdavi, M. Carvalho, and N. Tippett, "Aninvestigationinto cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying," Res. Brief, London, U.K., Tech. Rep. RBX03-06, Jul. 2006, pp. 1–69.

[38] M. Yao, C. Chelmis, and D.-S. Zois, "Cyberbullying ends here: Towardsrobust detection of cyberbullying in social media," in Proc. World Wide Web Conf. (WWW), 2019, pp. 3427–3433.

[39] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detectionin online communications: The experimental case of cyberbullying detec- tion in the Twitter network," Comput. Hum. Behav., vol. 63, pp. 433–443, Oct. 2016.

[40] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "A study on themethods to identify and classify cyberbullying in social media," in Proc. 4th Int. Conf. Adv. Comput., Commun. Autom. (ICACCA), Oct. 2018, pp. 1–6.

[41] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner,"Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," Psychol. Bull., vol. 140, no. 4, pp. 1073–1137, Feb. 2014.

[42] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resourcesand benchmark corpora for hate speech detection: A systematic review,"Lang. Resour. Eval., vol. 55, no. 2, pp. 477–523, Jun. 2021.

[43] C. C. Aggarwal and C. X. Zhai, "A survey of text classification algo- rithms," Min. Text Data, vol. 9781461432, pp. 163–222, Feb. 2012.

[44] A. Oma, T. A. El-Hafeez, and T. M. Mahmoud, Comparative Performanceof Machine Learning and Deep Learning Algorithms for Arabic Hate Speech Detection in OSNs, no. 1153. Cham, Switzerland: Springer, 2020.

[45] W. Dorris, R. Hu, N. Vishwamitra, F. Luo, and M. Costello, "Towards auto-matic detection and explanation of hate speech and offensive language," in Proc. 6th Int. Workshop Secur. Privacy Anal., Mar. 2020, pp. 23–29.

[46] M. Moh, T. S. Moh, and B. Khieu, "No 'love' lost: Defending hate speechdetection models against adversaries," in Proc. 14th Int. Conf. Ubiquitous Inf. Manag. Commun. (IMCOM), Jan. 2020, pp. 1–6.

[47] Z. Mossie and J.-H. Wang, "Vulnerable community identification usinghate speech detection on social media," Inf. Process. Manage., vol. 57, no. 3, May 2020, Art. no. 102087.

[48] M. Sajjad, F. Zulifqar, M. U. G. Khan, and M. Azeem, "Hate speechdetection using fusion approach," in Proc. Int. Conf. Appl. Eng. Math. (ICAEM), Aug. 2019, pp. 251–255.

[49] D. Suryakumar, A. H. Sung, and Q. Liu, "Determine the critical dimensionin data mining (experiments with bioinformatics datasets)," in Proc. 11th Int. Conf. Intell. Syst. Design Appl., Nov. 2011, pp. 481–486.

[50] N. Sharma and K. Saroha, "Study of dimension reduction methodologiesin data mining," in Proc. Int. Conf. Comput., Commun. Autom., May 2015, pp. 133–137.

[51] E. N. Sathishkumar, K. Thangavel, and T. Chandrasekhar, "A novelapproach for single gene selection using clustering and dimensionality reduction," vol. 4, no. 5, pp. 1540–1545, 2013, arXiv:1306.2118. [Online].Available: https://arxiv.org/abs/1306.2118

[52] L. Nanni, S. Brahnam, C. Salvatore, and I. Castiglioni, "Texture descrip- tors and voxels for the early diagnosis of Alzheimer's disease," Artif. Intell. Med., vol. 97, pp. 19–26, Jun. 2019.

[53] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? Thechallenging case of long tail on Twitter," Semantic Web, vol. 10, no. 5, pp. 925–945, Sep. 2019.

[54] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobichatespeech on social media," J. Inf. Technol. Politics, vol. 17, no. 1, pp. 66–78, Jan. 2020.

[55] H. Faris, I. Aljarah, M. Habib, and P. Castillo, "Hate speech detectionusing word embedding and deep learning in the arabic language con- text," in Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods, Jan. 2020, pp. 453–460.

[56] O. Oriola and E. Kotze, "Evaluating machine learning techniques fordetecting offensive and hate speech in South African tweets," IEEE Access, vol. 8, pp. 21496–21509, 2020.

[57] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo,M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," J. Inf. Sci., May 2020, Art. no. 016555152091765. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/0165551520917651

[58] E. Ombui, L. Muchemi, and P. Wagacha, "Hate speech detection in code- switched text messages," in Proc. 3rd Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT), Oct. 2019, pp. 1–6.

[59] K. Nugroho, E. Noersasongko, M. Purwanto, A. Z. Fanani, andR. S. Basuki, "Improving random forest method to detect hatespeech and offensive word," in Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT), Jul. 2019, pp. 514–518.

[60] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2018, pp. 69–76.

[61] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in Proc. 7th Brazilian Conf. Intell. Syst. (BRACIS), Oct. 2018, pp. 61–66.

[62] P. Burnap and M. L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," EPJ Data Sci., vol. 5, no. 1, pp. 1–5, Dec. 2016.

[63] A. Géron, Hands-On Machine Learning With Scikit-Learn and Tensor- Flow: Concepts. Tools, and Techniques to Build Intelligent Systems. CA, USA: O'Reilly Media, 2017.

[64] M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, and J. L. FernándezAlemán, "Reviewing ensemble classification methods in breast cancer," Comput. Methods Programs Biomed., vol. 177, pp. 89–112, Aug. 2019.

[65] P. Montebruno, R. J. Bennett, H. Smith, and C. V. Lieshout, "Machine learning classification of entrepreneurs in British historical census data," Inf. Process. Manage., vol. 57, no. 3, May 2020, Art. no. 102210.

[66] Z. Ding, Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and their Application in Bioinformatics. Atlanta, GA, USA: Georgia State Univ., 2011.

[67] A. Suárez-García, M. Díez-Mediavilla, D. Granados-López,D. González-Peña, and C. Alonso-Tristán, "Benchmarking of meteorological indices for sky cloudiness classification," Sol. Energy, vol. 195, pp. 499–513, Jan. 2020.

[68] Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, S. Nasrin, and V. K. Asari, "Comprehensive survey on deep learning approaches," 2017.

[69] L. Aristodemou and F. Tietze, "The state-of-the-art on intellectual property analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data," World Pat. Inf., vol. 55, pp. 37–51, Dec. 2018.

[70] Y. Zhang, Q. Wang, Y. Li, and X. Wu, "Sentiment classification based on piecewise pooling convolutional neural network," Comput., Mater. Continua, vol. 56, no. 2, pp. 285–297, Jan. 2018.

[71] R. Ong, "Offensive language analysis using deep learning architecture," 2019.P. M. Sosa, "Twitter sentiment analysis using combined LSTM-CNN models," Academia.edu, Jun. 2017.

[72] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proc. 26th Int. Conf. World Wide Web Companion (WWW) Companion, 2017, pp. 759–760.

[73] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," Social Netw. Comput. Sci., vol. 2, no. 2, pp. 1–15, Apr. 2021.

[74] S. Zimmerman, C. Fox, and U. Kruschwitz, "Improving hate speech detection with deep learning ensembles," in Proc. 11th Int. Conf. Lang. Resour. Eval. (LRED), 2019, pp. 2546-2553.

[75] P. Kapil and A. Ekbal, "A deep neural network based multi-task learn- ing approach to hate speech detection," Knowl.-Based Syst., vol. 210, Dec. 2020, Art. no. 106458.

[76] A. Botchkarev, "New typology design of performance metrics to measure errors in machine learning regression algorithms," Interdiscipl. J. Inf., Knowl. Manage., vol. 14, no. 113, pp. 13–21, 2019.

[77] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbully- ing detection on Twitter using big five and dark triad features," Personality Individual Differences, vol. 141, pp. 252–257, Apr. 2019.

[78] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect- based sentiment analysis: A comparative review," Expert Syst. Appl., vol. 118, pp. 272–299, Mar. 2019.

[79] C. Bhagat and D. Mane, "Survey on text categorization using sentiment analysis," Int. J. Sci. Technol. Res., vol. 8, no. 8, pp. 1189–1195, 2019.

[80] B. Zhang, S. Zhou, L. Yang, J. Lv, and M. Zhong, "Study on multi-label classification of medical dispute documents," Comput., Mater. Continua, vol. 65, no. 3, pp. 1975–1986, 2020.

**Research Article**

[81] M. Dholvan, A. K. Bhuvanagiri, and S. M. Bathina, "Offensive text detec- tion using temporal convolutional networks," Int. J. Adv. Sci. Technol., vol. 29, no. 6, pp. 5177–5185, 2020

[82] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, andM. Camacho-Collados, "Detecting and monitoring hate speech in twitter," Sensors, vol. 19, no. 21, pp. 1–37, 2019.

[83] Z. Mossie and J.-H. Wang, "Social network hate speech detection for amharic language," in Proc. Comput. Sci. Inf. Technol., Apr. 2018, pp. 41–55.

[84] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in Proc. Conf. Empirical Meth- ods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 4667–4676.

[85] K. Dong, T. Guo, X. Fang, Z. Ling, and H. Ye, "Estimating the num- ber of posts in SinaWeibo," Comput., Mater. Continua, vol. 58, no. 1, pp. 197–213, 2019.

[86] C. Wilhelm, S. Joeckel, and I. Ziegler, "Reporting hate comments: Inves- tigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation," Commun. Res., vol. 47, no. 6, pp. 921–944, 2019.

[87] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of hindi-english code-mixed social media text for hate speech detection," in Proc. 2nd Workshop Comput. Modeling People's Opinions, Personality, Emotions Social Media, 2018, pp. 36–41.

[88] C. Udanor and C. C. Anyanwu, "Combating the challenges of social media hate speech in a polarized society: A Twitter ego Lexalyt- ics approach," Data Technol. Appl., vol. 53, no.4, pp. 501–527, Oct. 2019.