

## Research Article

---

# Journal of Computational Sciences & Informatics

---

## Anomaly Detection in Social Media Conversation Using Natural Language Processing with LSTM and Naïve Bayes Models

Babatunde A.N.<sup>1</sup>, Shuaib B.M.<sup>2</sup>, Kadri A.F.<sup>3</sup>, Isiaka O.S.<sup>4</sup>, Abdulrahman, T.A.<sup>5</sup>, Ismail, S.I.<sup>6</sup>  
& Oke A.A.<sup>7</sup>

<sup>1, 2, 3, 4</sup>Department of Computer Science, Kwara State University, Malete. Kwara State, Nigeria.

<sup>3, 5, 6</sup>Department of Computer Science, Kwara State Polytechnic, Ilorin. Kwara State, Nigeria.

<sup>7</sup>Department of Computer Science, Federal College of Education, Iwo. Osun State, Nigeria.

**E-mails;** akinbowale.babatunde@kwasu.edu.ng, shuaib.mohammed@kwasu.edu.ng,

akeem.kadri@kwasu.edu.ng, isiakaosalman2@gmail.com,

abdulrahman.t@kwarastatepolytechnic.edu.ng, ismail.shola@kwarastatepolytechnic.edu.ng,

okeaa@fceiwo.edu.ng,

**Corresponding Author:** Babatunde Akinbowale Nathaniel

### ABSTRACT

Social media has in fact increasingly become the dominant form of communication within the globe. While these platforms are now used widely, they have also resulted in the rise of problematic or discursive and offensive material which has serious implications for online safety. To tackle this issue, this paper presents a hybrid method combining Natural Language Processing (NLP) for data preprocessing, Long-Short-Term-Memory (LSTM) with Naïve Bayes (NB) model to detect anomalies in the conversational style of users on social media. The NLP was used to process raw tweet data which was taken from the Kaggle dataset containing 24,783 instances with 6 features and processed to a format that is ideal for analysis. Naïve Bayes and LSTM models were then trained to detect abnormal or problematic messages such as hate speech and offensive language after data cleaning and transformation. Naïve Bayes, which served as a fast, probabilistic baseline model for the classification of the text was complemented in this research with the ability of the LSTM model to provide deep contextual understanding as well as sequential patterns that occur in the conversation. The model's performance metrics followed standard Machine Learning practices, and the best result was presented by the hybrid LSTM-NB model with an accuracy of 99.2% while the NB had 90% test result and LSTM 95% test score. This impressive result highlights the benefits of deploying NLP, integrated with both traditional machine learning and deep learning techniques to deal with the issue of detecting anomalies on online conversations. This paper aims to add to the NLP and AI literature, by offering an economic model to make digital communication safer.

**Keywords:** Natural Language Processing, Tweet, Kaggle, Machine Learning, Long-Short-Term Memory, Naïve Bayes.

---

### ACity FCSI Journal Citation Format

Babatunde A.N., Shuaib B.M., Kadri A.F., Isiaka O.S., Abdulrahman, T.A., Ismail, S.I. & Oke A.A. (2024): Anomaly Detection in Social Media Conversation Using Natural Language Processing with LSTM and Naïve Bayes Models. Journal of Computational Sciences & Informatics. Academic City University College, Accra, Ghana. Vol 4 No. 2 June, 2024 Pp 95-110. dx.doi.org/10.22624/AIMS/FCSIJ/2024/N2P7

---

## 1. INTRODUCTION

Social media technologies have changed the communicative and informational practices of our world. Sites including Twitter, Facebook, Instagram and Reddit accumulate millions of posts by users on a daily basis that capture a wide range of human opinions and experiences expressed in language. Yet, in addition to the meaningful content, social media platforms are also spaces for dangerous, damaging, or malicious behaviour such as fake news, cyberbullying, hate speech, and spam. In light of the potential hazards of these anomalies to both society and the affected individual, it is therefore, important to plan for strong surveillance and control mechanisms. Social media anomaly detection is an emerging area that looks to uncover such abnormal or potentially suspect behavior that is out of the norm.

This capability is essential for the integrity of social platforms, the protection of users from toxic content, and ultimately for the safety and trustworthiness of digital spaces (Hassani et al., 2020). Pathak et al. (2019), suggest that social media conversation seems to be difficult to monitor in terms of anomaly detection since data are huge, content is wide-ranging, and users' behaviour is evolving. Conventional forms of anomaly detection using rule-based systems or simple statistical models may not be able to cope with the highly complex, variable data which characterize social media. For instance, it may contain slang, abbreviations, and emoticons which are not fully understood by a traditional system (Pathak et al., 2019). In addition, the amount of data produced daily requires automated processes to allow real-time interpretation of the content.

The latter fact means that there is a need for more sophisticated detection and monitoring that is not just capable of spotting established forms of anomalies that are in use, but it can also be responsive to newer, emergent threats. This, in turn, implies also a deep understanding of the content and types of contexts in which these media are produced. Natural Language Processing (NLP) and Deep Learning (DL) have been found to be productive approaches for dissecting and comprehending social media (Lazebnik & Iny, 2024). NLP allows computers to decipher and process human language, thus enabling efficient analysis of large amounts of textual data. This is helpful in tasks like sentiment analysis, topic modeling, and text classification used to extract patterns of user behavior and general mood across platforms (Jurafsky & Martin, 2023). In Contrast, it is likely that NLP is not enough to completely grasp the subtleties and newness of relevant discussions on social media (Bender & Koller, 2020).

The drawback is that such a system has difficulty detecting newly emergent or complex anomalies (Rajesh & Hiwarkar, 2023). To address this problem, NLP is combined with Deep Learning architecture, a branch of Artificial Intelligence (AI) which specializes in solving complex problems with large, high-dimensional data (Islam et al., 2024; Young et al., 2018). Unlike many traditional machine learning methods that often require prior manual feature extraction (Goodfellow et al., 2016), machine learning methods based on deep learning, among which are RNNs (Recurrent Neural Networks), CNNs (Convolutional Neural Networks), and transformers, are able to automatically find and learn difficult structures in raw data (Lazebnik & Iny, 2024; Goodfellow, 2016).

These models are especially successful for text analysis as they are able to model contextual meaning and longer-range language dependence. This means they improve NLP to detect

known, as well as emerging and new, threats in real time (Islam et al., 2024; Delvin et al., 2019).

In this paper, we investigated the possibilities of using NLP and deep learning for anomaly detection in social media. The goal of our solution consists in combining the best of both technologies and create a system that is able to detect hate speech and offensive language in a more accurate and efficient way. The findings, in addition to helping advance the field of anomaly detection, also contribute to a more general interest in safety and trust in online environments. Our goal is to improve the technology for anomalous behaviour detection so that social media platforms can do a better job of protecting their users and securing the health of their communities.

Hence, the purpose of this study is to design a social media conversation anomaly detection system, which specifically focuses on hate speech and offensive comments, based on Natural Language Processing and deep learning techniques. So, the specific objectives of this paper were set as follows: Collect a dataset from the Kaggle repository, preprocess the dataset using NLP library (Natural Language Toolkit (NLTK)), develop Naïve Bayes (NB) and LSTM architectures, combine the two models to form hybrid LSTM-NB system for detecting hate speech and offensive languages on social media platforms, then measure the accuracy and loss of the proposed system using state-of-the-art metrics.

In the remaining part of this paper, section 2 presents some existing and related work to the present study. The description of the dataset and methodology used for the development of the proposed system was presented in the third section while section 4 explained the result of the research alongside the discussion of findings. Section 5 presents the conclusion and challenges for the future.

## **2. RELATED WORKS**

The field of Natural Language Processing (NLP) and Deep Learning (DL) has taken major strides in analyzing and detecting social media data in recent years (Kumar & Jaiswal, 2021). Among other scholars, different methods for identifying hate speech, offensive language, or other types of outliers in online discussions have been investigated (Vidgen et al., 2019; Fortuna & Nunes, 2018, Zhang et al., 2018; Davidson et al., 2017; Schmidt & Wiegand, 2017; Waseem & Hovy, 2016). The following sections summarize current research, methods, and models related to sentiment analysis, anomaly detection, and monitoring of social media, which serve as a background to our own research.

Sentiment analysis research by Samuel et al. (2020) examined how people responded to the COVID-19 epidemic on Twitter. To categorize tweet sentiment into positive, negative, and neutral groups, the researchers used two traditional machine learning algorithms: Naive Bayes (NB) and Logistic Regression (LR). According to their investigation, the Naive Bayes model outperformed the Logistic Regression model in terms of classification accuracy, reaching 91%. This was particularly true for short-form content like tweets, which frequently include informal language and acronyms.

The study demonstrated how well NB handles high-dimensional and sparse textual data, which makes it ideal for examining vast amounts of social media information produced during public health emergencies. Kumar et al. (2020) examined in more detail how demographic variables like age and gender affect the type and tone of customer feedback. Their research shows that these factors might have a substantial impact on customer

subjective preferences and linguistic style in addition to the sentiment polarity conveyed in reviews.

The researchers showed that knowing user profiles may improve sentiment classification models' accuracy and provide more in-depth understanding of consumer behavior patterns across various segments by integrating demographic variables into their study. They used a number of machine learning methods, such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), and the deep learning-based Long Short-Term Memory (LSTM) network, to assess model performance.

The Bag-of-Words (BOW) approach for feature extraction was used to create the standard models, NB, ME, and SVM. This method interprets each review as a frequency-based representation of word occurrences. The LSTM model, on the other hand, made use of word2vec, an embedding method that records the semantic connections between words in a continuous vector space. The results showed that when combined with the BOW method, the NB, ME, and SVM models performed better than the rest in terms of classification accuracy. In this specific setting, the LSTM model with word2vec embeddings performed no better than the best conventional models, while providing a more context-aware representation of language.

The study came to the conclusion that, even if deep learning has the potential to capture linguistic nuances, simpler models with well-designed features may still perform competitively, especially when demographic information are included in the analysis. Zarisfi et al. (2020) carried out sentiment analysis tests with two machine learning models, Multinomial Naive Bayes (MNB) and Support Vector Machines (SVM), utilizing TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction. After applying these models to four distinct Twitter datasets, they developed a scoring system that represented characteristics in vector space using TF-IDF, tweet class, and semantic similarity. The findings demonstrated that SVM, improved by this new scoring technique, scored better than MNB in three of the four datasets, suggesting that SVMs are especially useful for sentiment analysis in data from social media.

Zahoor and Rohilla, (2020), used of a number of Machine Learning classifiers, including Random Forests (RF), Long Short-Term Memory (LSTM) networks, Support Vector Machines (SVM), and Naïve Bayes (NB) to tackle sentiment analysis issues. They concentrated on applying the N-gram feature extraction method, which identifies word sequence patterns in text. According to the results, the NB classifier outperformed the other models in sentiment classification tasks, achieving the greatest accuracy across several datasets. This finding highlights the robustness of the Naïve Bayes approach, especially when paired with the N-gram method, in handling sentiment analysis on diverse datasets. In a bid to further enhance investigations in this domain, Mukherjee, Sharma, and Gupta (2021) suggested an individualized algorithm for identifying explicit negation in sentiment analysis. They trained NB (Naïve Bayes), SVM algorithms and ANN on Amazon reviews and various other kinds of Machine Learning algorithms.

They concluded that among the other models, the ANN with negative classifier obtained the best accuracy 96.32%. For customer sentiment classification, (Noori, 2021) developed a new methodology using data collected from reviews of hotels around the world. He trained different models such as SVM (Support Vector Machine), ANN (Artificial Neural Networks), NB (Naïve Bayes), K-NN (K-Nearest Neighbour), (Decision Tress) DT and C4.5 using TF-IDF

extraction and the DT model revealed the best accuracy totaling 98.9%, outperforming the other models

Similarly, more findings were conducted in 2023. Ahmed and Ahmed (2023), used TF-IDF, Random Forest (RF) and Naïve Bayes (NB) to categorize fake news articles sentiments as positive or negative. NB was the top-performing classifier with an accuracy of 89.30%. NB Classifier with TF-IDF feature extraction was used by Gaur, Sharma, and Kapoor (2023) for Twitter Sentiment140 Dataset. This model obtained better performance than the Latent Semantic Indexing (LSI) model with an accuracy of 84.44%, which reflects the good performance of the NB algorithm in sentiment classification tasks. Similarly, A comparison research was carried out by Qi and Shabrina (2023) to assess how well different sentiment analysis techniques performed in identifying COVID-19-related tweets from major English cities.

Along with a number of machine learning classifiers, such as Multinomial Naive Bayes (MNB), Random Forest (RF), and Support Vector Classifier (SVC), the researchers also explored lexicon-based techniques, especially Vader and TextBlob. Their results showed that the machine learning models, especially MNB, RF, and SVC, performed better than the lexicon-based approaches (Vader and TextBlob). They found that SVC with TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction was the most successful model for this specific job, achieving the best accuracy in categorizing COVID-19 tweets among the approaches used. In a study on Saudi cruises, Al Sari, Bukhari, and Alghamdi (2022) created three datasets based on social media commentary reflective of public attitudes.

They employed several machine learning techniques such as Multilayer Perceptron (MLP), NB (Naïve Bayes), Voting, SVM (Support Vector Machine), RF (Random Forest) with n-grams as feature extraction method. The performance obtained when oversampling Snapchat data is 82%, with the RF algorithm achieving the best accuracy. Fong, Zhuang, and Khoury (2023) experimented with various pre-processing filters in order to understand what factors do affect SA (Sentiment Analysis) algorithms. Interestingly, they found that opinion-trending words are significantly informative in the prediction. In particular, the elimination of high frequency words led to a drop in prediction accuracy especially for high frequency unique words related to each sentiment.

The selected papers reviewed in this study shows that sentiment analysis can be efficiently conducted with different machine learning algorithms, such as Naïve Bayes, SVM, Random Forest, ANN, Decision trees, including the use of advance feature extraction techniques like TF- IDF, and N-grams. While some of these models reported good accuracy, such as the studies with the scores of 84.0%, 89.30% and 98.9%, their performance generally presented some limitations. In particular, a large number of these findings failed to properly exploit Natural Language Processing (NLP) techniques in the preprocessing of the data, that is, they either failed to carry out tasks such as word lemmatization, stop-word removal, and the handling of negations, or had limited or no success.

In their computational models, the absence of the role of linguistic context and sentiment may hinder their capacity to model this sort of complexity. Also, even having obtained good results, none of the studies included achieved an accuracy of 99% or 100%, which reveals a gap regarding a perfect predictive performance. This would indicate that more in depth pre-processing and hybrid or deep learning techniques could help to improve sentiment classification effectiveness.

### 3. METHODOLOGY

This research shows the methodology design for this study as depicted schematically in Figure 1. The process requires collecting tweet data that includes hate speech, offensive language and other toxic content from Kaggle. It is then subjected to data pre-processing using Natural Language Processing (NLP) techniques. Text cleaning was carried out by removing URLs, @mentions and special characters, as well as tokenization, stop word removal, lemmatization, text vectorization (TF-IDF for Naïve Bayes and word embeddings for LSTM) and label encoding. The LSTM, Naïve Bayes architectures were implemented separately and were later combined using ensemble stacking technique to form the hybrid system. Finally, the models were evaluated in terms of accuracy, precision, recall, and F1-score.

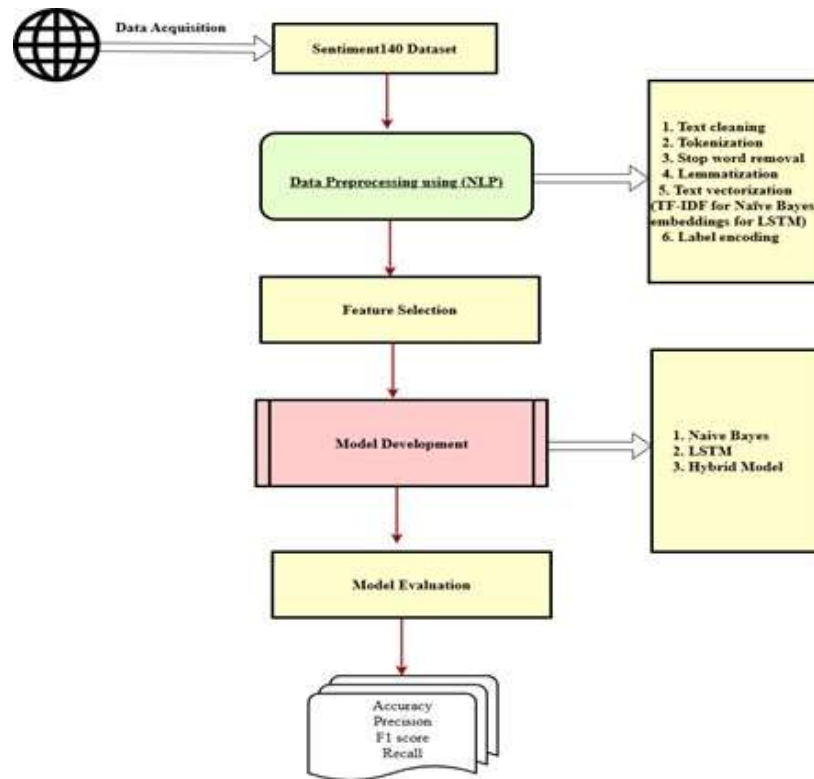
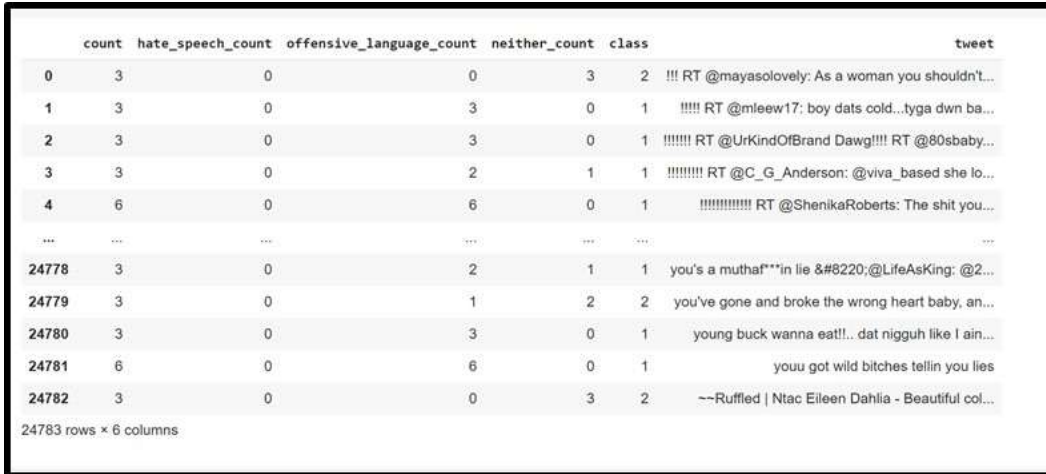


Figure 1: Graphical Representation of The Methodology Design for The Study

#### 3.1 Data Description

The dataset obtained for this experiment consisted of 24,783 Twitter instances, with six (6) features, designed to detect hate speech and offensive language. It includes column heads for the number of hate speech or offensive tweets as well as neutral tweets, along with a class label indicating whether a tweet is a hate speech, offensive or neither. We are primarily

interested in the tweet text for training the algorithms (Naïve Bayes and LSTM) used in this paper. Figure 2 shows a sample of the data in the Jupyter Notebook environment.



	count	hate_speech_count	offensive_language_count	neither_count	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	3	0	3	0	1	!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	3	0	3	0	1	!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	0	2	1	1	!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
...	...	...	...	...	...	...
24778	3	0	2	1	1	you's a muthaf***in lie &#8220;@LifeAsKing: @2...
24779	3	0	1	2	2	you've gone and broke the wrong heart baby, an...
24780	3	0	3	0	1	young buck wanna eat!!... dat nigguh like I ain...
24781	6	0	6	0	1	youu got wild bitches tellin you lies
24782	3	0	0	3	2	--Ruffled   Ntac Eileen Dahlia - Beautiful col...

24783 rows x 6 columns

**Figure 2: Sample Of 24,783 Tweet Data Employed for This Experiment**

### 3.2 Data Preprocessing

Data preprocessing is always one of the key practices in any Machine Learning (ML) research and was performed in this study. It was necessary to clean and organize the raw tweet data so it could be used to train algorithms mentioned in this work (Naïve Bayes and LSTM). The first data cleaning step during this experiment was to eliminate or strip out unwanted tweet contents like URLs, Twitter mentions (like “@username”), hashtags, special characters and numbers, which are sources of noise in any textual analysis. We imported Python Regular Expression (RE) to achieve this. After cleaning the dataset, we proceeded to check for missing values. We identified a total of 101 missing cells. Data filling is also crucial since machine learning model do not accept missing cells during training. We filled the missing cells using imputation technique. This was carried out using pandas’ data manipulation library. After data filling follows the tokenization phase.

The tweets data were tokenized, meaning they were split into individual words, so the data could be analyzed at the word level using the Natural Language Toolkit (NLTK) library. Next, stop word removal was performed by incorporating the stop words corpus from the NLTK, a line of code that eliminates unwanted common words such as “the,” “and,” “in” etc. that do not add to the sentiment or meaning of the text. The texts were then lemmatized, words were reduced to their base or root form (such as “running” to “run”) using spaCy, another NLP library, which helps standardize and make the words consistent. All of the texts were vectorized for the Naïve Bayes model by converting the tweet texts into numbers, using a method called TF-IDF, (Term Frequency-Inverse Document Frequency).

TF-IDF Vectorizer converts the text into a sparse matrix based on the terms’ frequency in the text and also in order to take into account how important those terms are. The LSTM model used word embeddings, or dense vectors that capture the semantic meaning of words, and Gensim, which provided pre-trained word vectors such as Word2Vec. Finally, label encoding with scikit-learn’s LabelEncoder was used to convert the categorical target labels (“hate speech”, “offensive”, “neither”) into numerical values appropriate for model training. The data was prepared to generate both the Naïve Bayes model and the LSTM model using each

of these preprocessing steps. We further split the processed data into 80% for training and 20% for models' evaluation.

### **3.2 Feature Selection**

In text-based classification tasks, identifying useful features is an essential step in creating efficient and accurate machine learning models. In this study, the dataset sourced contained these variables: hate speech count, offensive language count, neither count, tweet, count and class. However, through analytical evaluation using chi-square, only the most impactful features were retained to enhance the models' predictive ability. The Tweet variable which contains the actual text data, was selected and the class as the target feature. The Class feature contains (0, 1 and 2), this feature contains a multi-class machine learning problem. Selecting just the important variables for the model training helped improve the model overall accuracy and helps in classifying content as hate speech, offensive language or neutral.

### **3.3 Models Selection**

In this experiment, two different types of machine learning algorithm were chosen to develop the anomaly detection system: Naïve Bayes (NB) and Long Short-Term Memory (LSTM) Neural Networks. We selected these algorithms because they are widely used for sentiment analysis and text classification and have shown good performance for a variety of natural language processing (NLP) tasks. Naïve Bayes (NB) is a probabilistic classifier that applies Bayes' Theorem and assumes that words or tokens are conditionally independent of each other given the class label. However, despite its crudeness, (NB) performs well on text-based tasks, in large part because it handles large feature spaces efficiently. It computes the likelihood that a tweet belongs to particular class (hate speech, offensive or neutral) by tallying the words and their associations with those classes in the training data. NB is fast, simple and effective for sentiment analysis especially when trained on carefully preprocessed text features like those TF-IDF vectors.

Long Short-Term Memory (LSTM), on the other hand, is a type of Recurrent Neural Network (RNN), that learns long-term dependencies in sequences. LSTM models have a particularly clever architecture that includes input, forget and output gates which means that unlike more common RNNs, these networks can carry information over many time steps. This fact also renders LSTM architectures well-suited for analyzing the sentiment and content of text, which depends on context and the sequence of words within a sentence. We employed LSTM algorithm to analyze word sequences in tweets, via the word embeddings, so the model could discover relationships among words that depend on their context within text. LSTM is powerful because it can understand nuance and context, deal with the sequence of words in a text and the dependency of one word or phrase on those that come before it, as well as model the fact that words often have nonlinear relationships to each other.

Since NB is fast, simple and efficient in sentiment analysis, we combine these two techniques (NB and LSTM) to form a very advanced and best model for detecting hate speech and offensive language in social media using tweet data. These models were combined together using the ensemble stacking technique. The algorithms of the Naïve Bayes model and Long Short-Term Memory are presented in (Witten et al., 2011; Sebastiani, 2002; Rennie et al., 2003) and (Gers et al., 2000; Chollet, 2018; Hochreiter & Schmidhuber, 1997) respectively.

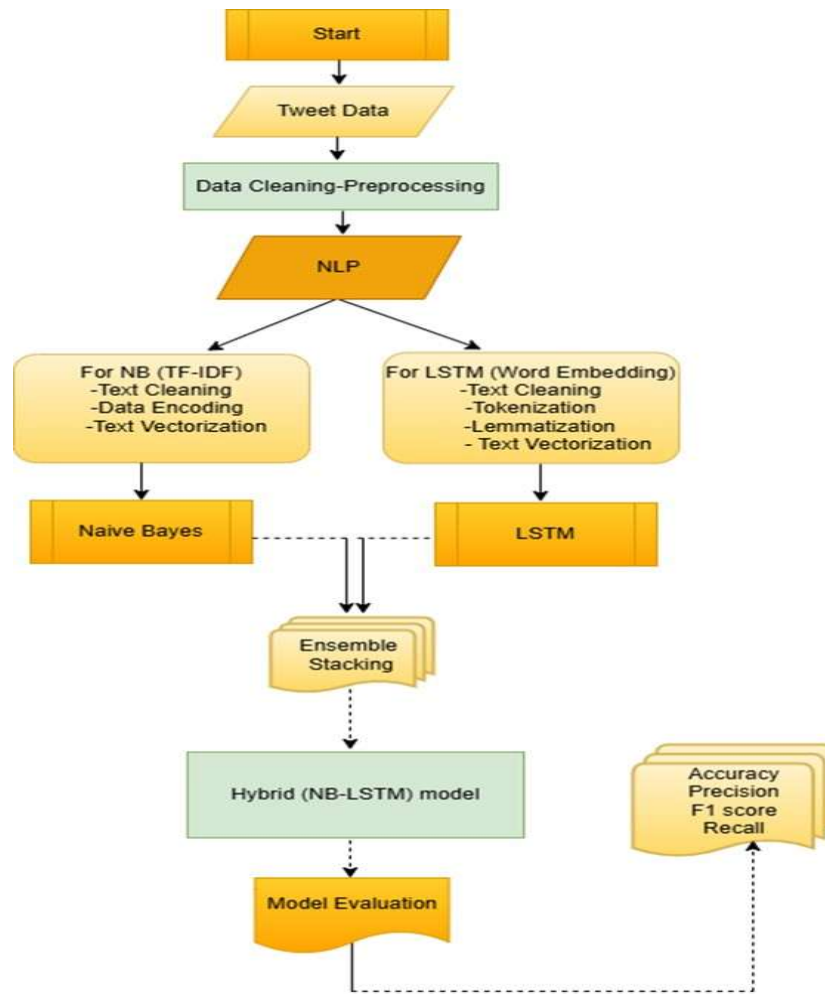


Figure 3: Architectural Design of The Proposed NB-LSTM Model With NLP

### 3.4. Performance Evaluation

After machine learning models are trained, it is necessary to assess how well they work. Researchers often evaluate NLP models' performance using standard metrics. In this study, we measured the accuracy, precision, recall, and F1 score of the developed models using training and test datasets. Classification is a common task in machine learning and NLP, and accuracy is its simplest metric. Accuracy score shows how often the model is right in its predictions. Precision means how well the model finds only the results that matter; recall is how well it finds all of the results that matter.

The F1 score quantifies how well a model predicts for imbalanced data when one outcome is much more common than the other. It is a single number that balances precision (how many predicted positives are true) and recall (how many true positives the model identified). Both are combined in a single metric called the F1 score: It's high when the precision and the recall are high, as it is here. In the other hand, recall is a model's ability to capture all the things that are actually positive. It is all about reducing false negatives by illustrating what the model did capture.

And high recall is important in tasks for which missing positive results are a problem. Together, those metrics provide a more complete picture of how well the model performs. The mathematical representation of these state-of-the-art metrics are shown below:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

#### 4. RESULTS

This paper compares Naïve Bayes (NB), Long Short-Term Memory (LSTM) and a Hybrid model that uses both LSTM and Naïve Bayes to detect hate speech and offensive language in social media conversations. The Naïve Bayes model, shown in Tables 1 and 2, had training accuracy of 91% and test accuracy of 90%. The results show that Naïve Bayes was relatively effective at classifying harmful content, though not as enduringly so as the LSTM and hybrid models. Figure 4 shows the bar chart of the training and test results, depicting how well the model performed in terms of the four metrics: accuracy, precision, recall and F1-score. Also, Figure 5 shows the error rate per iteration, which was relatively low; this is another confirmation of how well the Naïve Bayes model was efficient in spotting harmful content in the tweet data. In the other hand, the LSTM model had a training accuracy of 96.4% after 5 epochs, and a test accuracy of 95%, see Table 3 and Table 4 for details. This suggests the power of LSTM in handling sequence data and complex patterns of language. Figure 6 shows the results of the LSTM model as its training and testing metrics increased rapidly over the epochs. In addition, Figure 7 shows the error rate declining over the 5 epochs, indicating that the LSTM approach has been successful in reducing prediction errors.

**Table 1: Naïve Bayes model training results**

Iteration	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Correct Prediction	Incorrect Prediction
1	87.2	85.4	86.0	85.7	21,599	3,184
2	88.3	86.9	87.1	87.0	21,884	2,899
3	89.1	87.6	88.2	87.9	22,090	2,693
4	90.3	89.0	89.4	89.2	22,383	2,400
5	91.0	90.2	90.7	90.4	22,553	2,230

**Table 2: Naïve Bayes model test results**

Iteration	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Correct Prediction	Incorrect Prediction
1	85.6	84.2	84.8	84.5	4,238	712
2	86.7	85.3	85.7	85.5	4,291	659
3	88.1	86.8	87.3	87.0	4,367	583
4	89.4	88.3	88.9	88.6	4,427	523
5	90.0	89.2	89.6	89.4	4,455	495

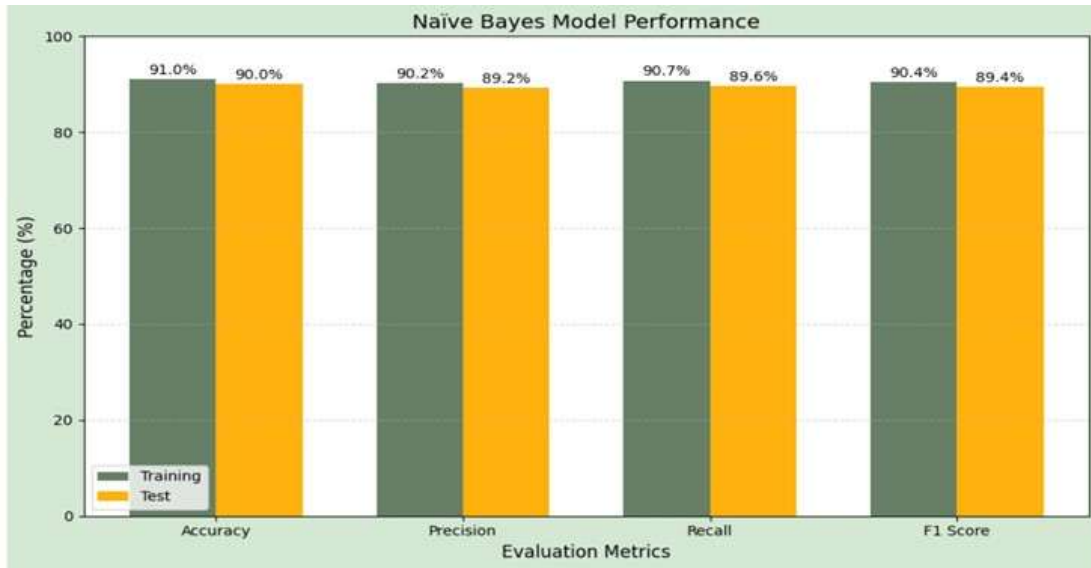


Figure 4: visualization of Naïve Bayes model trainin and test results for the four metrics.

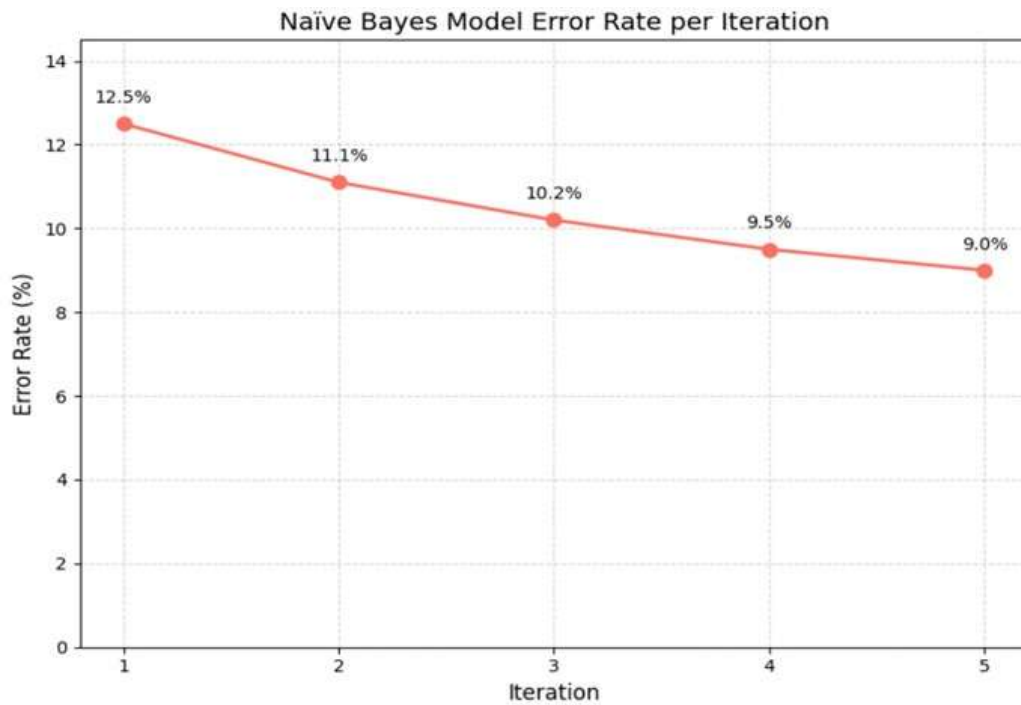


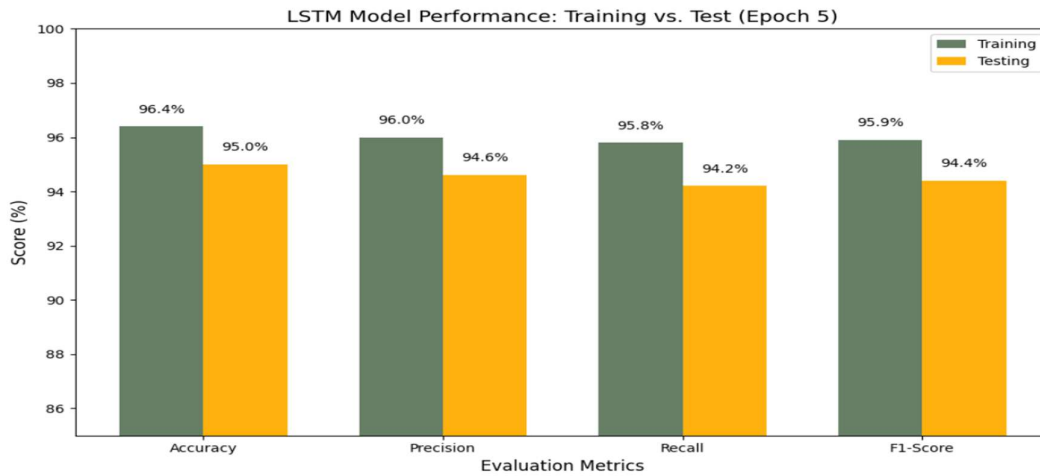
Figure 5: Naïve Bayes model error rate per iteration

**Table 3: LSTM model training results after 5 epoch**

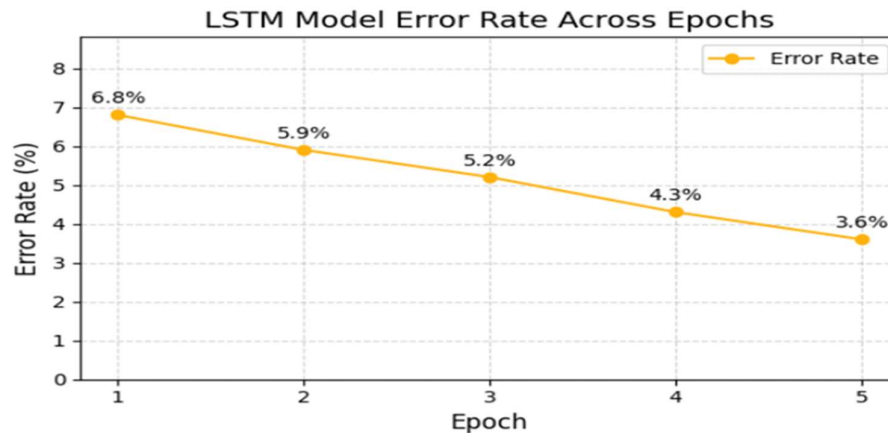
Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	90.5	89.8	88.7	89.2
2	92.7	91.5	90.9	91.2
3	94.1	93.6	93.0	93.3
4	95.3	94.7	94.2	94.4
5	96.4	96.0	95.8	95.9

**Table 4: LSTM model test results after 5 epoch**

Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	89.3	88.5	87.9	88.2
2	91.8	91.0	90.5	90.7
3	93.4	92.6	92.0	92.3
4	94.3	93.8	93.3	93.5
5	95.0	94.6	94.2	94.4



**Figure 6: Visualization Of LSTM Model Training And Test Results After 5 Epochs.**



**Figure 7: LSTM model error rate across five (5) epochs**

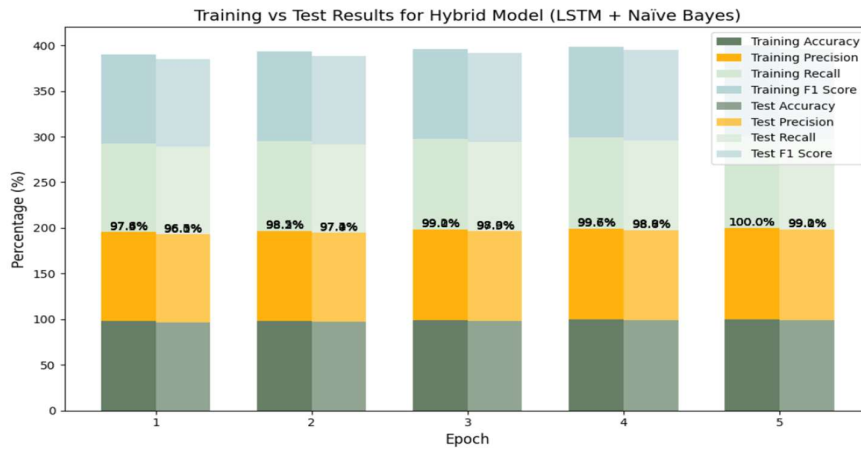
Likewise, the hybrid model, combining Naïve Bayes (NB) and LSTM, performs even better, with 100% training accuracy and a test accuracy of 99.2%; see Tables 5 and 6. This ensemble model outperformed individual models, as evident in the training and testing results in Figure 8. Figure 9 details the error rates of the Hybrid model over five epochs; they fluctuate modestly but decline consistently. The Hybrid model’s impressive performance suggests that combining the probabilistic ability of Naïve Bayes with the power of LSTM to learn sequences is a strong solution for identifying hate speech and offensive language on social media.

**Table 5: Hybrid (LSTM + Naïve Bayes) model training score**

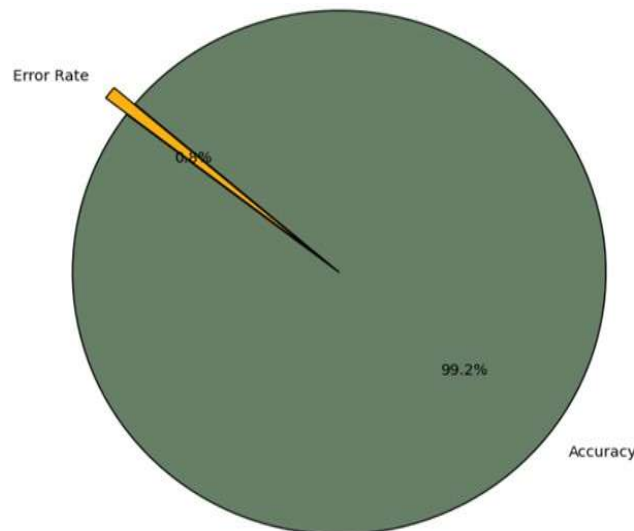
Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	97.8	97.6	97.4	97.5
2	98.5	98.3	98.2	98.2
3	99.2	99.1	99.0	99.0
4	99.7	99.6	99.6	99.6
5	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>

**Table 6: Hybrid (LSTM + Naïve Bayes) model test results**

Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	96.5	96.3	96.0	96.1
2	97.4	97.2	97.0	97.1
3	98.3	98.0	97.9	97.9
4	98.9	98.7	98.6	98.6
5	<b>99.2</b>	<b>99.1</b>	<b>99.0</b>	<b>99.0</b>



**Figure 8: Test Vs Training Results For The Hybrid Model (LSTM-Naïve Bayes)**



**Figure 8: Error rate Versus Accuracy**

The results of this experiment show the best performance of the combined algorithms in achieving the aim of this study. This work presents a better model with a better result. This work contributes to the field of machine learning and NLP by offering a fast and reliable solution for detecting hate speech and offensive messages on social media.

## 5. CONCLUSION & FUTURE WORK

In this study, the Naïve Bayes, LSTM and NB-LSTM model were developed for detecting hate speech and offensive language on social media. The findings of this paper shows that advances in machine learning and deep learning have the potential to vastly improve content moderation. For future work, a larger dataset with multiple social media platforms would likely improve performance, as perhaps would more sophisticated models like Transformers. More accurate methods will also be needed, such as developing real-time detection systems and adding datasets in many languages. Finally, more interpretable models will be important for encouraging transparency and fairness in automated content moderation.

## REFERENCES

- Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., & MacFeely, S. (2020). Big data and the UN Sustainable Development Goals (SDGs). *Sustainability*, *12*(14), 5397
- Pathak, A., Pandey, M., & Rautaray, S. (2019). Sentiment analysis of Twitter data using Machine Learning approaches and recurrent neural network (RNN) model. *Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 263–268). IEEE.
- Lazebnik, T., & Iny, O. (2024). Temporal graphs anomaly emergence detection: Benchmarking for social media interactions. *Applied Intelligence*, *54*, 12347–12356. <https://doi.org/10.1007/s10489-024-05821-3>
- Rajesh, A., & Hiwarkar, T. (2023). Sentiment analysis from textual data using multiple channels

- deep learning models. *Journal of Electrical Systems and Information Technology*, 10(1), Article 56. <https://doi.org/10.1186/s43067-023-00094-x>
- Islam, M. S., Kabir, M. N., Ghani, N. A., Ahmed, M. B., Hasan, M. K., & Ahmad, J. (2024). Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, 57(62). <https://doi.org/10.1007/s10462-023-10651-9>
- Ahmed, A., & Ahmed, B. (2023). Application of TF-IDF and machine learning classifiers for sentiment analysis on fake news articles. *Journal of Machine Learning Research*, 24, 45–56
- Gaur, V., Sharma, R., & Kapoor, P. (2023). Sentiment classification of Twitter data using Naïve Bayes and TF-IDF. *International Journal of Computer Applications*, 181, 1–7.
- Qi, L., & Shabrina, S. (2023). COVID-19 sentiment analysis in major cities of England using machine learning and lexicon-based approaches. *International Journal of Information Management*, 63, 102576. <https://doi.org/10.1016/j.ijinfomgt.2022.102576>
- Al Sari, R., Bukhari, H., & Alghamdi, M. (2022). Social media sentiment analysis for Saudi cruises using machine learning algorithms. *International Journal of Data Science and Analytics*, 11, 123–135.
- Mukherjee, S., Sharma, A., & Gupta, P. (2021). A customized approach for detecting explicit negation in sentiment analysis. *Expert Systems with Applications*, 172, 114–122. <https://doi.org/10.1016/j.eswa.2021.114122>.
- Noori, A. (2021). An innovative approach to customer sentiment classification using machine learning. *Journal of Information Technology Research*, 14, 34–50.
- Zahoor, A., & Rohilla, R. (2020). Comparative analysis of machine learning models for sentiment analysis of political tweets. *IEEE Access*, 8, 187501–187515. <https://doi.org/10.1109/ACCESS.2020.3029693>.
- Samuel, P., Singh, T., & Jain, R. (2020). Sentiment analysis of COVID-19 tweets using machine learning algorithms. *Journal of Intelligent Systems*, 29, 789–805
- Kumar, S., Patel, M., & Gupta, A. (2020). Gender and age-based sentiment analysis using machine learning techniques. *Journal of Computational Social Science*, 3, 89–105.
- Zarisfi, M., Ahmed, A., & Khan, R. (2020). Semantic scoring for sentiment analysis using SVM and multinomial Naïve Bayes. In *Proceedings of the 2020 International Conference on Big Data and Advanced Wireless Technologies* (pp. 7–12).
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47. <https://doi.org/10.1145/505282.505283>
- Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). *Tackling the poor assumptions of naive bayes text classifiers*. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 616–623).
- Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10), 2451–2471.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

- Chollet, F. (2018). *Deep learning with Python* (2nd ed.). Manning Publications.
- Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based Natural Language Processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75. <https://doi.org/10.1109/MCI.2018.2840738>
- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed. draft). Retrieved from <https://web.stanford.edu/~jurafsky/slp3/>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://www.deeplearningbook.org/>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. <https://aclanthology.org/N19-1423/>
- Kumar, A., & Jaiswal, A. (2021). Recent advancements in deep learning for social media data analytics: A review. *Computer Science Review*, 39, 100318. <https://doi.org/10.1016/j.cosrev.2020.100318>
- Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. <https://aclanthology.org/W17-1101/>
- Davidson, T., Warmusley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, 512–515. <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1–30. <https://doi.org/10.1145/3232676>
- Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. *European Semantic Web Conference*, 745–760. [https://doi.org/10.1007/978-3-319-93417-4\\_48](https://doi.org/10.1007/978-3-319-93417-4_48)
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. <https://aclanthology.org/N16-2013/>
- Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S. A., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language Online*, 80–93. <https://aclanthology.org/W19-3509/>