ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

A Framework for the Detection of Hate Speech in Unstructured Facebook Data using Sentiment and Emotion Analysis

Arya Sable[1], Manjiri Gulhane[2], Shivastha Rewaskar[3], Divesham Devdi[4], Priyanka Ingle[5], Dr. Rupali R. Deshmukh[6]

[1,2,3,4,5]*Student, CSE, HVPM College of Engineering and Technology, Amravati, India*

[6]*Assistant Professor, CSE, HVPM College of Engineering and Technology, Amravati, India*

Abstract: *One of the most popular communication and information sharing media has been the social media. Facebook is one of the platforms that produce a colossal user-formed textual information in the form of posts, comments, and messages. Though it is true that these platforms help in social interaction as well as exchange of knowledge, they also make possible the propagation of illicit content like hate speech, abusive language and offensive words. Social media hate speech has the potential to harm individuals, society, and social balance as a whole, thus its identification can be a relevant research issue. This study suggests a model of identifying hate speech using unstructured Facebook data on the basis of sentiment and emotion analysis and natural language processing methods. The suggested framework evaluates the user-generated content by utilizing a series of steps such as data collection, preprocessing, feature extraction, and classification. Sentiment analysis is applied to determine the polarity of a text, whereas emotion analysis identifies the emotional state of anger, fear, disgust, and sadness, which are typically related to hateful communication. These characteristics are coupled with machine-learning based text classification models that classify the social media text to either hate speech, offensive speech, or non-hate speech. The suggested system will enhance the precision of harmful communication detection through a combination of the emotional and sentiment-based features with the contextual language understanding. The framework will help social media sites to monitor the online interactions and detect abusive content automatically with a higher degree of effectiveness. Finally, this strategy will help create safer and more responsible digital spaces by aiding in the identification and detection of hate speech in massive social media information.*

Keywords: Hate Speech Detection, Social Media Analysis, Natural Language Processing (NLP), Sentiment Analysis, Emotion Analysis, Machine Learning, Facebook Data, Text Classification.

I. INTRODUCTION

Social media has been one of the most powerful means of communication and sharing of information in the recent years. Social media Like Facebook, enable individuals to exchange their ideas, views, and experiences by creating posts, comments, and messages. On the one hand, These Platforms Give the Chances to communicate and connect communities; however, on the other one, the platforms bring issues connected with the distribution of harmful materials. The Existence of Hate Speech is one of the most important problems within social media platforms.

 DOI: 10.5281/zenodo.19370658

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

Hate Speech A communication that is used to attack or attack a group of people with a hate speech directed against them or a group, on the basis of their character like race, religion, gender, nationality, or political views. Hate speech manifestation in Facebook is usually in the form of abusive comments, threatening, and aggressive postings. This kind of content may lead to poor environments and encourage intolerance among the users. It is challenging to identify hate speech since the information created on social media sites is not organized. Unstructured data contains the text that is not arranged according to a predetermined structure and can include slang, emojis, abbreviations, and informal words. Conventional data analysis techniques are non-practical when it comes to analysing such intricate textual data.

Advanced technologies that were deployed to handle this issue include Natural Language Processing (NLP), sentiment analysis, and emotion analysis. Sentiment analysis assists in determining the nature of a work of text, be it positive, negative, or neutral opinion. Emotion analysis also detects the particular emotions like anger, hatred, fear or sadness. With the combination of these methods, one can better identify harmful communication patterns.

The principal goal of the given research is to create a framework that will be able to analyse unstructured Facebook data, identifying hate speech with the help of sentiment and emotion analysis. The suggested framework is expected to enhance the methods of dangerous content detection and assist the social media outlets in ensuring the online environment remains safe and respectful.

II. LITERATURE REVIEW

A number of scholars have investigated the issue of hate speech identification on social media. The process of classifying harmful content has been a significant subject of study in data science and natural language processing with the growing application of social media.

Numerous works have been devoted to machine learning algorithms applied to hate speech detection in textual information. Methods used by researchers to make such classifications include Support Vector Machines (SVM), Naive Bayes, and Logistic Regression to identify hate and non-hateful posts in the social media. Such methods examine the textual characteristics including the keywords, word frequency, the linguistics patterns.

Hate speech detection has also recently been performed using deep learning techniques. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are models which have demonstrated good performance with complex textual data. These models can comprehend contextual data and discern trends within big data.

Another useful method of harmful content detection is sentiment analysis. It assists in the establishment of the positive or negative in a message. Sentiment analysis has been employed by many researchers to recognize negative or aggressive words in social media posts.

The analysis of emotion also improves the detection process as they detect different emotions, including those of anger, hatred, fear, and sadness. The fusion of sentiment analysis and emotion detection is more insightful to user behaviour and communication patterns.

Despite the contribution made by earlier researchers on hate speech detection, further refinements on the same are necessary to provide better frameworks capable of processing large amounts of

 DOI: 10.5281/zenodo.19370658

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

unstructured data. This study will seek to come up with a framework that incorporates both sentiment and emotion analysis methods to better specific the detection of hate speech on Facebook.

III. RELATED WORK

Among the initial papers by Del Vigna et al. (2017), there was one aimed at identifying hate speech on Facebook with machine learning. They used a method of extracting language characteristics using Facebook comments and use of supervised learning algorithms to determine hateful and non-hateful posts. The researchers have shown that textual statistical signals / n-grams and term frequency-inverse document frequency (TF-IDF) were effective in identifying explicit hate speech but were not effective with implicit/ sarcastic utterances.

Matamoros-Fernandez and Farkas (2021) used a systematic review of the literature on racism and hate speech in social media platforms. Their journal article outlined how hate speech on the internet has affected society and the necessity of automated ways of detecting hate speech that can handle vast amounts of unstructured social media information. It has also identified drawbacks of the conventional moderation methods in the study and suggested adoption of the artificial intelligence approaches to improve content control.

Sentiment analysis is another significant trend in hate speech detection. The sentiment analysis methods are used to categorize the textual information according to the polarity of emotions like positive, negative, or neutral. Nonetheless, it has been established that sentiment analysis cannot be used adequately to identify hate speech due to the fact that hateful information did not necessarily have high negative polarity. Consequently, a mixture of sentiment and other contextual characteristics is required to perform better.

New studies have been concentrated on emotion analysis, extending beyond the sentiment polarity to pinpoint certain emotional states including anger, fear, disgust, sadness and joy. Hate speech is commonly related to anger and disgust, which means that emotion detectors can offer more detailed information about the intent of the user. There is a combination of emotion-conscious models with machine learning classifiers and deep learning architectures to enhance detection accuracy.

Besides the conventional machine learning methods, recently transformer-based language models like BERT and XLM-R have been used to address hate speech detection problems. These models identify both contextual meaning and semantic connections among words, enabling them to deal with the complicated linguistic forms like sarcasm, slang, and multilingual phrases. Tuning of these models on data related to social media has demonstrated great advances in the classification tasks.

In addition, others have investigated the adoption of mixed structures that incorporate lexical, sentiment, emotion, and contextual embeddings. The goal of such frameworks along with the previous one is to achieve better detection accuracy and explainability since they emphasize the textual components that drive the classification.

In spite of these developments, there are still a number of challenges, such as multilingual material, code-switching, developing slang, and implicit hate speech. Thus, more detailed frameworks that combine sentiment analysis, emotion detection, and language context models are necessary to successfully identify hate speech in social media text that lacks any structure.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

IV. SYSTEM ARCHITECTURE

The suggested hate speech detection framework on Facebook is based on the multi-layered design that is established to handle unstructured data on social media and classify harmful information correctly. The architecture starts with data collection layer which gathers Facebook posts and comments through publicly available datasets like HASOC or some other social media corpus. These datasets consist of high quantities of unstructured text that comprises of slang, emojis, abbreviations, and multilingual information. After gathering, the data is sent to the data preprocessing layer, where noises and other unwanted components like URLs, special characters and stop words are eliminated. At this step, the text is normalized by tokenization, low-casing, and processing code switched language (e.g. Hindi-English) so that the data can be analysed further.

Once the cleaning is done, the text has been cleaned and then processed through the feature extraction layer which is also important in interpreting the semantic and emotional contents of the text. This layer identifies various kinds of features such as sentiment polarity (positive, negative, or neutral) and emotional reactions such as anger, fear, disgust, sadness or joy. Moreover, the contextual text embeddings are produced with the help of the high-end natural language processing models like the transformers-based construction. These embeddings are useful to capture the additional contextual meaning of words and phrases so that implicit or sarcastic versions of hate speech can be detected by system which would otherwise be overlooked by the traditional keyword-based systems. The resulting features are then fed into the classification layer where machine learning or deep learning algorithms classify the material into one of the previously defined categories, e.g. hate speech, offensive speech, or non-hate speech. The classification model combines emotion and sentiment signals and contextual representations to enhance the accuracy of predictions. After classification the findings are sent to the integration and visualization layer where the structured findings are stored and displayed in dashboards or analytical interfaces. This layer shows the identified hate speech, related sentiment and emotional patterns, which can be interpreted by the moderators or researchers. Lastly, the system performance is checked with the help of conventional evaluation metrics like accuracy, precision, recall, and F1-score to guarantee the reliability and the strength of the framework. The architecture allows the effective representation, detection and analysis of hate speech in large-scale unstructured data of Facebook and allows explainable and scalable content moderation systems.

V. TECHNOLOGIES USED

The suggested model of hate speech detection on Facebook based on sentiment and emotion analysis presupposes a set of programming implementations, machine learning packages, natural language processing systems, and visualization systems. The technologies are used to gather data, process unstructured text, train models, and demonstrate findings.

To begin with, the system is programmed with Python, one of the most popular programming languages to use when dealing with artificial intelligence and natural language processing tasks. Python includes a vast suite of libraries that support both text processing and machine learning as well as deep learning.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

Libraries like NumPy and Pandas are used in the data preprocessing and manipulation. These libraries contribute to the processing of datasets, cleaning of raw data, elimination of undesired characters, handling of missing values, and the structuring of input to be fed to machine learning models. Moreover, spaCy and NLTK (Natural Language Toolkit) are also applicable in the processing and techniques of natural language like tokenization, stop-word removal, stemming, and lemmatization.

Sentiment analysis may be conducted with the help of such tools as VADER Sentiment Analyzer or TextBlob. These tools examine the polarity of the text and label it as positive, negative or neutral sentiment. To detect emotions, emotion lexicon or pre-model can be applied to recognize emotional states in the text, including anger, sadness, fear, joy, and disgust.

Scikit-learn, TensorFlow and PyTorch are machine learning and deep learning frameworks that are used to build the hate speech detection model. Such frameworks can be used with algorithms such as Logistic Regression, Random Forest and more sophisticated neural network. Beyond this, models that rely on transformers like BERT (Bidirectional Encoder Representations from Transformers) can also be applied to contextual text embeddings and this is much better at making the model interpret semantic meaning and context.

In order to visualize data and present results, Matplotlib, Seaborn, and Plotly libraries are required to build charts, graphs, and dashboards, which provide sentiment distribution, emotion patterns, and classification results.

Such visualizations assist the moderators and researchers to comprehend the outcomes of the results. Jupyter Notebook, Google Colab or Visual Studio Code can be used as the development and experimentation environment, which enables machine learning models to be easily tested and trained. To store and manage the data sets, one can use CSV files or databases like MySQL or MongoDB depending on the dataset size.

VI. METHODOLOGY

The research proposed follows a systematic approach in identifying hate speech on Facebook posts by combining sentiment analysis and emotional recognition and using highly complex natural language processing strategies. The general process layout is aimed to handle unstructured social media data, then mine meaningful linguistic characteristics and categorize text information in pre-determined categories. The model has five broad steps, which include data collection, data pre-processing, feature extraction, classification, and evaluation.

Data Collection:

The initial action of the methodology is gathering textual information on publicly provided datasets of social media. As the access to the private Facebook data is limited by the privacy policies, benchmark datasets concerning the hate speech detection, including the HASOC (Hate Speech and Offensive Content) dataset and other publicly available social media corpora are applied. These datasets include labelled samples of user-written posts that are considered hate speech, offensive language, or non-hate. The most important input used in training and testing the proposed model is the collected data.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

Data Preprocessing:

The social media data are usually not structured and include all types of noises (URLs, emojis, hashtags, repeated characters, and special symbols). Consequently, the data needs to be cleaned and normalised by a preprocessing phase prior to the implementation of machine learning methods. During this phase, some text normalization procedures are performed such as the deactivation of URLs and special characters, the transformation of text to small letters, the deletion of stop words, and sentencing and dividing the sentences into separate words or tokens. Other preprocessing activities are code-switched language (e.g. Hindi English mixture), slang and abbreviation normalization, and the elimination of non-relevant textual elements. This phase is necessary to make the dataset fit the additional analysis and model training.

Feature Extraction:

After the preprocessing of the text, an extraction of meaningful features of the data comes in the next step. The process of feature extraction is significant towards the interpretation of the text semantics and emotion. There are three types of features that are taken into account in this research.

Originally, sentiment analysis is implemented to identify the polarity of every post. Sentiment analysis assigns the text to one of the three categories (positive, negative, or neutral), which give an idea of the overall attitude that is conveyed in the message.

Second, the analysis of emotions is conducted to determine certain emotional conditions in the text. Anger, disgust, fear, sadness, joy, and surprise are some of the emotions that are identified with the help of emotion lexicons or pre-trained models. As hate speech can be linked to numerous negative feelings like anger and disgust, emotion-detection can give a more insight to identify the intention of the user.

Third, transformer-based natural language processing models are used to create contextual text embeddings. These embeddings learn the semantic relation between words and assist the system to comprehend the contextual meaning of words, sarcasm and language variation. This sentiment coupled with emotion and the contextual features make a more detailed representation of the text to classify.

Classification Model:

Once the feature extraction is completed, the processed data is then used to train a classification model that classifies societal media posts into a set of predestined classes that include hate speech, offensive speech or non-hate speech. It is done with the help of machine learning and deep learning algorithms. The classification model trains on the training data and determines linguistic features of a hate speech. Considering the contextual text representations as well as the sentiment and emotion indicators, the classifier can enhance detection rates and eliminate false alarms.

Result Integration and Visualization:

The results of the classification model are incorporated in a system of interpretation and analysis. The system indicates the identified content of the hate speech as well as the sentiment and emotion

 DOI: 10.5281/zenodo.19370658

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

indicators. The distribution of the types of hate speech, sentiment polarity, and emotional patterns in the data are presented with the help of visualization tools in forms of charts and graphs. This step enhances the decipherability of the system and helps the moderators or researchers to grasp the online content phenomenon.

Performance Evaluation:

The last phase of the methodology is the review of the performance of the suggested framework. Measures of effectiveness of the hate speech detection model are standard evaluation measures like accuracy, precision, recall, and F1-score. These measures are used to evaluate how the model can rightly categorize hateful content and reduce the false positives and false negatives. The evaluation findings can be used to understand how reliable and robust the proposed system would be in the context of social media data in the field.

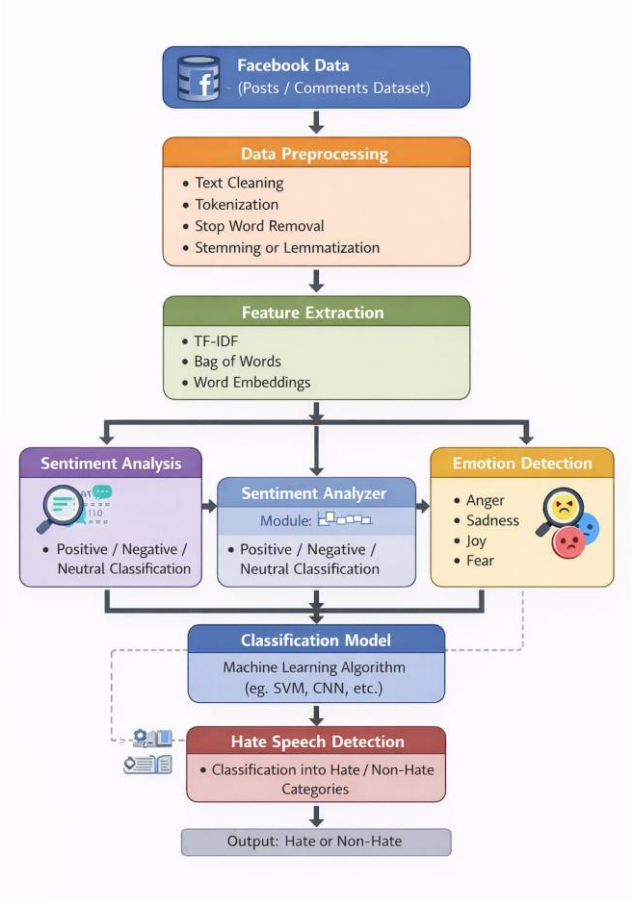

FIG 1: SYSTEM ARCHITECTURE FOR HATE SPEECH DETECTION FROM FACEBOOK DATA

VII. CONCLUSION

The study offers one of the frameworks of detecting hate speech in Facebook posts with the help of sentiment and emotion analysis with the help of the natural language processing method. As the social media sites continue to grow fast, the proliferation of bad and objectionable content has become a

 DOI: 10.5281/zenodo.19370658

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

big issue to the online communities. This is because such environments are hard to detect the hate speech because the social media data is unstructured and frequently encompass informal language, slang, emojis, and multiple languages.

The suggested framework will tackle these issues by combining sentiment analysis and emotion recognition with text classification tools that are based on machine learning. Facebook data is processed in the system using three steps namely data collection, preprocessing, feature extraction, classification and performance evaluation. Sentiment analysis is used to identify the polarity of user-generated content whereas emotion analysis is used to identify emotional states of anger, fear, disgust, and sadness that are frequently relevant to hateful phrases. Using these features together with contextual text representations, the suggested method offers better accuracy and precision of hate speech recognition.

The classification model presents the categorization of social media content based on hate speech, offensive speech and non-hate speech, which allows to monitor harmfully communicative content in social media and monitor it more efficiently. The analysis of the framework with the conventional performance metrics like accuracy, precision, recall, and F1-score proves the efficiency of the proposed framework in detecting hate speech among unstructured textual data.

On the whole, the suggested framework will aid in the creation of automated mechanisms that will facilitate safer and more responsible online communication. This framework can be expanded into multilingual datasets, multimodal data (images and videos) and using more sophisticated deep learning models in future work to improve the performance and scalability of detection in the real-world social media setting.

REFERENCES

[1] A. Matamoros-Fernandez and J. Farkas, "Racism, Hate Speech, and Social Media: A Systematic Review and Critique," Television and New Media, vol. 22, no. 2, p. 205-224, 2021.

[2] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, Hate Me, Hate Me Not: Hate Speech Detection on Facebook in Proceedings of the First Italian Conference on Cybersecurity (ITASEC), Venice, Italy, 2017, pp. 86-95.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, in Proceedings of the 11 th International AAAI Conference on Web and Social Media (ICWSM), 2017.

[4] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter with a Convolution-GRU based Deep Neural Network," in Proceedings of the European Semantic Web Conference, 2018.

[5] B. Liu, Sentiment Analysis and opinion mining. San Rafael, Californian, USA: Morgan and Claypool Publishers, 2012.

[6] C. Hutto and E. Gilbert, VADER: A Parsimonious Rule-Based Model of Sentiment Analysis of Text in Social Media, in Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM), 2014.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training Deep Bidirectional Transformers, Language Understanding," in the 2019 Conference of the North American Chapter of the Association of Computational Linguistics, 2019.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 4 | Issue 6 | December 2025

www.ijirid.in

[8] T. Pota, S. Garg, and M. Shakshuki, A Survey of Hate Speech Detection Using Natural Language Processing, Journal of Information Processing Systems, vol. 15, no. 4, pp. 885-901, 2019.

[9] M. Mozafari, R. Farahbakhsh, and N. Crespi, Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model, vol. 15, no. 8, 2020, PLOS ONE.

[10] HASOC, Hate Speech and Offensive Content Identification in Indo-European Languages, 2019, Forum of Information Retrieval Evaluation.

