Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

A FRAMEWORK FOR HATE SPEECH IDENTIFICATION USING OPTIMIZED TEXT FEATURES AND NATURAL LANGUAGE PROCESSING ON TWITTER DATASET

Irsa Manzoor[1], Muhammad Sajid Maqbool[*2], Faisal Shahzad[3], Muqadas Nadeem[4], Amna Zulfiqar[5], Syeda Qanitah Naqvi[6]

[1]*Department of Information Technology, Bahauddin Zakariya University, Multan*
[*2]*Department of Computer Science, NFC Institute of Engineering and Technology, Multan*
[3]*Department of Computer Science, Bahauddin Zakariya University, Multan*
[4]*Department of Computer Science, Emerson University, Multan, Pakistan*
[5]*Department of Computer Science, NFC Institute of Engineering and Technology, Multan*
[6]*Department of Software Engineering NUML, Multan*

[1]Irsamanzoor6@gmail.com, [*2]sajid.maqbool@nfciet.edu.pk, [3]faisal.shahzad.research@gmail.com, [4]nmuqadas587@gmail.com, [5]Amna.zufiqar@nfciet.edu.pk, [6]qanitah.naqvi@numl.edu.pk

[*2]ORCID: 0000-0001-5583-3550

DOI: https://doi.org/10.5281/zenodo.20677715

Keywords
Hate Speech Recognition, Sentiment Analysis, Tweets Prediction, Machine Learning

Article History
Received: 15 April 2026
Accepted: 27 May 2026
Published: 13 June 2026

Corresponding Author: *
Muhammad Sajid Maqbool

Abstract

Twitter has emerged as a prominent social media platform where users rapidly share opinions, emotions, experiences, and real-time events. Due to the increasing volume of user-generated textual content, sentiment analysis and hate speech detection have become important research areas in the fields of Natural Language Processing (NLP) and Machine Learning (ML). Although considerable research has been conducted on hate speech detection using Twitter data, the automatic identification of multilingual hate speech, particularly in Roman Urdu and English, remains a challenging task. This research proposes a hybrid NLP-based framework for multilingual sentiment analysis using a combined dataset of Roman Urdu and English tweets collected from publicly available hate speech datasets. The datasets are integrated into a unified corpus and processed using several NLP preprocessing techniques, including stop-word removal, punctuation removal, URL elimination, tokenization, and stemming. Furthermore, optimized textual features are extracted using Python-based NLP libraries to improve the quality of the dataset for machine learning applications. To enhance feature relevance and reduce dimensionality, Principal Component Analysis (PCA) is applied to eliminate less informative features while retaining the most significant attributes. The experimental implementation is carried out using Google Colab, where multiple machine learning classifiers, including Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT), are trained and evaluated. In addition, a Hybrid Ensemble Model (HEM) is proposed, which combines the predictions of all four classifiers to improve classification performance. The proposed system classifies users' sentiments into

three categories: Positive, Negative, and Neutral. The performance of the models is evaluated using standard evaluation metrics, including training accuracy, testing accuracy, precision, recall, and F1-score. A comparative analysis of all models is conducted to identify the most effective approach for multilingual sentiment analysis and hate speech detection on Roman Urdu and English Twitter datasets.

1 INTRODUCTION

In the digital age, the widespread use of social media platforms and online communication channels has given rise to a concerning phenomenon: hate speech. Any type of communication—verbal or written—that encourages violence, prejudice, or animosity towards one person or group of people because of that person's race, religion, ethnicity, gender, sexual orientation, or other traits is referred to as hate speech [1]. The proliferation of hate speech online poses significant challenges to maintaining inclusive and respectful digital spaces. Recognizing the urgency and importance of addressing hate speech, researchers and technologists have turned their attention to developing automated systems for hate speech recognition. These systems employ advanced machine learning algorithms and natural language processing techniques to detect and classify instances of hate speech in online content [2]. By automating the identification process, these systems aim to alleviate the burden on human moderators and enable timely interventions to curb the spread of hate speech. Research focuses on the development of a hate speech recognition system that combines state-of-the-art machine learning models with a comprehensive dataset specifically curated for hate speech detection. The ultimate goal is to design an accurate and efficient system capable of identifying hate speech instances with a high degree of precision and recall [3].

Twitter has become a quintessential platform for users to share their thoughts, feelings, and activities in real-time. It serves as a powerful broadcast medium that allows individuals to swiftly express where they are, what they are engaged in, their current thoughts, and their immediate emotions. In recent times, there has been a growing focus on the analysis of sentiment in Twitter data, particularly in the context of identifying and combatting hate speech. In ongoing research endeavors, innovative methods have been developed for sentiment analysis, primarily in the context of a hate-speech dataset. The objective is to harness the capabilities of Natural Language Processing (NLP) and Machine Learning (ML) techniques to better understand and combat the pervasive issue of hate speech on this popular social media platform. Despite considerable progress, there is still ample room for improvement in the automatic detection of hate speech on Twitter. The dynamic nature of language and the evolving forms of online abuse present ongoing challenges. Researchers are continually exploring new avenues for refining NLP and ML algorithms to enhance the accuracy and efficiency of hate speech detection. This is crucial not only for maintaining a safe and inclusive online environment but also for understanding the changing landscape of digital communication [4]. Moreover, as Twitter remains a vital source of information and communication for millions of users worldwide, the significance of robust and adaptable hate speech detection methodologies cannot be overstated. The development of more sophisticated models and strategies for recognizing hate speech will contribute to a safer and more respectful online discourse, thereby ensuring that Twitter fulfills its potential as a platform for meaningful expression and communication. As the digital landscape continues to evolve, the pursuit of effective hate speech detection methods remains a prominent research area, warranting continuous investigation and innovation [5].

A promising avenue for research lies in the amalgamation of two distinct datasets featuring hate-speech tweets in two different languages, namely Roman Urdu and English. Combining these datasets into a single CSV file presents a unique opportunity to examine the cross-linguistic

dimensions of online hate speech. To facilitate this study, advanced Natural Language Processing (NLP) techniques can be employed, leveraging the power of Python as a primary programming language. The initial phase of the research focuses on extracting and optimizing features from the hate-speech tweets, which will result in the creation of a refined dataset [6]. This dataset will subsequently serve as the basis for training and testing machine learning models. A key component of the methodology involves the application of Principal Component Analysis (PCA) to filter out less important features, ensuring that only the most informative ones are retained. By doing so, the research aims to enhance the efficiency and effectiveness of the machine learning models used for hate speech detection. In pursuit of this research, Google Collab, a powerful cloud-based tool for data analysis and machine learning, will be harnessed to construct and assess a variety of machine learning models. This approach not only streamlines the development process but also provides access to substantial computational resources, fostering the creation of robust and accurate models. By combining datasets from different languages, this research endeavors to shed light on the universality of hate speech characteristics and the effectiveness of NLP techniques in capturing them across linguistic barriers. Furthermore, the optimization of features through PCA and the application of various machine learning models will contribute to the ongoing efforts to combat online hate speech, ensuring a safer and more inclusive digital environment for users across diverse linguistic backgrounds [7].

2 Literature Review

The current study's goal is to build on earlier research that has demonstrated the existence of both favorable and unfavorable views towards the tweet's variant. Twitter sentiment research allows organizations a rapid and effective way to track public opinion about their brand, operation, executives, etc., according to [6,9,11] research efforts. In the proposed study, the works of different researchers are explained, that introduces novel technique for incorporating semantics as extra features in the training set for sentiment analysis. They add the semantic concept (such as "Apple product") for each retrieved entity (such as the "iPhone") from tweets as an extra feature and calculate the correlation between the represented concept and the positive/negative sentiment.

A study on the sentiments of tweets is conducted by [1]. In this study, they explain that Numerous people around the world have been impacted by the COVID-19 epidemic since 2019. It is starting to take on the characteristics of an infectious disease that sparked a catastrophe with far-reaching implications on areas like health, economics, and education. New COVID-19 mutations, including the Beta, Delta, and Omicron forms, appeared during the coronavirus outbreak, frightening and alarming the populace. According to World Meter, differences in COVID-19 caused around 6 million deaths. On November 24, 2021, the SARS-CoV-2 omicron strain was discovered for the first time in South Africa, and it has since spread to more than 57 countries. They investigate attitudes and behaviors towards the omicron variation. Put up a method for determining sentiment analysis for tweets from Twitter on Omicron. There is a lot of potential in the sentiment analysis of Twitter data. The methodology uses Python's NLP tools to extract the best qualities from the Omicron tweets, creating a dataset that can be utilized for learning the Models. Four machine learning (ML) classifiers, including "Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM)", used the generated dataset to accurately classify users' emotional behavior into three categories: neutral, negative, and positive. Based on the accuracy of the forecast level, the Class Neutral earns the highest score, and the Class Negative receives the lowest score. When all of the characteristics are used in the model's training, the SVM and RF classifiers perform better. SVM classifier accuracy is 89.8%, whereas RF classifier accuracy is 82%. The Class Neutral receives the best score, and the Class Negative receives the lowest score based on the accuracy of the forecast level. An innovative

method for the sentiment analysis of Twitter data was presented by [15]. The most focused study terminology used for sentiment analysis to extract this unknowable information from the linguistic data is NLP and input mining methodologies. A DL technique is introduced to public sentiment analysis by [4]. He states in his study piece that the microblog has become a popular platform for discussing current events in China. The volume of linked posts on microblogs generally jumps all at once when a coronavirus outbreak develops, providing an excellent opportunity to ascertain how the general population feels about the issue. In [7], during the Omicron, the SentiStrength software is used to offer a new Twitter sentiment analysis method. According to them, a new tweet variant known as "SARS-CoV-2 Omicron" emerged, even though many COVID-19 variants have a significant detrimental influence on the lives of millions of people worldwide. The public's perception of the propagation of the SARS-CoV-2 Omicron strain on Twitter is investigated in this study. The applied approach is based on text analytics of Twitter data, considering the main subjects of tweets, retweets, and hashtags, the pandemic's containment, the efficacy of tweet immunizations, transmissible variants, and the rise in infection.

In [12], a new study on social media text behavior analysis is introduced using machine learning models. In this study, they explain that social media use and the dissemination of knowledge have helped humanity. But this has also resulted in several issues, such as the growth of hate speech. Recent studies have used a number of feature engineering techniques and machine learning algorithms to tackle this new problem of hate speech on social media. It is unknown if a comparison of different methods for developing features and machine learning algorithms has been undertaken to discover which one performs better on a common dataset that is made available to the public. The test results showed that the support vector machine technique with the bigram feature set produced an overall accuracy of 79 percent. The results showed that the negative sentiment significantly increased on January 20 when the lockdown in Wuhan started. Mohammad Mahyoob et al. put forward a novel Twitter sentiment analysis technique during the Omicron outbreak by utilizing SentiStrength software. They showed that while the spread of the SARS-CoV-2 Omicron strain had a significant negative impact on the lives of millions of people worldwide, the method studied how the public felt about it. This method relied on text analytics of Twitter data, focusing on primary topics like tweets, retweets, and hashtags, the restriction of the pandemic, the efficacy of COVID-19 vaccinations, transmissible variations, and the increase in infection. The SentiStrength software was used to analyze 18,737 tweets from December 3, 2021, to December 26, 2021, using a set of linguistic criteria and a vocabulary of sentiment phrases with an average of 95% confidence intervals [29]. The analysis revealed that the strength of the negative sentiment was higher than that of the positive sentiment, with weak sentiment strength at 31.01%, moderate sentiment strength at 16.32%, strong sentiment strength at 5.36%, and very strong sentiment strength at 0.35%. On the other hand, the strength of the positive sentiment declined, with weak sentiment strength at 16.48%, moderate sentiment strength at 11.19%, strong sentiment strength at 0.80%, and very strong sentiment strength at 0.02%.

Table 1 Literature Review

Literature	Used Dataset	Algorithms	Software and Tools	Accuracy
Hassan Saif et al., 2017	3 Twitter datasets	SVM & NB	Python	75%
Unaiza Fazal et al., 2023	Omicron Tweets	SVM, NB & RF	Python	89%
Patel Ravikumar et al., 2020	Twitter dataset for the 2014 FIFA World Cup	SVM & KNN	Weka, Python	85%
Jintao Ling et al., 2020	1 million blog postings	BERT	Python	75.13%
Mohammad Mahyoob et al., 2022	18,737 tweets from Twitter	Software based	SentiStrength	71%
Lokesh Mandloi and Ruchi Patel, 2020	Twitter dataset	NB, SVM,	Python	86%
Doaa Mohey El-Din Mohamed Hussein, 2016	Online papers	KNN, NB, RF	Python	83%
Deepika Vatsa and Ashima Yadav, 2022	Two-month data set of tweets	N/A	Python	84%
Sanjeev Verma, 2022	Public services dataset	NB	HCI	86%
Oksana Tokarchuk et al., 2022	TripAdvisor online reviews.	N/A	SPSS	87%

3 Material and Methods

The proposed methodology consists of several phases designed to perform multilingual sentiment analysis on Roman Urdu and English Twitter data. In the first phase, an appropriate dataset containing tweets in both Roman Urdu and English were selected from the widely recognized Kaggle repository. The second phase involved preprocessing the collected data to extract meaningful textual features using various Natural Language Processing (NLP) techniques, including stop-word removal, punctuation removal, URL elimination, and stemming. In the third phase, the processed dataset was further refined through labeling, annotation, and cleaning procedures. This included transforming categorical attributes into numerical representations and removing null, empty, and noisy instances to improve data quality. The fourth phase focused on feature engineering, where two additional features, namely polarity and subjectivity, were extracted using the TextBlob NLP library in Python. In the fifth phase, multiple machine learning models were implemented for sentiment classification, including Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and a Hybrid Ensemble Model (HEM). The proposed HEM integrates the predictions of all four individual classifiers to generate a final prediction outcome. Finally, in the evaluation phase, the performance of all selected models was assessed using standard evaluation metrics such as training accuracy, testing accuracy, precision, recall, and F1-score. The comparative analysis of these results was then conducted to identify the most effective model for multilingual sentiment analysis on the given dataset.

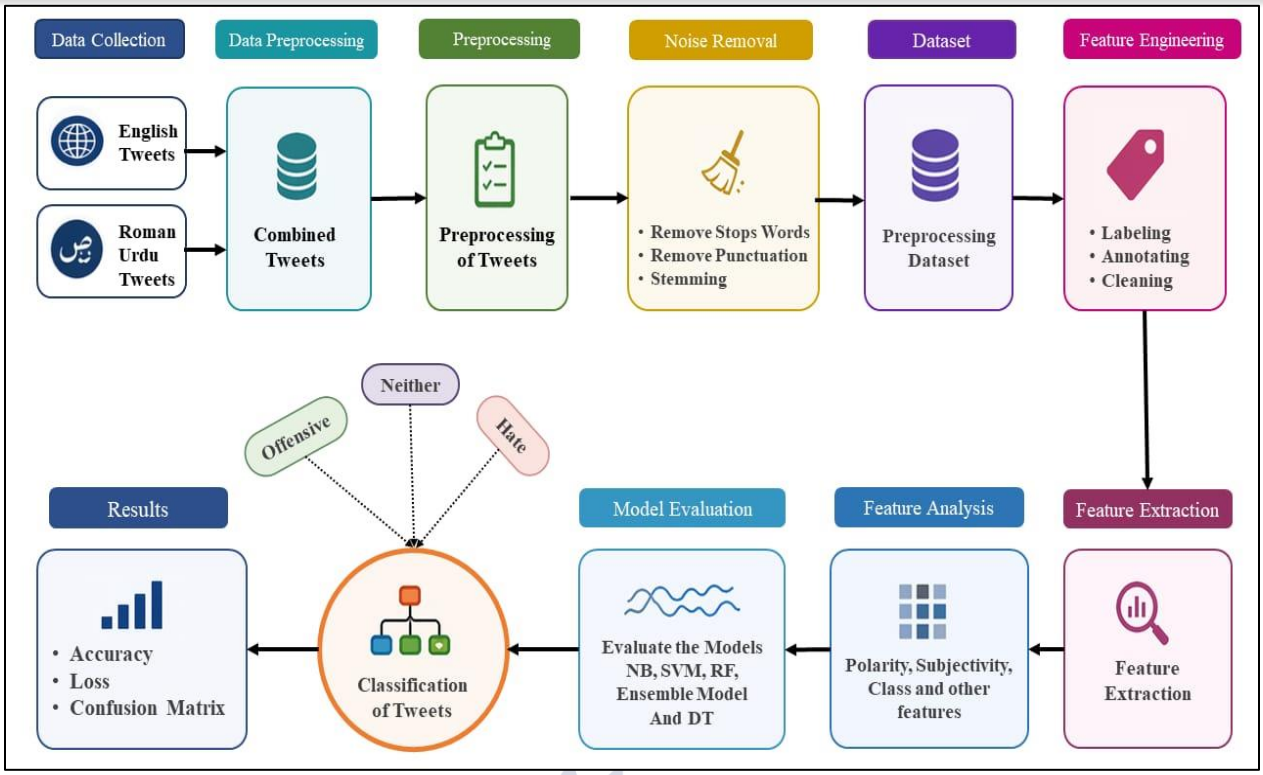

Figure 1 Proposed Methodology

The working of Hybrid Ensemble Models is explained in Figure 3. In Figure 3, X is the dataset that is used by every model, and (Y1, Y2, Y2 and Y4) are the predictions of simple models (NB, DT, RF, and SVM). The predictions (Y1, Y2, Y2 and Y4) of base models are then combined by using a voting classifier to give the final prediction. Y in Figure 3 is the final prediction of the Hybrid Ensemble Model (HEM).

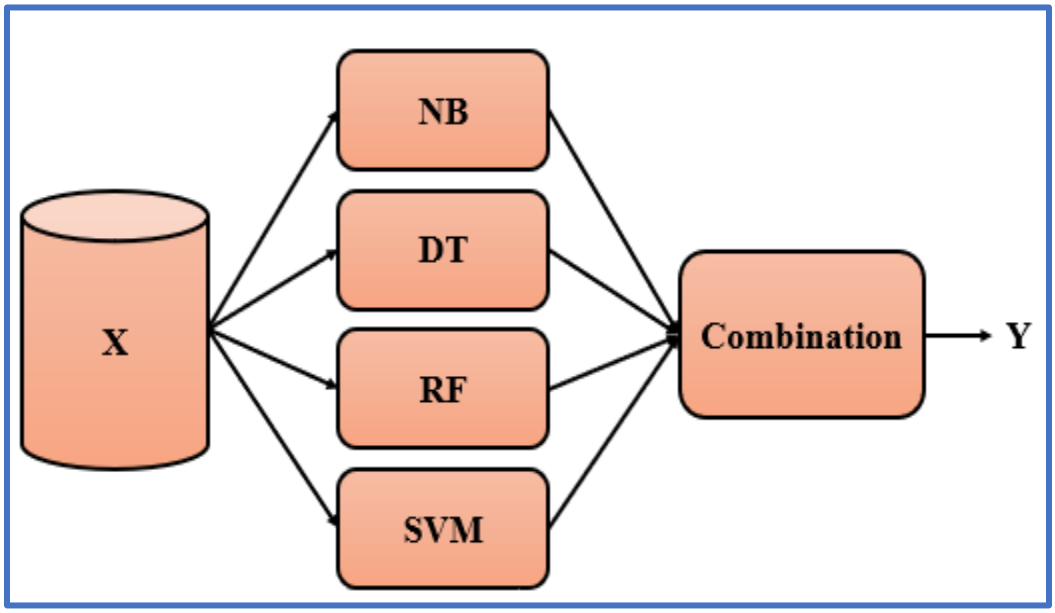

Figure 2 Proposed Model

3.1 Used Dataset

The dataset that is used for the experiment of our suggested model contains 31962 rows and 3 columns. Each row contains the ID of the tweet, the label of the tweet, and the text of the tweet. We will use 30 percent (9589 rows) of the data for the testing of our selected machine models and the remaining 70 percent (22373 rows) for the training of models.

3.2 Preprocessing of Tweets

Different preprocessing steps are applied to the combined dataset to create meaningful features from the tweets. Preprocessing Steps involve Natural Language Processing (NLP) techniques that process the data before extracting features. We use five NLP-based text preprocessing techniques by using the Natural Language Tool Kit (NLTK) library of the Python language. Text preprocessing techniques that we implement on combined tweets are:

- Creating Tokens of Tweets
- English Stopwords Removing Technique
- Roman Urdu Stopwords Removing Technique
- Punctuation Removing Technique
- Joining of Tokens

All of the mentioned techniques are applied by using an open-source GPU from the Google Collab platform.

3.2.1 Creating Tokens of Tweets

The first technique applied to the tweet's dataset is creating tokens of tweets for processing further tasks. Tokenization in natural language processing (NLP) is the process of breaking down a text into smaller units, typically words or subwords, which are referred to as tokens. These tokens are the basic building blocks for various NLP tasks, such as text analysis, language modeling, and machine learning. The main goal of tokenization is to split a text into meaningful units so that a computer can understand, process, and analyze the text effectively.

3.2.2 English Stopwords Removing Technique

Removal of English stopwords from the tweets is the next step after the tokenization of the text. Stopwords removal is a common preprocessing step in natural language processing (NLP) that involves eliminating common words that are considered to be of little value in text analysis. Stopwords are words that occur frequently in a language but typically do not carry significant meaning on their own. Examples of stopwords in English include words like "the," "and," "in," "of," "is," "a," and "an."

3.2.3 Roman Urdu Stopwords Removing Technique

The third NLP applied technique is removing stopwords of Roman Urdu. We defined a Roman Urdu stopwords list to remove them because the NLTK library does not support Urdu stopwords. The following stopwords list is defined from Roman Urdu words.

"ai", "ayi", "hy", "hai", "main", "ki", "tha", "koi", "ko", "sy", "woh", "bhi", "aur", "wo", "yeh", "rha", "hota", "ho", "ga", "ka", "le", "lye", "kr", "kar", "lye", "liye", "hotay", "waisay", "gya", "gaya", "kch", "ab", "thy", "thay", "houn", "hain", "han", "to", "is", "hi", "jo", "kya", "thi", "se", "pe", "phr", "wala", "waisay", "us", "na", "ny", "hun", "rha", "raha", "ja", "rahay", "abi", "uski", "ne", "haan", "acha", "nai", "sent", "photo", "you", "kafi", "gai", "rhy", "kuch", "jata", "aye", "ya", "dono", "hoa", "aese", "de", "wohi", "jati", "jb", "krta", "lg", "rahi", "hui", "karna", "krna", "gi", "hova", "yehi", "jana", "jye", "chal", "mil", "tu", "hum", "par", "hay", "kis", "sb", "gy", "dain", "krny", "tou" , "h" , "hai," , "he" , "k" , "ha" , "thy" , "thay" , "tum" , "tm" "kam", "tu", "wo", "hum", "aap", "ho", "hai", "thay", "kar", "karna", "karne", "kiya", "lekin", "aur", "yeh", "woh", "jab", "kyun", "kaise", "jaisa", "agar", "kuch", "bahut", "zyada", "kam", "yahan", "wahan", "ab", "kab", "kahan" , "hai", "se", "ka", "ke", "ko", "ki", "mein", "mein", "bhi", "par", "or", "ho", "aur", "kya", "hai", "nahi", "hota", "tha", "tha", "jab", "woh", "yeh", "uska", "uski", "kuch", "ab", "tak", "baat", "kab", "lekin", "aap", "mere", "mujhe", "agar", "tum", "hum", "jaise", "koi", "jao", "jana", "mil", "gaya", "ja", "raha", "kaam", "sab", "kuchh", "acha", "thodi", "zyada", "usne", "kisi", "din", "raat", "shayad", "ho", "waise", "jahan", "aaj", "kabhi", "kuch", "iss", "uss", "ek", "do", "teen", "char", "panch", "chhe", "saat"

Figure 3 Roman Urdu Stopwords

The figure contains a list of self-created stopwords of the Roman Urdu language.

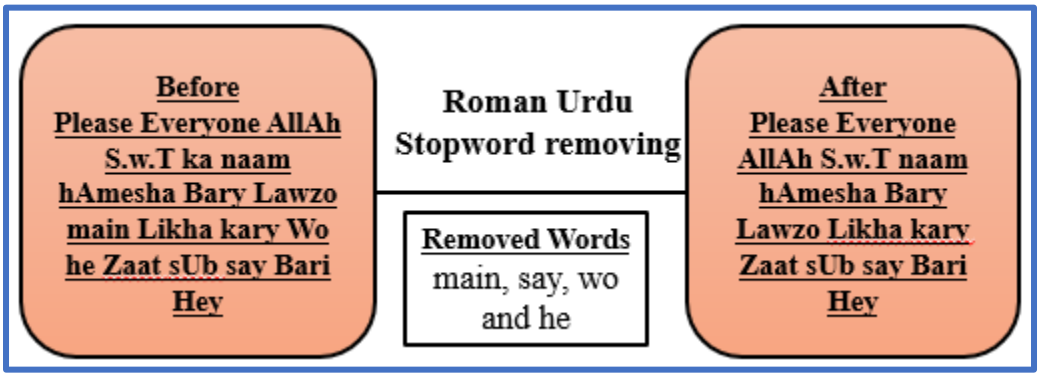

Figure 4 Example of Urdu Stopwords Removal

Figure 8 shows the example text of tweets before and after removing the Roman Urdu stopwords.

3.2.4 Removing Punctuations

The second last step of the NLP task is the removal of punctuation from the tweet text to extract the most meaningful words from the tweet. Removing punctuation is another common preprocessing step in natural language processing (NLP). Punctuation marks such as periods, commas, question marks, exclamation points, and various other symbols serve as important elements in text for conveying meaning and structure. However, in many NLP tasks. One example of how the text is before and after the punctuation removal is given in Figure 9.

The library of Python.

3.2.5 Removing Special Characters

In NLTK (Natural Language Toolkit), a special character typically refers to any character that is not a letter or a digit. Special characters include punctuation marks, symbols, whitespace characters (like spaces and tabs), and any other character that doesn't fall into the category of letters (alphabetic characters) or digits (numeric characters).

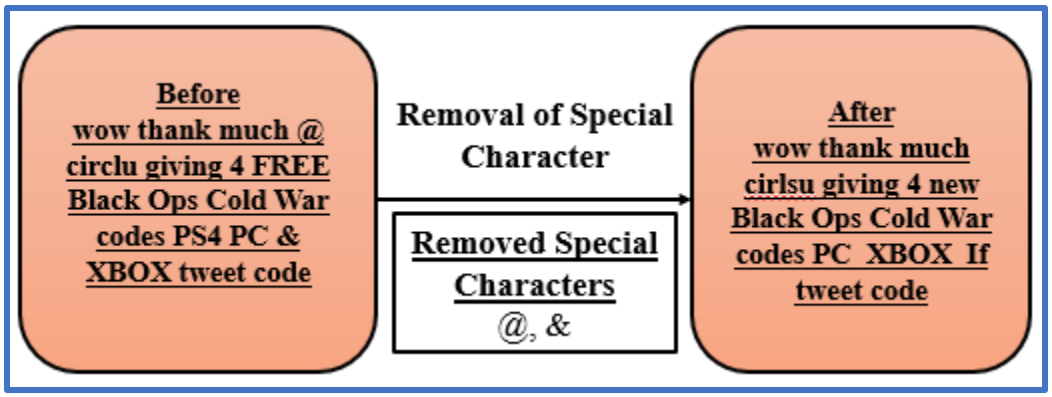

Figure 5 Example Text after Removal of Special Characters and Punctuation

3.3 Labeling and Annotating

Labeling of data is necessary to improve the accuracy and understanding of features for machine learning models. The act of recognizing raw data (pictures, text files, videos, etc.) and adding one or more relevant labels to give context and enable a machine learning model to learn from it is known as data labeling in machine learning.

All of the extracted features are in numeric form, except the class column. We change the class columns' values into categorical form by using the label encoder method given in the Python language. The labeling of class values Neutral, Positive, and Negative are used as 0, 1, and 2, respectively and also given in Table 3.

Table 2 Labeling of Class

Class	New Label
Neutral	0
Positive	1
Negative	2

3.4 Feature Extraction

In an NLP-based model, feature extraction and text preprocessing are an important part due to unstructured data. Feature extraction is a basic job in natural language processing (NLP) that entails transforming unprocessed text data into a format that machine learning algorithms can easily analyze. Two common text features (Polarity and Subjectivity) are extracted from tweets for the sentiment of text. We extract many other features like length of text in terms of words and in terms of characters, and details of each feature are given in this section.

3.4.1 Polarity

A sentence, phrase, or word's overall sentiment is referred to as its polarity. A "sentiment score" is a numerical rating that represents this polarity. This score, for instance, may be a figure between -100 and 100, where 0 would indicate a neutral mood. Text Blob library is used for extracting the sentiments of tweets by calculating the polarity and subjectivity values of each tweet.

3.4.2 Subjectivity

Subjectivity measures how much of the material is made up of factual facts and subjective opinions. Because of its increased subjectivity, the material is more likely to convey subjective viewpoints than accurate facts. Intensity is an additional setting for Text Blob. Text Blob uses the 'intensity' to compute subjectivity.

3.4.3 Length in Words

Number of words in the tweet text is also used as a feature in the dataset. The maximum number of extracted features is improving the accuracy of the models.

3.4.4 Length in Characters

The number of words in the tweet text is also used as a feature in the dataset. The maximum number of extracted features is improving the accuracy of the models.

Table 3 All Extracted Features

Text	Polarity	Subjectivity	No. of Words	No. of Characters
gearboxoffici sinc play borderland game friend i really natur taste someth more rememb i discov battleborn way buy stuff	0.22	0.44	172	30
gearboxoffici sinc play borderland game friend tast someth more i discov battleborn way buy now	-0.075	0.441667	126	23
Chri love borderland one two	0.5	0.6	28	5
Chri love Borderland one two	0.5	0.6	28	5
Chri love edg one two	0.5	0.6	21	5

3.5 Model Building and Evaluation

The Google Collab tool is used to build many ML (Naïve Bayes (NB), Random Forest (RF), SVM, and Decision Tree (DT) classifiers and one Hybrid Ensemble Model (HEM). To measure the accuracy of the proposed method, Naïve Bayes (NB), Random Forest (RF), SVM and Decision Tree (DT) classifiers and one Hybrid Ensemble Model (HEM) that combine the prediction of all 4 models are applied on the created dataset to classify users' emotional behavior into "neither," "hate," and "offensive" based on some features. Research also focuses on performing a complete comparison of ML models and evaluating these models on the basis of training accuracy and testing accuracy, and also calculating the precision, recall, and F1-Score for each model.

3.6 Classification

To measure the accuracy of the proposed method, Naïve Bayes (NB), Random Forest (RF), SVM and Decision Tree (DT) classifiers and one Hybrid Ensemble Model (HEM) that combine the prediction of all 4 models are applied on the created dataset to classify users' emotional behavior into "neither," "hate," and "offensive" based on some features.

4 Results and Discussions

The results of the proposed model are analyzed by calculating the Training accuracy, testing accuracy, and confusion matrix of all ML models. This study evaluated built machine learning models based on following parameters:

4.1 System Configuration and Tools

The method used for the experiments and implementation is explained in this section. Windows 10 operating system is installed on the dell system with core 15 and 6th generation laptop. The used system includes RAM of 8GB, hard disk of 320GB and SSD of 512GB. Different tools are used for implementing the and the model and writing of the original manuscript, such as Google Colab used for coding and MS Word used for writing.

4.2 Accuracy

A measure of the general accuracy of a model's predictions is accuracy. Out of all the occurrences in the dataset, the percentage of accurately predicted cases is calculated. The accuracy equation is:

Table 8 shows the training and testing accuracies of all the applied models. Figure 12 depicts the Confusion matrix graph of NB, and Figure 13 depict the evaluation results of NB. Training

accuracy RF is highest as compared to other models, with percentage of 97 and training accuracy of SVM is lowest as compared to the other 4 models.

Table 4 Training & Testing Accuracy of ML Models

Models	Training Accuracy	Testing Accuracy
NB	81	81
RF	97	97
DT	85	85.13
SVM	65.89	65.34
Hybrid	90	90

Testing accuracy of RF, DT, SVM, NB, and Hybrid is given in Table 8 with percentages of 97, 85.13, 65.34, 81, and 90, respectively.

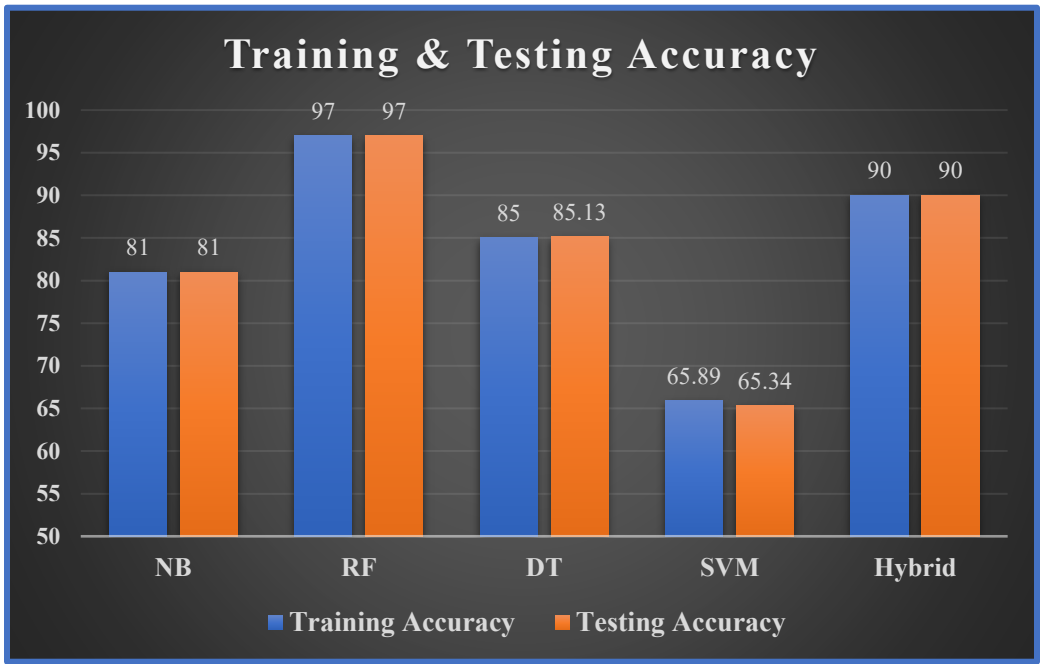

Figure 13 shows the screenshot for the values of precision, recall, and f1-score of the naïve Bayes model for every class 0,1 and 2. The precision of classes 0, 1, and 2 is 92 percent, 48 percent, and 83 percent, respectively. The recall for the model NB of classes 0, 1, and 2 is 70 percent, 30 percent and 96 percent, respectively. The f1-score for the model NB of class 0, 1, and 2 is 80 percent, 37 percent, and 89 percent, respectively.

Classification Report:

	precision	recall	f1-score
0	0.92	0.70	0.80
1	0.48	0.30	0.37
2	0.83	0.96	0.89

Figure 6 Screenshot of Evaluation Results of NB

Figure 14 shows the screenshot for the values of precision, recall, and f1-score of the RF model for every class 0,1 and 2. The precision of classes 0, 1, and 2 is 87 percent, 100 percent and 100 percent, respectively. The recall for the model RF of classes 0, 1 and 2 is 100 percent, 79 percent and 100 percent, respectively. The f1-score for the model RF of classes 0, 1, and 2 is 93 percent, 88 percent and 100 percent, respectively.

Classification Report:

	precision	recall	f1-score
0	0.87	1.00	0.93
1	1.00	0.79	0.88
2	1.00	1.00	1.00

Figure 7 Screenshot of Evaluation Results of RF

Figure 18 shows the screenshot for the values of precision, recall, and f1-score of the SVM model for every class 0,1 and 2. The precision of classes 0, 1, and 2 is 0 percent, 0 percent and 65 percent, respectively. The recall for the model SVM of classes 0, 1 and 2 is 0 percent, 0 percent and 100 percent, respectively. The f1-score for the model SVM of classes 0, 1 and 2 is 0 percent, 0 percent and 79 percent, respectively.

Classification Report:			
	precision	recall	f1-score
0	0.00	0.00	0.00
1	0.00	0.00	0.00
2	0.65	1.00	0.79

Figure 8 Screenshot of Evaluation Results of SVM

Classification Report:			
	precision	recall	f1-score
0	0.58	1.00	0.73
1	0.00	0.00	0.00
2	1.00	1.00	1.00

Figure 9 Screenshot of Evaluation Results of DT

Figure 21 shows the screenshot for the values of precision, recall, and f1-score of the Hybrid model for every class 0,1 and 2. The precision for the hybrid model of classes 0, 1, and 2 is 83 percent, 100 percent, and 91 percent, respectively. The recall for the model hybrid of classes 0, 1, and 2 is 100 percent, 30 percent, and 100 percent, respectively. The f1-score for the model RF of classes 0, 1, and 2 is 91 percent, 46 percent, and 95 percent, respectively.

Classification Report:			
	precision	recall	f1-score
0	0.83	1.00	0.91
1	1.00	0.30	0.46
2	0.91	1.00	0.95

Figure 10 Screenshot of the Evaluation Results of the Proposed

5 Conclusion

The proposed research focused on the combination of two different datasets of hate-speech tweets based on two different languages, namely Roman Urdu and English. The datasets were consolidated into a single file, and various NLP techniques were applied to process the dataset. In the intended methodology, NLP

techniques were applied using the Python language to extract optimized features from the hate-speech tweets. A dataset was then created that could be used by machine-learning tools to train and test the models. The extracted dataset was filtered by removing less important features through the application of the Principal Component Analysis (PCA) technique, and only the filtered and most informative features were retained. The Google Collab tool was employed to build multiple ML models. To measure the accuracy of the proposed method, Naïve Bayes (NB), Random Forest (RF), SVM, and Decision Tree (DT) classifiers, along with one Hybrid Ensemble Model (HEM) that combined the predictions of all four models, were applied to the created dataset. This classification aimed to categorize users' emotional behavior into "Neither," "Positive," and "Negative" based on certain features. The research also focused on performing a comprehensive comparison of ML models and evaluating these models based on training accuracy and testing accuracy. Additionally, precision, recall, and F1-score were calculated for each model.

REFERENCES

Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment Analysis of Omicron Tweets by using Machine Learning Models. "Journal of VFAST transaction on Software Engineering".

H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7649 LNCS, no. PART 1, pp. 508–524, 2012, doi: 10.1007/978-3- 642-35176-1_32.

Iqbal, M. K., Abid, K., U Din Ayubi, S., & Aslam, N. (2023). Omicron Tweet Sentiment Analysis Using Ensemble Learning. Journal of Computing & Biomedical Informatics, 4(02), 160-171.

J. Ling, "Coronavirus public sentiment analysis with BERT deep learning," Information, Commun. Soc., vol. 22, no. 13, pp. 2037–2038, 2019, DOI: 10.1080 /1369118x.2019.1620824.

Jacobs, C., Rakotonirina, N. C., Chimoto, E. A., Bassett, B. A., & Kamper, H. (2023). Towards hate speech detection in low-resource languages: Comparing ASR to acoustic word embeddings on Wolof and Swahili. arXiv preprint arXiv:2306.00410.

M. Mahyoob, J. Algaraady, M. Alrahiali, and A. Alblwi, "Sentiment Analysis of Public Tweets Towards the Emergence of SARS-CoV-2 Omicron Variant: A Social Media Analytics Framework," Eng. Technol. Appl. Sci. Res., vol. 12, no. 3, pp. 8525–8531, 2022, doi: 10.48084/etasr. 4865.

Munir, M. S., Parveen, K., Farooq, U., Shaalan, K., Abualkishik, A. Z., & Mohammed, A. S. (2022, October). Use of Different Machine Learning Algorithms for Hate Speech Detection. In the 2022 International Conference on Cyber Resilience (ICCR) (pp. 1-7). IEEE.

Pawar, A. B., Gawali, P., Gite, M., Jawale, M. A., & William, P. (2022, April). Challenges for Hate Speech Recognition System: Approach based on Solution. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 699-704). IEEE.

R. Patel, "Sentiment Analysis on Twitter Data Using Machine Learning by Ravikumar Patel A thesis submitted in partial fulfillment of the requirements for the degree of MSc Computational Sciences the Faculty of Graduate Studies," 2017.

William, P., Gade, R., esh Chaudhari, R., Pawar, A. B., & Jawale, M. A. (2022, April). Machine Learning based Automatic Hate Speech Recognition System. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 315-318). IEEE.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), 330-338.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2023). Optimized feature extraction and cross-lingual text reuse detection using ensemble machine learning models. Journal of Computing & Biomedical Informatics, 5(01), 26-40.

Abid, K., Aslam, N., Fuzail, M., Maqbool, M. S., & Sajid, K. (2023). An efficient deep learning approach for the prediction of student performance using a neural network. VFAST Transactions on Software Engineering, 11(4), 67-79.

Kanwal, F., Abid, M. K., Maqbool, M. S., Aslam, N., & Fuzail, M. (2023). Optimized classification of cardiovascular disease using machine learning paradigms. VFAST Transactions on Software Engineering, 11(2), 140-148.

Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). Understanding Urban Expansion Through Multi-Temporal Satellite Data Analysis. Kashf Journal of Multidisciplinary Research, 2(09), 252-273.

Hasnain, M. A., Ali, S., Malik, H., Irfan, M., & Maqbool, M. S. (2023). Deep learning-based classification of dental disease using X-rays. Journal of Computing & Biomedical Informatics, 5(01), 82-95.

Basit, A., Hanif, I., Maqbool, M. S., Qayyum, W., Hasnain, M. A., & Nazeer, R. (2023). Cross-lingual information retrieval in a hybrid query model for optimality. Journal of Computing & Biomedical Informatics, 5(01), 130-141.

Hasnain, M. A., Ali, Z., Maqbool, M. S., & Aziz, M. (2024). X-ray image analysis for dental disease: A deep learning approach using EfficientNet. VFAST Transactions on Software Engineering, 12(3), 147-165.

Rafiqee, M. M., Qaiser, Z. H., Fuzail, M., Aslam, N., & Maqbool, M. S. (2023). Implementation of an efficient deep fake detection technique on a video dataset using a deep learning method. Journal of Computing & Biomedical Informatics, 5(01), 345-357.

Maqbool, M. S., Fatima, N., Nazeer, R., Aslam, N., Abbas, F., Sumra, U., & Nadeem, M. (2025). A hybrid dataset-based ensemble strategy for efficient breast cancer detection. Kashf Journal of Multidisciplinary Research, 2(12), 39-57.

Muhammad Noman, Muhammad Sajid Maqbool, Dr. Naeem Aslam, Muqadas Nadeem, Hira Saleem, & Hanzla. (2026). Sleep disorder scoring is automated using advanced data science and machine learning techniques. Policy Research Journal, 4(3), 853–868. Retrieved from https://policyrj.com/1/article/view/1713

Zainab Naveed, Rubaina Nazeer, Muhammad Sajid Maqbool, Dr. Naeem Aslam, Hira Saleem, & Muqadas Nadeem. (2026). An end-to-end orthopedic disease image classification system using convolutional neural networks.

Mahnoor Zaman, Nosheen Fatima, Muhammad Sajid Maqool, Dr. Naeem Aslam, Rubaina Nazeer, & Hira Saleem. (2026). Ingredient: Intelligent CNN for food ingredient recognition and classification. Policy Research Journal, 4(3), 789–805.

Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). Understanding Urban Expansion Through Multi-Temporal Satellite Data Analysis. Kashf Journal of Multidisciplinary Research, 2(09), 252-273.

M. A., Ali, Z., Maqbool, M. S., & Aziz, M. (2024). X-ray image analysis for dental disease: A deep learning approach using EfficientNet. VFAST Transactions on Software Engineering, 12(3), 147-165.

Aslam, N., Maqbool, M. S., Nadeem, M., & Saleem, H. (2026). DEEPFAKESHIELD: ENHANCED VIDEO AUTHENTICITY DETECTION VIA CONVOLUTIONAL VISION TRANSFORMER. Spectrum of Engineering Sciences, 4(3), 1650-1665.

Aslam, M., Maqbool, M. S., Aoun, M., Aslam, N., Razzaq, A. M., & Ali, S. (2026). HIGH-PERFORMANCE AND EFFICIENT BRAIN TUMOR SEGMENTATION FOR ENHANCED CLINICAL ANALYSIS. Spectrum of Engineering Sciences, 4(3), 195-210.

Maqbool, M. S., Zahra, S. R., Ismail, S., Nadeem, M., Fatima, N., & Ahmad, J. (2026). A CNN-BASED FRAMEWORK FOR EFFICIENT DETECTION OF EYE DISEASE IN FUNDUS IMAGES. Spectrum of Engineering Sciences, 4(4), 1157-1169.

Farwa Zainab, Farwa Nazim, Muhammad Kashaf, Naeem Aslam, & Muhammad Sajid Maqbool. (2026). PREDICTIVE ANALYTICS FOR CUSTOMER CHURN IN SUBSCRIPTION-BASED BUSINESSES USING MACHINE LEARNING. Spectrum of Engineering Sciences, 4(4), 596–618. Retrieved from https://thesesjournal.com/index.php/1/article/view/2460.

Meiraj Aslam, Mohammad Sajid Maqbool, Muhammad Aoun, Naeem Aslam, Abdul Manan Razzaq, Abdul Manan Razzaq, & Salman Ali. (2026). HIGH-PERFORMANCE AND EFFICIENT BRAIN TUMOR SEGMENTATION FOR ENHANCED CLINICAL ANALYSIS. Spectrum of Engineering Sciences, 4(3), 195–210. Retrieved from https://thesesjournal.com/index.php/1/article/view/2169.

Syeda Qanitah Naqvi, Syeda Rabail Zahra, Muhammad Sajid Maqbool, Muqadas Nadeem, Hira Saleem, & Mahnoor Zaman. (2026). AN AUTOMATED AND ARTIFICIAL INTELLIGENCE-BASED SYSTEM FOR THE DIAGNOSIS OF SKIN CANCER. Policy Research Journal, 4(4), 58–72. Retrieved from https://policyrj.com/1/article/view/1769.

Sarim Javed, Muhammad Sajid Maqbool, Dr. Naeem Aslam, Muhammad Haseeb Ur Rehman, Muqadas Nadeem, & Hira Saleem. (2026). HIGH ACCURACY INTRUSION DETECTION IN IOT VIA HYBRID ML DL MODELS. Policy Research Journal, 4(4), 73–85. Retrieved from https://policyrj.com/1/article/view/1770.

Rabia Hassan, Muhammad Sajid Maqbool, Dr. Naeem Asalam, Ariba Afzal, Hira Saleem, & Muqadas Nadeem. (2026). AN IN-DEPTH STUDY ON STUDENTS' PERFORMANCE EVALUATION USING MULTIPLE MACHINE LEARNING CLASSIFIERS AND DATA ANALYTICS APPROACHES. Policy Research Journal, 4(4), 86–99. Retrieved from https://policyrj.com/1/article/view/1771.

Rabia Ikhlaq, Muhammad Sajid Maqbool, Hira Saleem, Dr. Naeem Aslam, Zeeshan Manzoor, & Muqadas Nadeem. (2026). DEEP CONVOLUTIONAL NEURAL NETWORKS FOR AUTOMATED BREAST CANCER DIAGNOSIS. Policy Research Journal, 4(4), 170–182. Retrieved from https://policyrj.com/1/article/view/1795.

Izhar, Dr. Naeem Aslam, Muhammad Sajid Maqbool, Muqadas Nadeem, & Hira Saleem. (2026). DEEPFAKESHIELD: ENHANCED VIDEO AUTHENTICITY DETECTION VIA CONVOLUTIONAL VISION TRANSFORMER. Spectrum of Engineering Sciences, 4(3), 1650–1665. Retrieved from https://thesesjournal.com/index.php/1/article/view/2360

Tanveer, K., Aslam, N., Saleem, H., Maqbool, M. S., Akhter, M., & Rashid, M. H. (2026). INTEGRATING BLOCKCHAIN AND MACHINE LEARNING FOR ROBUST IOT DATA INTEGRITY IN CRITICAL INFRASTRUCTURES. Spectrum of Engineering Sciences, 4(3), 1875-1886.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). A practical guide to sentiment analysis (Vol. 5). Cham: Springer International Publishing.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143-157.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR), 49(2), 1-41.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 538-541).

Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2023). Optimized Feature Extraction and Cross-Lingual Text Reuse Detection using Ensemble Machine Learning Models. Journal of Computing & Biomedical Informatics, 5(01), 26-40.

Kanwal, F., Abid, M. K., Maqbool, M. S., Aslam, N., & Fuzail, M. (2023). Optimized Classification of Cardiovascular Disease Using Machine Learning Paradigms. VFAST Transactions on Software Engineering, 11(2), 140-148.

Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.

Goularas, D., & Kamis, S. (2019, August). Evaluation of deep learning techniques in sentiment analysis from twitter data. In 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML) (pp. 12-17). IEEE.