INTERNATIONAL JOURNAL OF ELECTRONICS AND COMPUTER APPLICATIONS

ORIGINAL ARTICLE

Enhancing Social Media Analytics with NLP Pipeline for Trend Detection and Sentiment

Payal Ladhane[1*], Shivani Talole[1], Shraddha Gadsing[1], Nirjala Pund[1], Barkha Shahaji[2]

1 Student, Computer Engineering Department, Vidya Prathishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati, Maharashtra, India.
2 Assistant Professor, Computer Engineering Department, Vidya Prathishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati, Maharashtra, India.

OPEN ACCESS
Received: 29-07-2025
Accepted: 28-11-2025
Published: 12-12-2025

Citation: Ladhane P, Talole S, Gadsing S, Pund N, Shahaji B. Enhancing Social Media Analytics with NLP Pipeline for Trend Detection and Sentiment. 2025; 2(2):74-80. https://doi.org/10.70968/ijeaca.v2i2.ML117

* **Corresponding author.**
payal.ladhane@gmail.com

Funding: None

Competing Interests: None

ISSN
Electronic: 3048-8257

Abstract

Due to the rising popularity of social media, the necessity of automated systems that can monitor, detect harmful content, find sentiment, and detect emerging trends is becoming more urgent [1,2]. A unified Natural Language processing(NLP)[6] pipeline performing sentiment analysis and trend extraction using deep learning methods is proposed in this work. Through public APIs, live textual data is gathered and passed through preprocessing steps after which TF-IDF based feature extraction and initial polarity checking [6,7]. For a better understanding of sentiment patterns, an adapted RoBERTa model is used, and its performance is evaluated compared to widely used Machine Learning methods. According to the experimental observations, RoBERTa shows more robustness and deals with informal social media language better than others. The framework not only entails the sentiment classification of user tweets but also detecting trending topics by the trend in the frequency of keywords and temporal topic model[5,10]. This combination allows to keep a record of the discussions and its time patterns. In general, the suggested method provides a scalable and efficient solution for social media analytics as well as automated content monitoring[12,16].

Keywords: Natural Language Processing (NLP), Sentiment Analysis, Trend Detection, Deep Learning, RoBERTa, Text Classification, Machine Learning, RealTime Data Analytics

Introduction

The advent of social media sites including Twitter, Reddit, Instagram, and YouTube has dramatically increased their importance as key sources of publicly shared opinion, realtime knowledge exchange, as well as social engagement. User generated content in the form of millions of posts every day embodies the emotions, opinions, and responses of the masses regarding various aspects of business, politics, medicine, and governance. Such mass amounts of dynamic, unstructured, and constantly changing text data hold immense prospects of analysis of the public sentiment and the developing trends, although the challenges of analysis lie in their unstructured, dynamic, and velocity nature. This study introduces an improved social media analytics framework designed to not only analyze sentiment but also identify emerging trends, providing a more complete understanding of online discussions. The system gathers real-time textual data through APIs, cleans and preprocesses it, and applies TF-IDF techniques along with polarity scoring to extract meaningful features. A finetuned RoBERTa model is used for more accurate sentiment classification. The model is evaluated against popular machine learning techniques, including Na ̈ıve Bayes, Support Vector Machines (SVM), Logistic Regression, Random Forest, k-Nearest Neighbors (k-NN), Artificial Neural Networks, as well as the transformer- based BERT model[6, 7, 14]. The findings suggest that the RoBERTa model is able to better understand context and achieve better classification accuracy compared to the other models especially for informal and rapidly evolving social media language[14, 15]. Recent studies suggest that NLP systems should not only concentrate on sentiment identification but also monitor how discussions and opinions shift over time[5, 16]. Considering this idea, the framework uses keyword frequency, time, and topic clustering to capture an event on a real-time basis as it breaks out[5, 10]. The system provides a practical and flexible solution for monitoring social media by linking a deep learning-based sentiment analysis[3, 4] with real-time trend identification.

Table 1: Literature Review Summary

Ref	Year	Concept	Parameter	Dataset	Findings	Limitation
(1)	2023	Trend Detection + NLP	Accuracy, Classification	Twitter Dataset	Detects manipulation trends	No deep sentiment analysis
(2)	2024	NLP Topic Detection	Accuracy, Semantic Similarity	News Dataset	Identifies emerging topics	Weak on noisy text
(3)	2024	Sentiment Analysis	Accuracy, Interpretability	Social Media Dataset	Improves implicit sentiment	High computation cost
(4)	2023	NLP + Mobility Analysis	Prediction Accuracy, Efficiency	Social Media Data	Extracts mobility patterns	Needs location data
(5)	2023	Word2Vec + Text Mining	Accuracy, Similarity	Big Text Data	Detects semantic trends	Requires large dataset
(6)	2023	CNN + Autoencoder	Accuracy, Temporal Analysis	Social Media Dataset	Tracks opinion trends	Complex model
(7)	2022	BERT-CNN	Accuracy, Recall	Twitter Dataset	Improves sentiment detection	Poor short text handling
(8)	2023	Deep Learning Trend Prediction	Accuracy, Popularity Score	Weibo Dataset	Predicts opinion spread	Depends on interaction
(9)	2024	Emoji + ML	Precision, Recall, F1-score	Twitter Dataset	Enhances sentiment detection	Limited dataset diversity
(10)	2024	BERT + Emoji Embedding	Accuracy, Precision	Urdu Tweets Dataset	Better contextual sentiment	High resource usage
(11)	2024	BiLSTM + GRU	Accuracy, F1-score	Twitter Dataset	Captures emotional context	High complexity
(12)	2024	Hybrid Sentiment Model	Accuracy, ROC-AUC, F1-score	Review Dataset	Improves minority sentiment	Data dependent
(13)	2025	BERT + GCN	Accuracy, Classification	Twitter Dataset	Improves classification	High complexity
(14)	2024	Pretrained Models	Accuracy, Model Comparison	Multiple Datasets	High performance	High computation cost
(15)	2024	Deep NN + LLM	Accuracy, Analysis Depth	Aspect Dataset	Better opinion analysis	Needs large data
(16)	2021	BERT-LDA	Accuracy, Topic Coherence	Social Media Dataset	Combines topic + sentiment	Large data needed
(17)	2024	LSTM Model	Accuracy, Precision, Recall	Twitter Dataset	Detects hate speech	Limited scope
(18)	2023	Aspect-Based SA	Accuracy, Aspect Precision	Review Dataset	Provides aspect insights	Complex features
(19)	2024	Lexicon-Based SA	Accuracy, Processing Time	Multiple Datasets	Fast processing	Low consistency
(20)	2023	D-RNN	Accuracy, Classification	Twitter Dataset	Improves sentiment accuracy	High computation

Related Work

The rapid increase in the user-generated posts on sites like Twitter, YouTube, and Reddit has led to increasing research on social media analytics to understand public opinion sentiment and developing trends[1, 2]. The large volume, unstructured nature, and diverse languages used in the information can generate noise, contextual ambiguity, and linguistically varied issues that cause the need for automatic analysis frameworks. The initial research work on sentiment analysis made use of dictionary-based approach and statistical techniques, in which pre-defined sentiment word lists were in turn used to identify the polarity of text[3]. The straightforwardness and interpretability of these methodologies made it difficult for them to capture context, sarcasm, or the informal language often used in social media discourse[4]. To surmount earlier limitations, research works attempted to combine machine learning methods with the natural language processing to build a proper pipeline for sentiment classification[5]. These pipelines usually start with preprocessing steps like cleaning the text, tokenization, removing stop words, and normalizing the content to improve data quality[6]. Methods like Bag of Words, n-grams, and Term Frequency - Inverse Document Frequency are employed to transform textual data into numerical representations, which are then used by categorizer like Na ̈ıve Bayes.

As representation learning progressed, semantic embedding methods were introduced to better capture word meaning in context[8]. Techniques such as Word2Vec helped model relationships between words, which improved the quality of sentiment prediction[8]. Newly , transformer based models like BERT are used to produce contextual embeddings, leading to stronger performance on sentiment analysis tasks across varied and noisy social media data[9, 14].

Beyond sentiment classification, identifying trends has become a key focus in social media analytics[5]. Topic modeling techniques, including Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF)[10] are mostly used to identify underlying topics and evolving patterns within large-scale text data. When supported by frequency analysis and clustering, these approaches help detect popular hashtags, keywords, and discussion themes as they emerge in real time[5].

Recent studies highlight the benefit of combining sentiment analysis with trend identification to better understand patterns in social media activity[12]. Linking sentiment polarity with emerging topics allows such integrated systems to reveal clearer insights into public perception and emotional reactions, supporting modern NLP pipelines used for real-time monitoring and analysis[5, 16].

Ease of Use

Activities like data collection and cleaning, feature extraction, sentiment analysis, trend detection, and visualization, all perform their function independently but work well together in harmony. Due to this arrangement users can change or add to one part of the system, without disrupting the rest of the workflow. The implementation is based on the widely used open-source tools like Python, NLTK, Scikitlearn, TensorFlow/Keras, and Tweepy, therefore, anyone with rudimentary knowledge of ML and NLP can use or adapt it. Once set up, the system operates with minimal manual intervention. It can automatically take in live social media data and deliver useful analytical results. So, it becomes easier to interpret the analytical outcomes with the inclusion of visual elements such as sentiment charts, word clouds, and trend graphs. These dashboards convey complicated information in an uncomplicated manner thus making it easy for an analyst or decision-maker to comprehend how public opinion has shifted without the need for advanced technical skills. The simplicity of the setup, automation of function and the easy to read outputs of the system render the system practical, flexible and apt for doing real-world social media analytics tasks.

Proposed Work

The proposed work of study including the proposed framework, and algorithmic detail:

A. Dataset

For preparing the dataset, live Twitter posts were collected using the Tweepy API. The retrieved raw data was then organized into Pandas Data Frames in Python, making it easier to clean, transform, and remove unwanted noise before analysis.

B. Data Acquisition

The process begins by collecting live posts through public APIs, such as the Twitter API, using specific keywords, hashtags, or topics as search filters. This allows the system to continuously receive fresh content that reflects ongoing public conversations. The fetched data, usually available in JSON or text form, includes the actual post, time details, and other useful metadata, which are then arranged into a structured format to make further analysis easier and more organized.

C. Data Preprocessing

Social media data is highly noisy and unstructured, containing emojis, URLs, special characters, slang, and redundant content[3, 4]. To address these issues, preprocessing techniques are applied[6, 7]. Text normalization ensures consistency across linguistic variations, while language filtering removes irrelevant or unsupported content. This stage significantly improves data quality and prepares the text for feature extraction and model training.

D. Feature Extraction

At this stage, unprocessed text data transformed into numerical representations that can be utilized by machine learning algorithms[6]. Methods like TF-IDF and n-gram modeling are applied to measure the significance of words and phrases within the text[6, 7]. For richer semantic understanding, embedding techniques such as Word2Vec and contextual representations from transformer models like BERT are used[8, 14]. These embeddings preserve contextual relationships between terms, improving the quality of sentiment detection and trend analysis[8, 16].

E. Trend Detection

The trend detection component tracks rising topics, popular hashtags, and influential keywords through topic modeling and clustering approach such as Latent Dirichlet Allocation and Non-negative Matrix Factorization. Through these techniques presumably concealed themes and shifts in public attention are revealed. Analyzing these subject patterns along with sentiment polarity the system gives even a clearer idea about how people respond to and make sense of social ills.

F. Sentiment Analysis

The sentiment analysis module automatically assigns each social media post a positive, negative or neutral class[7]. Deep learning models in the form of LSTM networks are leveraged in conjunction with transformer architectures such as BERT[14, 17] and RoBERTa which are inherently suited to sequence and contextual understanding in text. This method helps the system interpret user sentiments, feelings, and attitudes toward various topics at different times[12, 16].

G. Visualization and Insight Generation

Makes Complex Analysis simpler and clearer. Graphs such as bar charts, word clouds and time series graphs illustrate things like sentiment distribution, most common hashtags discussed and changes in sentiment over time. Take the outputs of the analysis and produce insights from these which people can interpret without understanding the technique. Consequently, the evolving sentiments enable companies, officials, analysts, researchers and many more to quickly comprehend the changing public opinion, track the social conversation in real-time and make informed, data-led decision-making.

Algorithms Used

A. Sentiment Analysis Algorithms

1. Logistic Regression:

It is a supervised learning technique mainly used for tasks like classification. Unlike linear regression, which estimates continuous numerical outputs, logistic regression calculates the likelihood that a given input belongs to a particular class.

It is commonly applied to problems such as binary classification, where outcome has only two possible categories, such as Yes/No, True/False, or 0/1. This method uses sigmoid function for transform input values into probabilities ranging between 0 and 1.

In logistic regression, the output of a linear equation is passed through the sigmoid function to convert continuous values into discrete class predictions. The sigmoid (logistic) function maps any real-valued input into a probability value between 0 and 1, enabling classification.

$$z = w^T x + b \qquad (1)$$

The probability estimation in logistic regression is given as follows:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}}$$

$$J(\theta) = -\frac{1}{m}[y^T \log(h) + (1-y)^T \log(1-h)] \qquad (2)$$

In context of sentiment classification, Logistic Regression maps high-dimensional TF-IDF vectors to sentiment labels by determining an optimal decision boundary between classes. Its advantages include computational efficiency, interpretability, and effectiveness in handling sparse and high-dimensional text data[6, 7]. As a linear classifier, it has limited ability to represent intricate contextual patterns when compared with deep learning techniques. Still, because it is straightforward to implement and consistently dependable, Logistic Regression is often used as a baseline for assessing the performance of more sophisticated machine learning and transformer-based models in social media sentiment analysis[14].

2. RoBERTa:

RoBERTa is a deep learning model based on Transformer architecture, developed for high-performance natural language understanding tasks. It builds using original BERT model by introducing improved training techniques and utilizing larger and more diverse datasets.

The model relies entirely on the Transformer encoder and uses self-attention to understand relationships between words within a sentence. Compared to BERT, RoBERTa eliminates the Next Sentence Prediction (NSP) task and instead applies dynamic masking during training. These changes help the model learn more detailed and effective language representations.

Mathematically, RoBERTa is trained using the Masked Language Modeling (MLM) objective, where randomly selected tokens in a sentence are masked and predicted based on surrounding context.

- Masked Language Modeling Loss

$$LMLM = -X\log P(xi \mid x \setminus M)$$
$$i \in M$$

where L_{MLM} denotes the masked language modeling loss. Here, M denotes the set of masked token indices, x_i represents the original token at position i, and $x_{\setminus M}$ indicates the input sequence in which the masked tokens have been replaced, $P(\cdot)$ represents the probability function, log denotes the natural logarithm, and | represents conditional probability.

- Classification Layer Formula

$$y\hat{} = \text{softmax}(Wh_{[CLS]} + b)$$

where $\hat{y}$ represents the predicted probability distribution, softmax(·) is the activation function that converts logits into probabilities, W denotes the weight matrix of the classification layer, $h_{[CLS]}$ represents the contextual embedding of the special classification token $[CLS]$, and b denotes the bias vector[14].

Sentiment analysis and text classification tasks, RoBERTa is fine-tuned by adding task-specific classification layer on top of the pretrained model. During fine-tuning, the model adjusts its weights based on labeled data, enabling accurate prediction of sentiment polarity[15]. Compared to traditional machine learning models and earlier transformer architectures, RoBERTa demonstrates superior contextual understanding, robustness to informal language, and higher classification accuracy. Due to these advantages, it is widely adopted in research applications involving sentiment analysis, trend detection, question answering, and other NLP tasks.

B. Trend Detection Algorithms

1. Normalized Frequency Formula

A statistical approach that identifies trending keywords and hashtags based on their occurrence frequency within defined time windows.

$$F_{t,i} = \frac{f_{t,i}}{N_i} \tag{4}$$

where $F_{t,i}$ represents the normalized frequency of term t within time window i is represented using $f_{t,i}$, which denotes the raw count of term t observed in that specific time interval. and N_i represents total number of documents in time window i[5, 10].

2. Average Growth Rate (AAGR) Formula

Using the Word2Vec are applied to uncover latent semantic structures and cluster tweets into coherent topics, enabling effective detection of emerging discussion trends.

$$AAGR_t = \frac{\sum_{i=\frac{T}{2}+1}^{T} F_{t,i} - \sum_{i=1}^{\frac{T}{2}} F_{t,i}}{\sum_{i=1}^{\frac{T}{2}} F_{t,i}} \tag{5}$$

where $AAGR_t$ denotes the average growth rate of term t, T represents the total number of time windows, and $F_{t,i}$ is the normalized frequency of term t at time window i[5, 8].

Results And Discussion

The developed system provides an interactive dashboard for real-time sentiment analysis and trend detection. (Fig. 1) illustrates the user interface of the application where users can select trending topics such as Renewable Energy, Artificial Intelligence, and Climate Change. The system fetches realtime news data and performs sentiment analysis using the RoBERTa model.

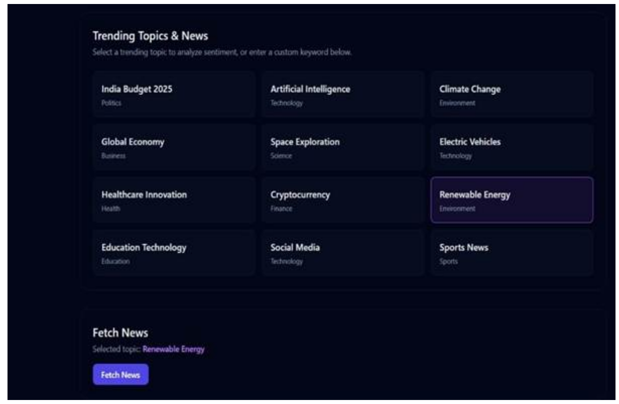

Fig. 1: Trending Topics Selection Dashboard

(Fig. 2) shows the fetched news articles related to the selected topic. The system allows users to select a specific news item for detailed sentiment evaluation.

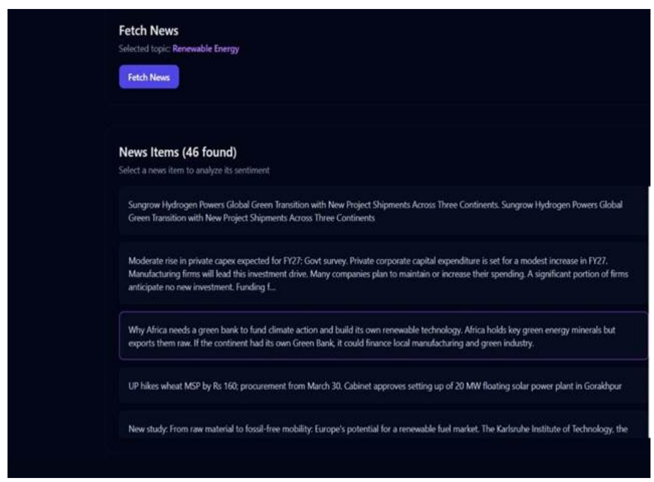

Fig. 2: News Fetching Module

(Fig. 3) illustrates the sentiment analysis outcomes, showing how positive, negative, and neutral sentiments are distributed. In the selected example, the majority of the sentiment is neutral (63.38%), followed by positive (30.94%) and negative (5.68%).

The system also provides an explanation of the predicted sentiment along with confidence scores, enhancing interpretability and user understanding.

Fig. 3: Sentiment Analysis Results

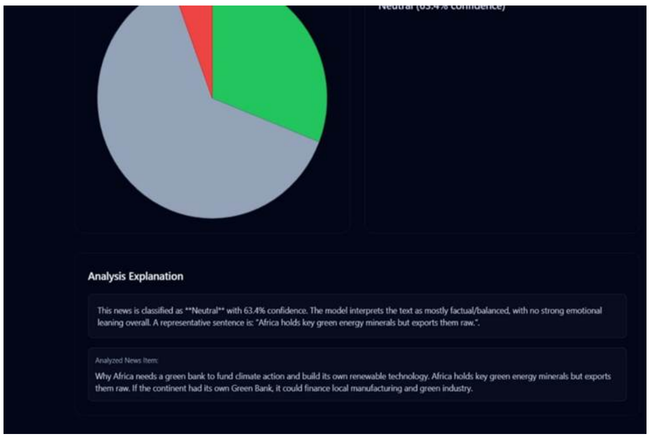

Fig. 4: Detailed Sentiment Explanation

Conclusion

This study demonstrates how a combined NLP pipeline can effectively analyze sentiment while simultaneously detecting emerging trends from live social media data[5]. The proposed framework integrates multiple components, including data acquisition, text cleaning and preprocessing, feature extraction, sentiment classification, and trend analysis, into a cohesive pipeline. This integrated design supports the effective handling and examination of high-volume, real-time social media streams by leveraging classical machine learning techniques and modern deep learning models, like RoBERTa, the system achieves strong classification performance while effectively capturing contextual and sequential relationships within user generated text.

The application shows the connection between the sentiment categories (positive, negative, neutral) and the emerging elements e.g. topics, hashtags, keywords, etc., which is useful to get changing public opinions. Different from manual observation or fixed analysis, this solution is more automated, scalable, and able to perform real-time analysis. Even though it currently faces challenges such as noisy data, sarcasm and multi-linguality, the system can be usefully applied in market research, monitoring public opinion, and analysing social media trends.

In conclusion, the implemented NLP pipeline shows how unstructured social media data can be systematically converted into meaningful and actionable insights, supporting the creation of intelligent, scalable, and real-time analytics systems.

References

1. Kausar S, Tahir B, Mehmood MA. Manipify: An Automated Framework for Detecting Manipulators in Twitter Trends. *Journal of Social Computing*. 2023;4(1):46-61. Available from: 10.23919/jsc.2023.0001
2. Khan ZA, Xia Y, Ali S, Khan JA, Askar SS, Abouhawwash M, *et al.* Identifying Hot Topic Trends in Streaming Text Data Using News Sequential Evolution Model Based on Distributed Representations. *IEEE Access*. 2023;11:98787-98804. Available from: 10.1109/access.2023.3312764
3. Wang A, Jiang H, Ma Y, Jiang J, Liu A, Okazaki N. Improving Implicit Sentiments Analysis via Explanations of Multiple Perspectives. *IEEE Access*. 2025;13:61136-61148. Available from: 10.1109/access.2025.3556762
4. Fontes T, Murços F, Carneiro E, Ribeiro J, Rossetti RJF. Leveraging Artificial Intelligence to Write Strategic Content on Social Media Platforms: A Tool-Based Approach. *IEEE Open Journal of Intelligent Transportation Systems*. 2023;4:663-681. Available from: 10.1109/OJITS.2023.3308210
5. Lobanova PA, Kuzminov IF, Karatetskaia EY, Sabidaeva EA, Anpilogov VV. Trend Detection Using NLP as a Mechanism of Decision Support. *Scientific and Technical Information Processing*. 2023;50(5):440-448. Available from: 10.3103/s0147688223050106
6. Sun W, Xiao R, Li X. Online Public Opinion Situation Assessment by Deep Sparse Autoencoder and Time-Series Modeling Convolutional Neural Network. *IEEE Access*. 2025;13:152198-152218. Available from: 10.1109/access.2025.3602491
7. Gupta T, Rai A, Roy S. SentimentMapper: A framework for mapping of sentiments towards disaster response using social media data. *Applied Intelligence*. 2025;55(7). Available from: 10.1007/s10489-025-06442-0
8. Yang Y, Fan C, Gong Y, Yeoh W, Li Y. Forwarding in Social Media: Forecasting Popularity of Public Opinion With Deep Learning. *IEEE Transactions on Computational Social Systems*. 2025;12(2):749-763. Available from: 10.1109/tcss.2024.3468721
9. Alfreihat M, Almousa OS, Tashtoush Y, AlSobeh A, Mansour K, Migdady H. Emo-SL Framework: Emoji Sentiment Lexicon Using Text-Based Features and Machine Learning for Sentiment Analysis. *IEEE Access*. 2024;12:81793-81812. Available from: 10.1109/access.2024.3382836
10. Narejo KR, Zan H, Oralbekova D, Dharmani KP, Orken M, Mukhsina K. Enhancing Emoji-Based Sentiment Classification in Urdu Tweets: Fusion Strategies With Multilingual BERT and Emoji Embeddings. *IEEE Access*. 2024;12:126587-126600. Available from: 10.1109/access.2024.3446897
11. Sherin A, Jeya IJS, Deepa SN. Enhanced Aquila Optimizer Combined Ensemble Bi-LSTM-GRU With Fuzzy Emotion Extractor for Tweet Sentiment Analysis and Classification. *IEEE Access*. 2024;12:141932-141951. Available from: 10.1109/access.2024.3464091
12. Razali MN, Manaf SA, Hanapi RB, Salji MR, Chiat LW, Nisar K. Enhancing Minority Sentiment Classification in Gastronomy Tourism: A Hybrid Sentiment Analysis Framework With Data Augmentation, Feature Engineering and Business Intelligence. *IEEE Access*. 2024;12:49387-49407. Available from: 10.1109/access.2024.3362730
13. Li X, Yang H, Sun X, Wei F. A Multimodal Aspect-Level Sentiment Analysis Model Based on Syntactic-Semantic Perception. *IEEE Access*. 2025;13:184220-184232. Available from: 10.1109/access.2025.3624546
14. Shafikuzzaman M, Islam MR, Rolli AC, Akhter S, Seliya N. An Empirical Evaluation of the Zero-Shot, Few-Shot, and Traditional Fine-Tuning Based Pretrained Language Models for Sentiment Analysis in Software Engineering.

IEEE Access. 2024;12:109714-109734. Available from: 10.1109/access.2024.3439450

15. Mughal N, Mujtaba G, Shaikh S, Kumar A, Daudpota SM. Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis. *IEEE Access*. 2024;12:60943-60959. Available from: 10.1109/access.2024.3386969
16. Tan X, Zhuang M, Lu X, Mao T. An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model. *IEEE Access*. 2021;9:15860-15871. Available from: 10.1109/access.2021.3052566
17. Roy SS, Roy A, Samui P, Gandomi M, Gandomi AH. Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach. *IEEE Transactions on Computational Social Systems*. 2024;11(4):5028-5037. Available from: 10.1109/tcss.2023.3260217
18. Maroof A, Wasi S, Jami SI, Siddiqui MS. Aspect-Based Sentiment Analysis for Service Industry. *IEEE Access*. 2024;12:109702-109713. Available from: 10.1109/access.2024.3440357
19. Mahmoudi A, Jemielniak D, Ciechanowski D. Assessing Accuracy: A Study of Lexicon and Rule-Based Packages in R and Python for Sentiment Analysis. *IEEE Access*. 2024;12:20169-20180. Available from: 10.1109/access.2024.3353692
20. Durga P, Godavarthi D. Deep-Sentiment: An Effective Deep Sentiment Analysis Using a Decision-Based Recurrent Neural Network (D-RNN). *IEEE Access*. 2023;11:108433-108447. Available from: 10.1109/access.2023.3320738