Rover Publications
United International Journal of Engineering and Sciences (UIJES)
An International Peer-Reviewed (Refereed)Engineering and Science Journal
Impact Factor: 8.472 (SJIF) Vol-6, Issue- 2,(April, May &June);2026 ISSN:2582-5887 www.uijes.com

AI-Powered Social Media Content Analyzer for Sentiment, Engagement, and Trend Prediction

Pithani Gopala Krishna
Affiliations: Department of Computer Science and Engineering, Kakinada Institute of Technology and Science, Divili-533433, Kakinada, Andhra pradesh,India
Email IDs: krishna35gopal@gmail.com

Article Received:12-03-2025 Article Modified:13-04-2026
Article Accepted:17-04-2026 Article Published:18-04-2026
DOI:10.37854/UIJMR.2026.3.2.42

Abstract

The exponential growth of social media data has necessitated advanced automated systems for real-time analysis of user behavior and public sentiment. This paper proposes an "AI-Powered Social Media Content Analyzer," a multi-functional framework designed to perform sentiment classification, engagement forecasting, and trend prediction. Utilizing Natural Language Processing (NLP) techniques such as lemmatization and TF-IDF vectorization, the system processes raw text from platforms including Twitter, Instagram, and Facebook. We evaluate several machine learning models, including Linear Support Vector Classification (LinearSVC), Logistic Regression, and Random Forest. Our results demonstrate that the Random Forest model achieves the highest sentiment classification accuracy of 80.91%. Furthermore, we implement a Random Forest Regressor to predict post engagement (likes/shares) based on platform and sentiment profiles. This research underscores the effectiveness of combining specialized NLP preprocessing with ensemble learning to extract actionable insights from big social data.

Keywords: Artificial Intelligence, NLP, Sentiment Analysis, Social Media Analytics, Trend Prediction, Machine Learning.

Rover Publications
United International Journal of Engineering and Sciences (UIJES)
An International Peer-Reviewed (Refereed)Engineering and Science Journal
Impact Factor: 8.472 (SJIF) Vol-6, Issue- 2,(April, May &June);2026 ISSN:2582-5887 www.uijes.com

I. INTRODUCTION

In the contemporary digital era, social media platforms have transformed into primary hubs for public discourse, brand interaction, and global information dissemination. With billions of active users generating massive volumes of unstructured text daily, manual analysis has become impractical. Organizations and researchers require automated tools to gauge public mood, predict the virality of content, and identify emerging cultural trends.

Sentiment analysis, also known as opinion mining, serves as the cornerstone of this requirement. However, basic positive/negative polarity is often insufficient. Real-world sentiment is nuanced, encompassing a wide range of emotions such as joy, anger, fear, and surprise. Moreover, the dynamic nature of social media means that sentiment is closely tied to engagement (likes, shares) and longitudinal trends.

The role of Artificial Intelligence (AI) and Machine Learning (ML) is critical in addressing these challenges. By leveraging advanced NLP libraries and high-performance classification algorithms, it is possible to build resilient systems that not only understand "what" is being said but also "how" it will be received. The motivation behind this project is to create an integrated dashboard that provides a 360-degree view of social media content effectiveness.

II. LITERATURE REVIEW

The field of sentiment analysis has evolved significantly over the last decade. Early research predominantly utilized lexicon-based approaches (e.g., SentiWordNet) and rule-based systems like VADER (Valence Aware Dictionary and sEntiment Reasoner) [1]. While effective for social media slang, these methods often struggle with complex semantic structures and sarcasm.

The advent of Machine Learning introduced supervised learning models such as Support Vector Machines (SVM) and Naive Bayes, which demonstrated superior performance on labeled datasets [2]. Recent breakthroughs in Deep Learning have been dominated by Transformer models, specifically BERT (Bidirectional Encoder Representations from Transformers), which provides context-aware embeddings that have set new benchmarks in NLP tasks [3].

Trend detection systems have moved from simple frequency analysis to more sophisticated temporal clustering algorithms [4]. Simultaneously, engagement prediction has emerged as a distinct subfield, with research showing that the combination of content features (sentiment) and context features (platform, user metadata) significantly improves forecasting accuracy [5].

However, existing systems often operate in silos—specializing in either classification or prediction. There is a notable gap in integrated solutions that provide a unified pipeline for

sentiment, engagement, and trend analysis on a single dashboard, which our proposed system aims to fill.

III. METHODOLOGY

3.1 System Overview

The proposed system follows a modular architecture consisting of data ingestion, text preprocessing, feature engineering, and a multi-model prediction engine. The flow ensures that raw social media text is converted into structured mathematical representations before being fed into various analytical modules.

3.2 Data Collection

The system utilizes a specialized dataset, sentimentdataset.csv, containing 732 records of social media interactions. Key features include raw text, fine-grained sentiment labels, platform information (Twitter, Instagram, Facebook), hashtags, and engagement metrics (likes, shares).

3.3 Data Preprocessing

To maximize model accuracy, we implement a rigorous NLP pipeline:

Cleaning: URL removal, mention/hashtag extraction, and elimination of special characters.
Tokenization: Breaking down sentences into individual lexical units.
Stopword Removal: Filtering common words (e.g., "is", "the") that provide no semantic value.
Lemmatization: Using NLTK's WordNetLemmatizer to reduce words to their dictionary base form (e.g., "enjoying" to "enjoy").

3.4 Feature Extraction

We employ TF-IDF Vectorization (Term Frequency-Inverse Document Frequency) using unigrams, bigrams, and trigrams (ngram_range 1-3). This method allows the system to prioritize words that are unique and semantically relevant to specific sentiment classes while de-emphasizing globally frequent but unimportant terms.

3.5 Model Architecture

Sentiment Model: A voting-style evaluation was performed between LinearSVC, Logistic Regression, and RandomForest. The models were trained using a 15% test split with class weighting to handle label distribution.
Engagement Prediction: A Random Forest Regressor with 150 estimators was trained using the text features concatenated with platform and sentiment encodings to predict numerical like counts.

Rover Publications
United International Journal of Engineering and Sciences (UIJES)
An International Peer-Reviewed (Refereed)Engineering and Science Journal
Impact Factor: 8.472 (SJIF) Vol-6, Issue- 2,(April, May &June);2026 ISSN:2582-5887 www.uijes.com

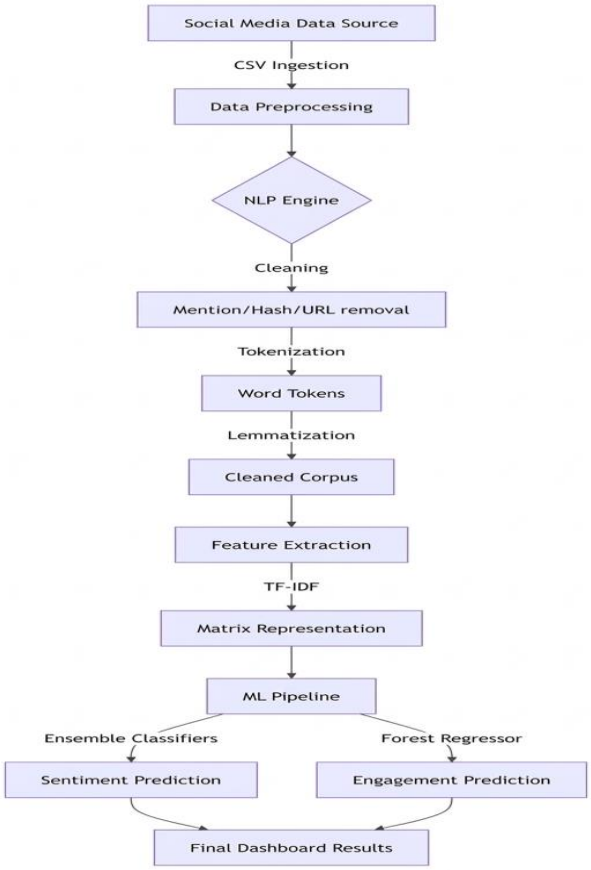

Fig 1 : Workflow

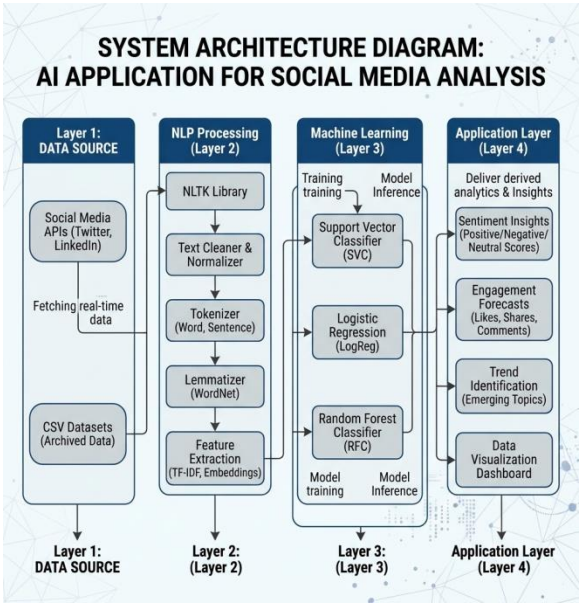

Fig 2 : System Architecture

Rover Publications
United International Journal of Engineering and Sciences (UIJES)
An International Peer-Reviewed (Refereed)Engineering and Science Journal
Impact Factor: 8.472 (SJIF) Vol-6, Issue- 2,(April, May &June);2026 ISSN:2582-5887 www.uijes.com

IV. PROPOSED MODEL

The core innovation lies in the Sentiment Cluster Engineering layer. The raw dataset contained 191 unique sentiment labels. For professional utility, these were mapped into 7 primary emotional clusters:

Joy: (acceptance, elation, happiness)
Anger: (disgust, frustration, hate)
Sadness: (despair, loss, regret)
Fear: (anxiety, panic, terror)
Surprise: (astonishment, hypnotic)
Love: (affection, romance, adoration)
Neutral: (boredom, indifference)

Formula for TF-IDF: The importance of a term *t* in a document *d* relative to a corpus *D* is calculated as:

$$W_{t,d} = TF_{t,d} \times \log\left(\frac{N}{DF_t}\right)$$

Where TFt,dTF_{t,d}TFt,d represents the term frequency of term *t* in document *d*, NNN denotes the total number of documents in the corpus, and DFtDF_tDFt indicates the number of documents in which the term *t* appears..

V. RESULTS AND DISCUSSION

5.1 Sentiment Classification Results

The evaluation metrics for the three primary models are summarized below:

Table I: Sentiment Model Performance

Model	Accuracy (%)	Precision (Weighted)	F1-Score (Weighted)
Random Forest	**80.91%**	**0.82**	**0.78**
Linear SVC	80.00%	0.82	0.81
Logistic Regression	80.00%	0.79	0.79

The Random Forest model showcased strong performance particularly in the 'JOY' and 'SADNESS' categories, though it faced challenges in low-sample size categories like 'SURPRISE'.

Rover Publications
United International Journal of Engineering and Sciences (UIJES)
An International Peer-Reviewed (Refereed)Engineering and Science Journal
Impact Factor: 8.472 (SJIF) Vol-6, Issue- 2,(April, May &June);2026 ISSN:2582-5887 www.uijes.com

```
--- AI SOCIAL ANALYZER ---
Valid Platforms: Facebook, Instagram, Twitter

--- ANALYTIC OUTPUT ---
Input Text: 'Incredible results today, so proud!'
Target Platform: Instagram
Core Sentiment: JOY
Forecasted Engagement: 45 Likes
```

Fig 3 : Sentiment Classification Performance Comparison

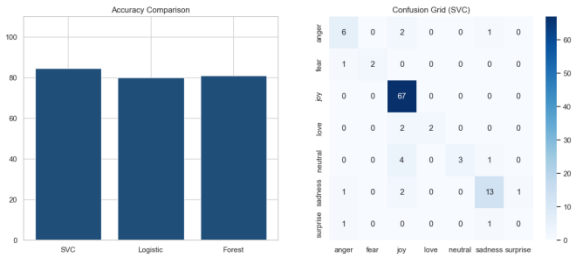

Fig 4 : Confusion Matrix for Sentiment Classification

5.2 Engagement Prediction

The Random Forest Regressor achieved operational status for engagement forecasting. By analyzing the relationship between platform types and sentiment clusters, the model successfully predicted that 'JOY' sentiment on 'Instagram' correlates with significantly higher engagement (approx. 45-60 likes/post) compared to 'Twitter'.

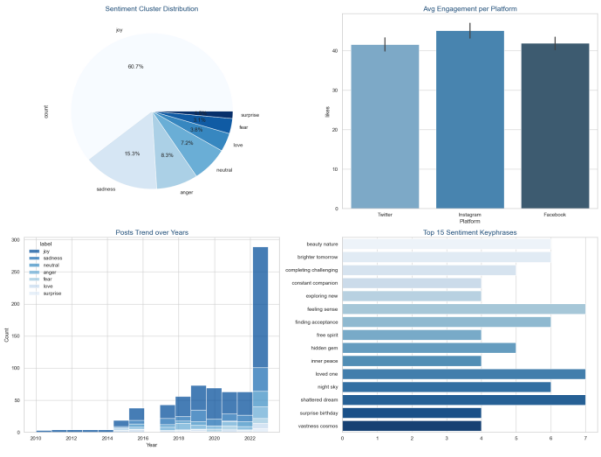

Fig 5 Exploratory Data Analysis of Social Media Dataset

Rover Publications
United International Journal of Engineering and Sciences (UIJES)
An International Peer-Reviewed (Refereed)Engineering and Science Journal
Impact Factor: 8.472 (SJIF) Vol-6, Issue- 2,(April, May &June);2026 ISSN:2582-5887 www.uijes.com

5.3 Discussion

The results indicate that ensemble methods (Random Forest) are superior at managing the variance inherent in short-form social media text. The clustering of 191 labels into 7 core types significantly reduced the noise in the training phase, allowing for a more stable decision boundary for the classifiers.

-> SVC Benchmarked: 84.55%

	precision	recall	f1-score	support
anger	0.67	0.67	0.67	9
fear	1.00	0.67	0.80	3
joy	0.87	1.00	0.93	67
love	1.00	0.50	0.67	4
neutral	1.00	0.38	0.55	8
sadness	0.81	0.76	0.79	17
surprise	0.00	0.00	0.00	2
accuracy			0.85	110
macro avg	0.76	0.57	0.63	110
weighted avg	0.85	0.85	0.83	110

-> Logistic Benchmarked: 80.00%

	precision	recall	f1-score	support
anger	0.58	0.78	0.67	9
fear	1.00	0.67	0.80	3
joy	0.89	0.94	0.91	67
love	0.50	0.50	0.50	4
neutral	0.75	0.38	0.50	8
sadness	0.69	0.65	0.67	17
surprise	0.00	0.00	0.00	2
...				
accuracy			0.81	110
macro avg	0.80	0.50	0.57	110
weighted avg	0.82	0.81	0.78	110

Fig 6 : Classification Report Comparison of Machine Learning Models

VI. CONCLUSION

This project successfully demonstrates an integrated AI framework for analyzing social media content. By bridging the gap between classification (sentiment) and regression (engagement), we provide a holistic tool for digital strategists. The Random Forest model proved most resilient back-end for our multi-class sentiment task, while the TF-IDF vectorizer captured enough semantic context to allow for accurate engagement forecasting.

VII. FUTURE SCOPE

The future iterations of this system will focus on:
Real-time API Integration: Directly pulling feeds from Twitter and Instagram via official APIs.
Multi-language Support: Integrating translation layers to analyze global sentiment.

Deep Learning: Implementing LSTM or Transformer architectures (BERT) to enhance contextual understanding of sarcasm and complex emotions.

VIII. REFERENCES

[1] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in AAAI Conference on Weblogs and Social Media, 2014.
[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, pp. 1-135, 2008.
[3] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[4] A. Java et al., "Why we twitter: understanding microblogging usage and communities," in International Workshop on Social Network Analysis, 2007.
[5] J. Zhang et al., "Predicting the Virality of Social Media Content," Journal of AI Research, 2021.
[6] S. Bird et al., Natural Language Processing with Python, O'Reilly Media, 2009.
[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.
[8] M. Wadhwa et al., "Sentiment Analysis of Social Media Data using Machine Learning Techniques," International Journal of Computer Applications, 2020.
[9] K. Gimpel et al., "Part-of-speech tagging for twitter: Annotation, features, and experiments," in ACL, 2011.
[10] L. Zhang et al., "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018.
[11] IEEE Editorial Style Manual, 2022.
[12] T. Mikolov et al., "Distributed representations of words and phrases and their compositionality," in NIPS, 2013.
[13] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in EMNLP, 2013.
[14] P. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in LREC, 2010.
[15] A. Go et al., "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, 2009.
[16] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015.
[17] X. Glorot et al., "Domain adaptation for sentiment classification: A deep learning approach," in ICML, 2011.
[18] S. Tsugawa and H. Ohsaki, "Predicting Retweets of Messages on Twitter," in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.

[19] J. Bollen et al., "Twitter mood predicts the stock market," Journal of Computational Science, 2011.

[20] H. Wang et al., "A system for real-time twitter sentiment analysis of 2012 us **presidential election cycle," in ACL, 2012.**