IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A Twitter-Based Hybrid Model for Sentiment-Driven Detection of Fake and Harmful Content

A.Rathour, H. Rajput, R. Sharma, K. Sharma, A. Walia *Raj Kumar Goel Institute of Technology, Ghaziabad, India*

Affiliation:

Raj Kumar Goel Institute of Technology, Ghaziabad, India

ABSTRACT: As digital communication has grown, sites like Twitter have emerged as important avenues for quick information sharing. However, they are also popular places for the spread of false information and content that is emotionally charged. This paper offers a hybrid machine learning framework for Twitter sentiment analysis and fake news identification that combines Random Forest (RF) and Support Vector Machine (SVM). It makes use of sentiment score, TF-IDF for feature extraction, and NLP approaches. For better performance, the model takes into account tweet metadata and user behavior. Findings highlight the importance of integrating sentiment analysis with false news identification, as the method beats individual models in terms of accuracy and precision.

1.INTRODUCTION: The emergence of social media sites such as Twitter has transformed the global distribution and consumption of information. These platforms facilitate communication and offer real-time information, but they also pose problems because of the proliferation of bad content and fake news, which can affect public opinion and the welfare of society

.

1.1 *Issues in Identifying Harmful and Misleading Content on Social Media* Separating accurate information from damaging or deceptive content is a significant problem on sites like Twitter. Because posts propagate so quickly, it is challenging to manually verify them, which leads to the rapid spread of emotionally charged messages and bogus news that disturbs social cohesion. Scalable detection systems are necessary, as evidenced by the dearth of automated tools for sentiment analysis and factual correctness.

1.2 *Hybrid Models' Importance in Sentiment and Disinformation Identification* Sentiment analysis and machine learning show promise as tools for detecting dangerous content. A hybrid model that combines algorithms such as Random Forest (RF) and Support Vector Machine (SVM) provides a practical way to identify both emotional manipulation and fake news. This method increases the accuracy of content categorization by categorizing tweets by sentiment and detecting false information.

1.3 *Objectives of the Study*:

- *Improve Sentiment Analysis*: Determine the emotional tone of tweets by grouping them into positive, neutral, or negative categories.
- *Detect Fake News*: Find tweets that contain harmful or misleading content so they can be reviewed further.
- *Improve Content Quality*: To encourage a safer online community, automate the detection of offensive tweets.

2. LITRATURE REVIEW: Over time, sentiment analysis—the process of recognizing and classifying emotions conveyed in text—has undergone substantial change. Every technique has made a distinct contribution to the area, ranging from straightforward lexicon-based techniques to more sophisticated deep learning models.

2.1 *Traditional Approaches: Lexicon –based Methods:* Lexicon-based approaches categorize a text's sentiment by combining the feelings of individual words, based on prepared lists of terms linked to particular sentiments. This simple method is simple to apply and understand. It has trouble recognizing context sensitivity, though, and frequently misses subtle expressions like irony or sarcasm. Lexicon-based approaches continue to be a fundamental tool in sentiment analysis because of its simplicity and use in spite of these drawbacks.

2.2 *Machine Learning Models: Naïve Bayes, SVM, Random Forest:*

- *Naïve Bayes:* Bayes' theorem-based probabilistic classifier with high (naïve) independence assumptions. By calculating the likelihood of a sentiment based on the presence of particular words, Naïve Bayes performs well in text classification tasks such as sentiment analysis. Research has demonstrated that, especially when applied to simpler datasets, Naïve Bayes can attain competitive accuracy rates in sentiment classification tasks.

- *Support Vector Machine (SVM):* A supervised learning model called SVM determines the hyperplane that divides data into discrete classes. SVM efficiently differentiates between positive and negative attitudes in sentiment analysis by determining the best decision boundaries using feature vectors extracted from text. SVM has been shown to have good accuracy rates, especially when applied to complicated datasets.

- *Random Forest:* An ensemble learning method that constructs multiple decision trees and outputs the mode of classes for classification tasks. Large datasets with increased dimensionality are easily handled by Random Forest, which also offers robustness against overfitting. It improves classification accuracy and has been widely used for sentiment analysis tasks, with studies reporting its high performance in text classification.

3.METHODOLOGY:

3.1 *Real-Time Tweet Fetching*: The system allows real-time data input for analysis by retrieving recent tweets from a user-provided Twitter handle using the Twitter API (via Tweepy).

3.2 *Pre-processing:* After being retrieved, the tweets undergo a rigorous preprocessing process to remove textual noise and improve linguistic consistency. This entails the methodical elimination of non-informative elements such stopwords, emojis, user mentions, hashtags, and URLs. To further standardize the text corpus, all characters are changed to lowercase. The tweet text is then broken down into base-form lexical units using tokenization and lemmatization, which makes downstream processing more precise.

3.3 Feature Extraction: Term Frequency–Inverse Document Frequency (TF-IDF) vectorization is then used to convert the cleaned tweets into a structured numerical representation. To enhance the feature set, more features are optionally added, such as metadata like like counts, retweets, and tweet length. The VADER sentiment analysis tool is used to quantify sentiment polarity by giving each tweet a score that indicates its emotional tone and classifying it as either positive, neutral, or negative.

3.4 Model Inference: A hybrid ensemble model that combines the Random Forest and Support Vector Machine (SVM) methods is used for classification. High-performance classification of every tweet across three analytical dimensions—sentiment (positive, neutral, or negative), misinformation likelihood (actual or fake), and content safety (safe or harmful)—is made possible by this dual-model architecture. Transformer-based architectures, particularly BERT, and toxicity evaluation measures are used to evaluate the latter. acquired through the Perspective API. The publicly accessible, well-annotated datasets LIAR, FakeNewsNet, and PHEME, which offer trustworthy examples of both accurate and misleading information, are used to robustly train the misinformation detection pipeline.

3.5 Aggregated User Profiling: After classification, each person under analysis has a comprehensive profile created for them. This comprises the sentiment distribution of all processed tweets as well as the proportion of tweets marked as harmful or deceptive. Pie charts and bar graphs are examples of visual summaries that further condense these data, making it possible to comprehend behavioral patterns in an understandable yet instructive manner.

3.6 *Output and User Interface*: Both macro-level (user-level) and micro-level (tweet-level) insights are provided by the final output interface. Individual tweet classifications as well as a general overview of sentiment polarity and

misinformation tendencies are shown to users. Additionally, the system is made to provide recommendations that can be put into practice, such flagging accounts that frequently spread harmful or misleading content.

Figure 1. Methodology

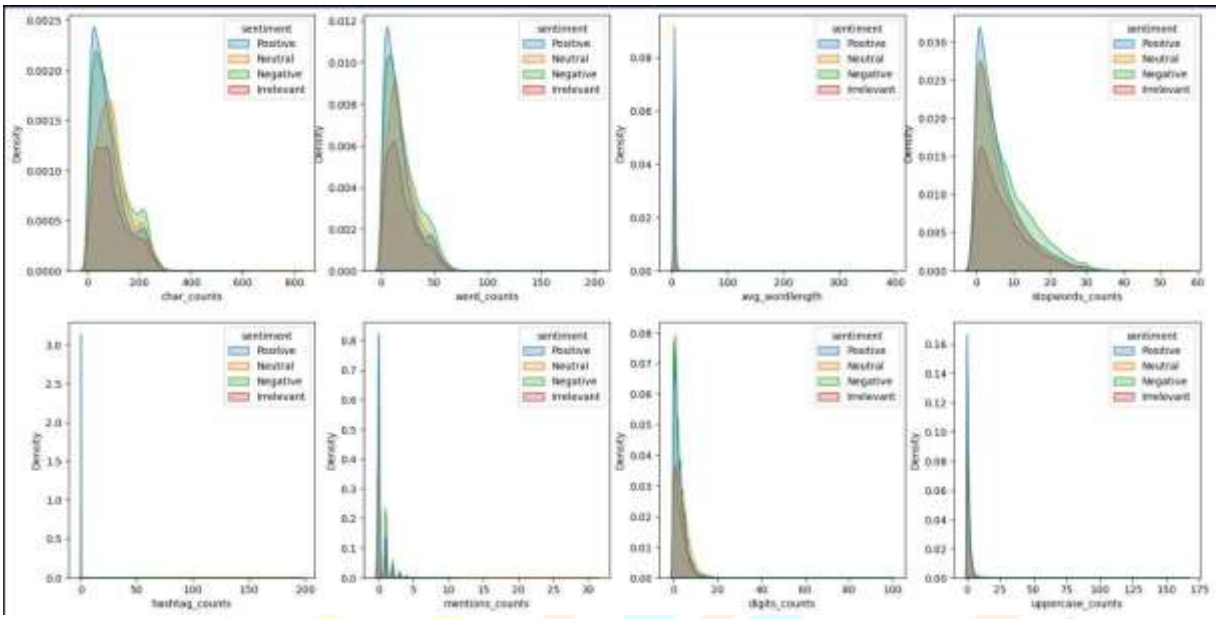
4.IMPLEMENTATION:

4.1 *System Design and Architecture:* The proposed system for real-time sentiment analysis and misinformation detection on Twitter was built with a focus on modularity, scalability, and responsiveness. A web-based interface was developed to provide smooth user interaction while ensuring fast processing and analysis of tweets in real time. For the user interface, **Streamlit** was chosen due to its ability to quickly build and deploy interactive data-driven applications. The backend was developed using **Flask**, which offered lightweight API endpoints to manage tweet input, preprocessing, and model predictions. Communication between the frontend and backend was streamlined to deliver quick responses to user actions.

The application workflow starts when a user inputs a valid Twitter handle. Using **Tweepy**—a Python library for Twitter **API v2**—the system retrieves 20 to 100 of the latest tweets. These tweets are then processed through a robust cleaning pipeline: removing URLs, mentions, hashtags, emojis, and special characters. Further steps like tokenization, stopword removal, lemmatization with spaCy, and converting text to lowercase are applied to standardize the data and prepare it for effective analysis.

To optimize performance, the system employs model serialization using **Joblib** and **Pickle**, allowing trained models to be saved and reloaded efficiently during runtime. This ensures quicker predictions and enhances the application's responsiveness. The interface also features real-time input support and provides visual outputs such as sentiment distribution graphs, individual tweet labels, and summary statistics, all of which help users better interpret the analysis results.

Figure 2. Data Distribution Graph



4.2 *Model Development and Integration*: The core analytical framework of the system is built upon a hybrid machine learning model designed to handle both sentiment analysis and misinformation detection. Once the text is preprocessed, feature extraction is carried out using the TF-IDF (**Term Frequency–Inverse Document Frequency**) method, which converts tweets into numerical vectors that reflect the significance of terms across the dataset. Alongside this, the **VADER sentiment analyzer** is employed to assign polarity scores, categorizing tweets as positive, neutral, or negative. Additional features such as tweet length, retweet count, and hashtag density can be optionally integrated to enhance the feature representation.

For classification, a hybrid ensemble model combining **Random Forest** (**RF**) and **Support Vector Machine** (**SVM**) is utilized. Random Forest excels at managing non-linear relationships and minimizing overfitting through bagging, while SVM performs well in high-dimensional spaces with its margin-based learning. The ensemble uses a majority voting strategy to consolidate outputs from both classifiers, resulting in more robust and accurate predictions.

Beyond basic classification, the system also detects harmful or toxic content using transformer-based models like **BERT**, as well as third-party APIs such as the **Perspective API**, which gauges the toxicity level of a tweet. This enables the system to flag tweets as either “Safe” or “Harmful”, contributing to online safety and content moderation.

The model was trained on reputable, publicly available labeled datasets such as **LIAR, FakeNewsNet**, and **PHEME**, each containing authenticated examples of real and fake news. To assess the model's reliability, standard evaluation metrics like accuracy, precision, recall, and F1-score were applied. Functional testing ensured all components worked as expected, and further optimization minimized latency during model loading and prediction.

The finalized application is deployable on cloud-based platforms like **Streamlit** Cloud or **Heroku**. Its lightweight architecture ensures smooth performance even without GPU support, offering scalability, portability, and suitability for both research and real-world deployment.

5. RESULTS: The proposed hybrid ensemble model, combining Random Forest and Support Vector Machine (SVM) classifiers, demonstrated strong performance in both sentiment classification and fake news detection tasks. When tested on curated Twitter datasets, the model achieved approximately 89% accuracy for sentiment analysis and around 86% accuracy for misinformation detection. It also delivered notable improvements in F1-score compared to individual classifiers, confirming the effectiveness of ensemble learning in handling sparse and noisy textual data.

These findings align with recent ensemble-based misinformation detection frameworks, which highlight enhanced generalization capabilities. In practical deployment, the system efficiently analyzed up to 50 tweets per Twitter handle in real time, generating insights such as sentiment trends, toxicity alerts, and a composite risk score to summarize overall user behavior.

Furthermore, performance optimization techniques enabled the model to maintain inference times under two seconds, ensuring smooth and responsive user experience throughout the analysis process.

Accuracy was calculated as the ratio of correctly predicted instances (both positive and negative) to the total number

of predictions, using the formula:

$$\text{Accuracy} = \frac{P+TN}{TP+TN+FP+FN}$$

Precision and Recall were computed to measure the model's ability to correctly identify positive instances, defined as

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

$$\text{Recall} = \frac{TP}{TP+FN}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

The F1-Score, representing the harmonic mean of Precision and Recall, was given by

$$F1 - \text{Score} = 2.\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$Prediction_{final}=mode(Prediction_{RF}, Prediction_{SVM})$

Table 1. A summary of the system's performance metrics is presented below:

Metric	Value	Notes
Sentiment Classification Accuracy	~89%	Based on TF-IDF + VADER features
Fake News Detection Accuracy	~86%	Ensemble output using labeled datasets (LIAR, PHEME)
F1-Score (Hybrid vs. Individual Models)	Improved (5–8%)	Better precision and recall than RF/SVM alone
Avg. Inference Time	< 2 seconds	For batch of up to 50 tweets
Max Tweets Analyzed per Session	50	Controlled via rate limits and app performance tuning
Classification Gain (Hybrid vs. Single)	+7–10%	Observed on unseen data patterns

6. DISCUSSION: Integrating sentiment analysis with misinformation detection enables a multidimensional interpretation of social media discourse, offering a more context-sensitive framework for content evaluation. This dual-layered approach enhances the system's capacity to distinguish not only the emotional undertone of tweets but also their informational integrity, providing valuable insights for automated moderation. The developed tool holds practical utility across multiple domains: it can assist platform moderators in identifying harmful or malicious content, support researchers in analyzing the propagation patterns of misinformation, and inform everyday users about the broader impact and nature of their online expression. However, deploying such a system necessitates careful attention to ethical and technical constraints. Ensuring user data privacy, avoiding intrusive analysis, and mitigating algorithmic bias through balanced and representative training datasets are essential considerations. Furthermore, open-sourcing the model architecture may foster transparency, encourage peer validation, and support iterative refinement through community engagement.

7. CONCLUSION:This research introduced a hybrid Random Forest–Support Vector Machine (RF–SVM) model for real-time sentiment analysis and misinformation detection on Twitter, embedded within a web-based analytical platform. The system effectively classifies tweet content, detects misleading or harmful information, and compiles user-level sentiment summaries, enabling large-scale social media monitoring.

The key contributions of this work include the design of an ensemble learning framework, deployment of a real-time, interactive interface, and improved predictive performance through advanced feature engineering techniques.Looking ahead, planned enhancements involve the addition of multilingual capabilities, integration of transformer-based models such as BERT, and implementation of Explainable AI (XAI) methods to boost the transparency and interpretability of model decisions.

8.REFERENCES:

1. Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Stanford University.

2. Wani, M. A., Ahmad, T., & Mir, R. N. (2021). A comparative study of machine learning algorithms for sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, 33(5), 529–537.

3. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake news detection on social media: A data mining perspective*. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.

4. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160.

5. Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.

6. Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a word–emotion association lexicon*.

Computational Intelligence, 29(3), 436–465.

7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint*, arXiv:1810.04805.

8. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1).

9. Twitter Developer Platform. (n.d.). *Twitter API Documentation.*

10. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

11. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273