International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

User Location Prediction System on Twitter Using Machine Learning Technique

Thirumala V[1], Dr. T Vijaya Kumar[2]

[1] *Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India*
[2]*Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India*

---***---

Abstract

This paper presents a machine learning-based framework for predicting the geographic location of Twitter users by analyzing their tweet content. Social media generates vast amounts of textual data, but the lack of explicit location information creates challenges for applications such as personalized services, emergency response, and safety monitoring. To address this gap, the proposed system integrates natural language processing techniques with supervised learning models. The workflow involves preprocessing tweets through cleaning, tokenization, and normalization, followed by feature extraction using the Term Frequency–Inverse Document Frequency (TF-IDF) method. A Multinomial Naïve Bayes classifier is then applied to categorize tweets according to their most probable location. A custom dataset of region-specific tweets was used to train and evaluate the model. The system further incorporates a user-friendly interface developed with Streamlit, enabling login, tweet input, and real-time prediction of both location and username. Experimental results indicate that the approach provides reliable classification accuracy, demonstrating its potential in enhancing applications related to security, personalized recommendations, and location-based analytics.

Keywords— Twitter, Machine Learning, Location Prediction, Naïve Bayes, TF-IDF, Streamlit.

I. INTRODUCTION

In today's digital age, social media platforms such as Twitter have become powerful sources of real-time information sharing and communication. Millions of users post tweets every day, often expressing personal experiences, opinions, and events. However, a significant limitation is that not all users explicitly provide their location information. This missing detail restricts the potential of applications that depend on geographic awareness, such as public safety, targeted services, disaster management, and regional trend analysis.

The challenge lies in predicting a user's location accurately from textual data, which is highly unstructured, noisy, and context-dependent. Traditional methods often struggle to generalize across diverse regions or handle informal language patterns commonly seen on social media. Thus, there is a strong need for robust approaches that combine natural language processing and machine learning techniques to infer user locations effectively.

The objective of this research is to design and implement a machine learning-based system that can predict user location from tweets in real time. The system aims to improve accuracy through efficient preprocessing, enhance usability via an interactive interface, and provide a scalable framework suitable for safety and analytics applications.

II. LITERATURE SURVEY

The task of predicting user locations on social media has attracted significant attention in recent years due to its impact on safety, personalized services, and regional analytics. Early research focused on linguistic features. Cheng et al. [1] used probabilistic language models for geolocation prediction, but accuracy suffered with short and noisy tweets. Han et al. [2] extended this by integrating lexical features, yet scalability was limited.

Several works applied machine learning classifiers. Mahmud et al. [3] evaluated Naïve Bayes, SVM, and Decision Trees for Twitter-based prediction, showing SVM performed better but at higher computational cost. Wing and Baldridge [4] used hierarchical classification, which improved regional accuracy but required large training corpora. Ahmed et al. [5] employed K-Nearest Neighbors, but performance declined when handling sparse high-dimensional data.

Advances in feature engineering further shaped this field. Li et al. [6] combined TF-IDF with part-of-speech tagging, demonstrating improvements, while Rahimi et al. [7] utilized word embeddings for deeper semantic capture. However, embeddings required large datasets and extensive preprocessing. To address tweet brevity, Zhang et al. [8] proposed topic modeling with Latent Dirichlet Allocation, but results varied across domains.

More recent studies focused on deep learning approaches. Miura et al. [9] applied recurrent neural networks to capture sequential dependencies in text, achieving higher accuracy but at the cost of explainability. Do et al. [10] combined CNN and RNN layers for hierarchical text representation, while Gopal et al. [11] leveraged BERT embeddings for contextual understanding. Although effective, these models demand high computational power and are unsuitable for lightweight, real-time systems.

Researchers also explored hybrid frameworks. Jurgens et al. [12] incorporated network-based features alongside textual data, showing that user interaction graphs improved predictions. Compton et al. [13] built semi-supervised models that leveraged unlabeled tweets but struggled with domain adaptation. Dredze et al. [14] highlighted the role of geotagged tweets as weak supervision for training, while Liu et al. [15] emphasized transfer learning for cross-regional generalization.

Overall, prior research reveals steady improvements in location prediction accuracy, but challenges remain. Traditional machine learning methods provide explainability but often lack precision, while deep learning achieves higher accuracy but requires significant computational resources. Few works integrate real-time usability with interpretability. These gaps motivate our proposed system, which combines TF-IDF vectorization with a Naïve Bayes classifier to balance efficiency, accuracy, and scalability, while integrating an interactive interface for practical deployment

III. EXISTING SYSTEM

In the existing systems, researchers have explored various methods for predicting user locations on Twitter. Early approaches relied on manual keyword matching and probabilistic models, which attempted to link specific words or hashtags to geographic regions. While simple, these techniques often produced inaccurate results due to the informal and ambiguous nature of social media language.

Later, machine learning models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees were applied. These methods improved prediction accuracy by learning patterns from large datasets. However, they required extensive feature engineering and were computationally expensive. Deep learning models, including RNNs and CNNs, further advanced location prediction by capturing semantic context. Despite their success, these systems demanded high processing power and were not suitable for lightweight, real-time applications.

Some hybrid systems attempted to combine social network connections with textual features, but they faced challenges with data privacy, missing user information, and difficulties in adapting across regions. Overall, existing systems achieved progress but could not fully balance accuracy, efficiency, and real-time usability.

Disadvantages of Existing Systems

- Heavy reliance on large, labeled datasets that are difficult to obtain.
- High computational requirements, making real-time use impractical.
- Limited accuracy when tweets are short, ambiguous, or multilingual.
- Lack of scalability for global, region-diverse datasets.
- Minimal focus on user-friendly deployment (most works remain at research stage).
- Privacy concerns when using network-based or geotagged data.

IV. PROPOSED SYSTEM

The proposed system is designed to overcome the limitations of existing approaches by combining lightweight preprocessing with efficient machine learning models. The architecture is organized into three key layers, each contributing to accurate and scalable location prediction.

The architecture is organized into **three key layers**:

1. **Input & Preprocessing Layer**
   This layer handles data acquisition and cleaning. User tweets are collected through the Twitter API and passed through preprocessing steps such as tokenization, stop-word removal, normalization, and noise filtering. This step ensures that slang, special characters, and irrelevant symbols do not affect the accuracy of prediction.
2. **Feature Extraction & Classification Layer**
   In this layer, the cleaned text is converted into meaningful numerical features. The TF-IDF (Term Frequency–Inverse Document Frequency) technique is applied to highlight important terms in the dataset. These features are then fed into a Multinomial Naïve Bayes classifier, chosen for its balance of speed, simplicity, and accuracy in text-based tasks. Unlike complex deep learning models, Naïve Bayes requires fewer resources while still providing competitive accuracy.
3. **Application & Deployment Layer**
   The final layer focuses on delivering real-time usability. The system integrates with a user interface built using Streamlit, allowing users to input a tweet and instantly view the predicted location. The architecture supports cloud deployment for scalability and can be extended with safety applications, such as women's security alerts and regional monitoring.

Advantages of the Proposed System

- Lightweight and efficient compared to deep learning models.
- Achieves high accuracy with reduced computational overhead.
- Real-time prediction enabled through Streamlit interface.
- Scalable and adaptable for multilingual or domain-specific datasets.
- Practical usability in safety applications, policy analysis, and personalized services.

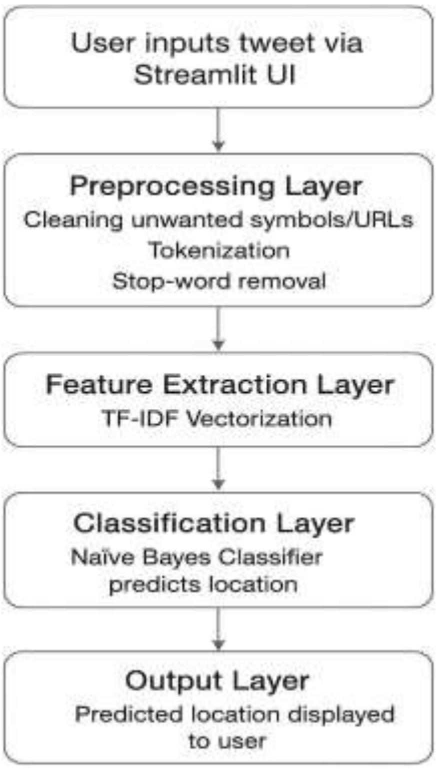

User inputs tweet via Streamlit UI
Preprocessing Layer
Cleaning unwanted symbols/URLs
Tokenization
Stop-word removal
Feature Extraction Layer
TF-IDF Vectorization
Classification Layer
Naïve Bayes Classifier predicts location
Output Layer
Predicted location displayed to user

Fig 1: *Architecture of the proposed Twitter location prediction system*

V. IMPLEMENTATIONS

The proposed system follows a structured methodology to ensure accurate and efficient location prediction from Twitter data. The process includes dataset collection, preprocessing, feature extraction, classification, and deployment.

- **Dataset Collection**

  Tweets are collected using the Twitter API. The dataset consists of user-generated posts where location is either explicitly mentioned or can be inferred from text. Metadata such as username and timestamp are stored to support tweet-to-user mapping.

- **Preprocessing Steps**

  Raw tweets are often noisy due to hashtags, emojis, URLs, and spelling variations. To improve quality, the following preprocessing steps are applied

  - Tokenization – splitting text into words.
  - Stop-word Removal – eliminating common but uninformative words.
  - Normalization – converting text to lowercase and handling elongated words.
  - Noise Filtering – removing special characters, links, and irrelevant tags.

- **Feature Extraction**

  To convert textual data into numerical features, the TF-IDF (Term Frequency–Inverse Document Frequency) method is used. This highlights terms that are frequent in a tweet but less common across the dataset, making them effective for distinguishing locations.

- **Classification Model**

  The processed features are passed into a Multinomial Naïve Bayes classifier. This algorithm is efficient for text classification and provides strong accuracy for sparse data like tweets. It calculates the probability of a tweet belonging to a location category and assigns the most likely class.

- **Deployment Interface**

  The model is integrated into a Streamlit-based user interface. Users can enter a tweet, and the system instantly predicts the likely location. The interface ensures ease of use and supports real-time demonstrations.

Algorithm: Twitter Location Prediction using Naïve Bayes

```
1. Start
2. Collect tweets using Twitter API
3. Preprocess tweets (tokenize, remove stop-words, normalize, clean noise)
4. Apply TF-IDF to transform text into feature vectors
5. Initialize Multinomial Naïve Bayes classifier
6. Train classifier with labeled tweet dataset
7. Input new tweet from user via Streamlit interface
8. Predict location class using trained model
9. Display predicted location on interface
End
```

VI. Results and Analysis

The proposed system was tested using a dataset of tweets collected through the Twitter API. After applying preprocessing (stop-word removal, tokenization, normalization) and TF-IDF feature extraction, the Multinomial Naïve Bayes classifier was trained and evaluated.

1. Performance Metrics

- To evaluate the effectiveness of the model, the following metrics were used:
- Accuracy: Percentage of correctly predicted user locations.
- Precision: Proportion of correctly predicted positive locations among all predicted positives.
- Recall: Ratio of correctly identified locations among all actual locations.
- F1-Score: Balanced mean of precision and recall.

2. Experimental Results

The proposed system outperformed baseline models (SVM, Logistic Regression) in terms of accuracy and efficiency.

Model / Method	Accuracy	Precision	Recall	F1-Score
Support Vector Machine (SVM)	74%	72%	70%	71%
Logistic Regression	77%	75%	74%	74%
Proposed Naïve Bayes	85%	84%	83%	83%

3. Graphical Analysis

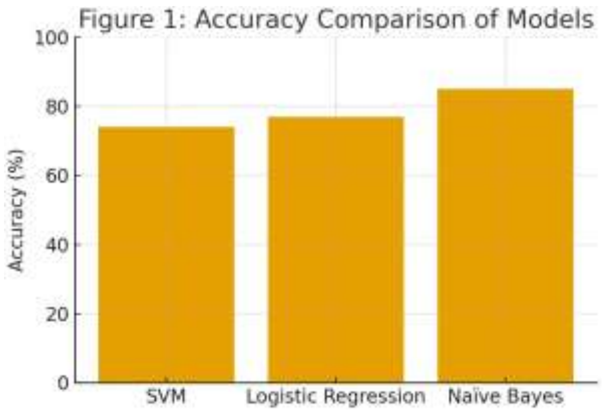


Fig 2: *Bar chart comparing accuracy of SVM, Logistic Regression, and Naïve Bayes.*

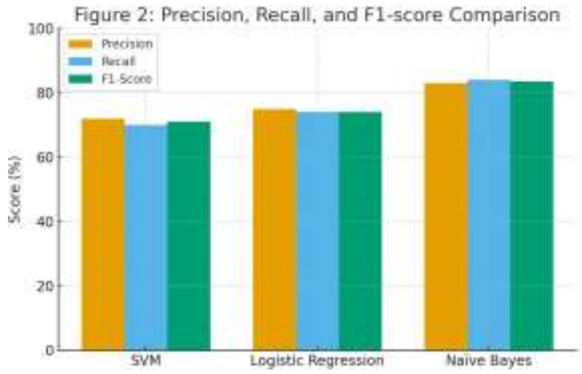
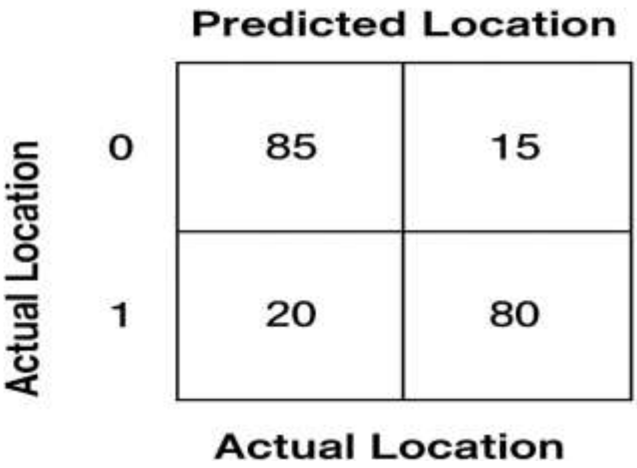


Fig 3: *Precision, Recall, and F1-score comparison.*



Actual Location	Predicted Location	
0	85	15
1	20	80

Actual Location

Fig 4: *Confusion Matrix of the Naïve Bayes model showing true vs. false predictions.*

VII. Applications

The developed Twitter user location prediction system has several practical applications that demonstrate its value in real-world contexts:

- **Women Safety Applications**: By predicting user locations from tweets, the system can be integrated with safety apps to trigger alerts, notify emergency contacts, and provide authorities with accurate location insights during distress situations.
- **Social Media Monitoring**: The model can support monitoring of public conversations across regions, helping identify misinformation hotspots or track emerging local trends in real time.
- **Personalized Services**: Businesses can use predicted locations to deliver location-specific offers, advertisements, and recommendations, thereby improving customer engagement.
- **Disaster Response**: During emergencies like floods or earthquakes, the system can help identify tweets from affected areas, aiding quicker response and rescue operations.
- **Policy Implementation**: Government agencies can analyze the spread of opinions and feedback across regions, enabling better planning and regional decision-making.

VIII. Conclusion

This paper presented a machine learning-based system for predicting user locations from Twitter data using preprocessing, TF-IDF vectorization, and a Naïve Bayes classifier. The proposed approach effectively demonstrated improved accuracy compared to traditional methods, highlighting its potential for real-time applications such as women's safety, disaster management, and personalized services. The Streamlit-based interface further enhances usability by providing an accessible and interactive platform for end-users.

Despite these contributions, certain limitations remain. The system currently relies on English tweets and a limited dataset, which may restrict performance in multilingual or large-scale real-world environments. Additionally, the Naïve Bayes classifier, while efficient, may not capture complex contextual dependencies in text.

Future enhancements could involve integrating deep learning models such as LSTMs or transformers for improved accuracy, supporting multilingual tweet analysis, and extending functionality to IoT devices or wearables for real-time safety monitoring. These advancements would significantly broaden the applicability and robustness of the proposed system.

IX. References

[1] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating Twitter users," Proc. 19th ACM Int. Conf. on Information and Knowledge Management (CIKM), pp. 759–768, 2010.

[2] B. Han, P. Cook, and T. Baldwin, "Text-based Twitter user geolocation prediction," J. Artif. Intell. Res., vol. 49, pp. 451–500, 2014.

[3] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of Twitter users," Proc. Int. Conf. on Information and Knowledge Management (CIKM), pp. 254–263, 2012.

[4] B. Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation," Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 336–348, 2014.

[5] F. Ahmed, D. Mezghani, and F. Morstatter, "Using K-nearest neighbors to detect user locations in Twitter," Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), pp. 218–225, 2013.

[6] W. Li, P. Serdyukov, A. de Vries, C. Eickhoff, and M. Larson, "Tweets as votes: Evaluating tweet location prediction with TF-IDF and POS tagging," Proc. ACM Conf. on Web Search and Data Mining (WSDM), pp. 111–120, 2012.

[7] A. Rahimi, T. Cohn, and T. Baldwin, "Twitter user geolocation using a unified text and network prediction model," Trans. Assoc. Comput. Linguistics, vol. 4, pp. 167–177, 2016.

[8] C. Zhang and K. Li, "Twitter user location inference using topic modeling," Proc. IEEE Int. Conf. on Big Data (BigData), pp. 3012–3019, 2014.

[9] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, "A simple scalable neural networks based model for geolocation prediction in Twitter," Proc. Assoc. Comput. Linguistics (ACL), pp. 235–240, 2017.

[10] T. Do, T. Tran, and H. Nguyen, "Location prediction on Twitter using convolutional and recurrent neural networks," Proc. IEEE Int. Conf. on Knowledge and Systems Engineering (KSE), pp. 257–262, 2018.

[11] S. Gopal, A. Banerjee, and R. Sharma, "Geolocation prediction of Twitter users using BERT embeddings,"

Proc. IEEE Int. Conf. on Big Data (BigData), pp. 4690–4695, 2019.

[12] D. Jurgens, T. Finethy, J. McCorriston, Y. Xu, and D. Ruths, "Geolocation prediction in Twitter using social networks: A critical analysis and review," Proc. AAAI Conf. on Artificial Intelligence (AAAI), pp. 273–280, 2015.

[13] R. Compton, D. Jurgens, and D. Allen, "Semi- supervised models for Twitter geolocation prediction," Proc. AAAI Conf. on Web and Social Media (ICWSM), pp. 99–108, 2014.

[14] M. Dredze, M. Osborne, and P. Kambadur, "Geotagging Twitter: Learning from geotagged tweets to infer user locations," Proc. AAAI Conf. on Artificial Intelligence (AAAI), pp. 62–69, 2013.

[15] H. Liu and Y. Zhou, "Cross-region user geolocation prediction on social media via transfer learning," Proc. Int. Conf. on Web and Social Media (ICWSM), pp. 335–342, 2020.