Sentiment-Based Machine Learning Approach for Mapping Citizen Problems

Dr. Madhu B K[1], Mahanthesha H R[2], Gowtham J V[3], Mohith B N[4], Nadir Durrani[5]

[1] *Dean and Professor, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Karnataka, Mysuru*
[2] *Student, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Karnataka, Mysuru*
[3] *Student, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Karnataka, Mysuru*
[4] *Student, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Karnataka, Mysuru*
[5] *Student, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Karnataka, Mysuru*

ABSTRACT

In today's digital era, social media platforms have become a key medium for citizens to express concerns and grievances. This project proposes a real-time system for detecting and classifying citizen problems from social media posts using a sentiment-based machine learning approach. Due to the massive volume and dynamic nature of user-generated content, manual analysis is impractical. To address this, we leverage machine learning and natural language processing (NLP) techniques to filter spam, detect relevant problems, and classify issues in real time. The system follows an end-to-end pipeline involving data extraction, preprocessing, problem detection, sentiment analysis, and location mapping. Social media APIs and web scraping methods gather real-time data, which is refined through text cleaning, tokenization, and stemming. Supervised learning models detect and classify problems, while NLP techniques like named entity recognition and geoparsing extract location information to map citizen concerns geographically. The proposed system demonstrates the potential to automate public grievance monitoring and provide actionable insights. By combining sentiment analysis, machine learning, and geolocation, it supports faster response to emerging issues, improves public services, and enhances citizen engagement through data-driven decision-making.

Keyword: Sentiment Analysis, Citizen Problem Detection, Machine Learning, Natural Language Processing, Social Media Analytics.

1. INTRODUCTION

India, with its vast population exceeding 1.3 billion people, faces a wide range of challenges, many of which go unheard by the authorities. Despite the government's efforts to address public issues, the sheer magnitude of problems across the nation makes it difficult to prioritize and resolve every concern. This often results in smaller or localized problems being neglected, leaving citizens frustrated and disheartened. The bureaucratic inefficiencies, complex procedures, and lack of transparency further exacerbate the situation, making it increasingly difficult for individuals to bring their issues to the attention of relevant authorities. These barriers lead to a sense of helplessness among the public, as their grievances often fail to be addressed in a timely manner.

In recent years, social media platforms have emerged as a powerful tool for citizens to voice their concerns and highlight their problems. Social media allows individuals to break free from geographical and bureaucratic barriers, offering a direct line to a wider audience, including the authorities. Platforms such as Twitter, Facebook, and

Instagram have become spaces where citizens can share their experiences, tag relevant officials, and use hashtags to amplify their issues. The viral nature of social media ensures that these problems gain momentum and, at times, attract the attention of mainstream media and government bodies.
Additionally, social media serves as a catalyst for collective action. Citizens who resonate with others' problems can unite to demand justice and solutions through mass tagging, online petitions, and hashtag campaigns. The collective pressure generated on social media platforms creates an avenue for authorities to take notice and act upon the issues raised. Given the power of social media in facilitating this communication, it is critical to develop systems that can automatically detect, analyze, and map citizen problems, enabling authorities to respond more efficiently.

2. RELATED WORK

In recent years, various studies have explored the use of social media data to detect and classify citizen problems. Abalı et al. [1] proposed a method for detecting citizen problems and their locations using Twitter data, demonstrating the potential of social media as a source for understanding real-time public concerns. Dongo et al. [2] conducted a comparative study between web scraping and API methods for Twitter data extraction, highlighting the challenges and efficiencies involved in acquiring credible information for analysis.

Sentiment analysis techniques have also been widely researched for real-time event detection and opinion mining. Goel et al. [3] presented a real-time sentiment analysis approach for tweets using the Naive Bayes classifier, showcasing the effectiveness of machine learning models in processing large-scale social media data. Similarly, Wanichayapong et al. [4] and Hasby et al. [5] explored social media-based traffic information extraction and classification, underlining the application of social sensing for urban management and smart cities.

In addition, Guo et al. [7] examined the relationship between happiness and urban factors such as jobs, children, and transportation by analyzing Twitter data, emphasizing the value of social media analytics for understanding citizen well-being. Furthermore, Sakaki et al. [8] demonstrated how Twitter could act as a real-time social sensor for detecting events like earthquakes, paving the way for early warning systems based on public posts. While these studies have laid a strong foundation, most existing approaches focus either on sentiment analysis or event detection without integrating location-based mapping of citizen problems. To address this gap, the proposed system combines sentiment-based classification with real time location extraction, enabling a more comprehensive and actionable understanding of public grievances.

3. PROPOSED METHODOLGY

The proposed system is designed as an end-to-end pipeline to detect, classify, and geographically map citizen problems from social media posts in real time. It integrates data extraction, preprocessing, machine learning-based problem detection, sentiment analysis, and location mapping, ensuring a comprehensive understanding of citizen concerns.

A. Data Collection

Data is collected from social media platforms using APIs and web scraping techniques. APIs allow structured access to user-generated content in real-time, while web scraping supplements the data collection process where API limitations exist. This dual approach ensures a continuous and robust stream of posts for analysis [2].

B. Data Preprocessing

The collected data undergoes preprocessing to enhance its quality for analysis. Noise such as URLs, hashtags, mentions, emojis, and non-textual elements are removed. Tokenization is performed to split the text into individual tokens, and stemming is applied to reduce words to their base form. These steps standardize the text and improve the performance of subsequent machine learning models.

C. Problem Detection and Classification

Machine learning algorithms, particularly supervised learning models, are employed to detect the presence of problems within social media posts. The system is trained on labeled datasets where posts are categorized into

predefined problem types such as infrastructure issues, public services, healthcare, and safety. Feature extraction techniques are used to identify patterns, keywords, and linguistic cues indicative of a citizen-reported problem [1][3].

D. Sentiment Analysis

Sentiment analysis is integrated to assess the emotional tone of posts, helping to prioritize the urgency or severity of detected problems. A Naive Bayes classifier, known for its efficiency in text classification tasks [3], is utilized to determine whether a post reflects positive, negative, or neutral sentiment.

E. Location Extraction and Mapping

Location data extraction is crucial for understanding the geographical distribution of citizen problems. Named Entity Recognition (NER) and geoparsing techniques are applied to identify and extract location names from the text. Posts with valid location information are then mapped to corresponding regions, allowing authorities to visualize the spread of issues across different geographic areas [1][4][5].

4. SYSTEM ARCHITECTURE

The overall architecture of the proposed system is designed to perform real-time detection, classification, and geographic mapping of citizen problems reported on social media platforms. The system follows a modular and scalable design to ensure efficient processing of large volumes of unstructured data and timely extraction of meaningful insights.

The first module is responsible for data acquisition, where social media posts are collected in real-time using official APIs and web scraping techniques. This ensures a continuous stream of user-generated content reflecting emerging citizen concerns.

The second module focuses on data preprocessing, where the raw text is cleaned, tokenized, and normalized through stemming and removal of unwanted elements. This step ensures that the data is standardized and suitable for machine learning analysis.

The third module is dedicated to problem detection and classification. A supervised machine learning model, trained on labeled datasets, identifies posts containing citizen reported problems and classifies them into predefined categories such as public services, healthcare, infrastructure, and safety.

The fourth module integrates sentiment analysis, enabling the system to determine the emotional polarity (positive, negative, or neutral) associated with each problem report. This helps prioritize issues based on the urgency and sentiment intensity reflected in the posts.

The fifth module handles location extraction and mapping. Named Entity Recognition (NER) and geoparsing techniques are employed to identify location references from the post content. The extracted location information is then mapped onto geographic visualizations to display the spatial distribution of citizen problems.

Finally, the system compiles the analyzed results into a structured database and visualization dashboards, offering real-time insights to concerned authorities for proactive decision-making. The modular architecture allows seamless updates and scaling as new social media platforms or new categories of citizen problems are incorporated into the system.

5. EXPERIMENTAL SETUP AND RESULTS

A dataset of approximately 50,000 social media posts was collected using official APIs and supplemented by web scraping techniques. The dataset was manually annotated into various problem categories such as healthcare, infrastructure, public services, and safety. Preprocessing techniques, including text cleaning, tokenization, stopword removal, and stemming, were applied to prepare the data for model training and evaluation.

For problem detection and classification, multiple supervised machine learning algorithms were evaluated, including Naive Bayes, Support Vector Machine (SVM), and Logistic Regression. Among these, the Support Vector Machine

classifier achieved the highest classification accuracy of 88.5%. Sentiment analysis was performed using a Naive Bayes classifier, yielding an accuracy of 85.2% in identifying the sentiment polarity of posts.

Location extraction was successfully implemented using Named Entity Recognition (NER) and geoparsing techniques, with an extraction accuracy of approximately 82%. The extracted location information enabled the visualization of citizen problems on geographic heatmaps, providing insights into region-specific concerns.

To further validate the robustness of the system, cross validation techniques were employed during model evaluation, ensuring that the results were not biased by any specific data partitioning. Hyperparameter tuning was conducted using grid search methodology, optimizing parameters such as the SVM kernel type and regularization strength to achieve peak performance. Additionally, the system's scalability was assessed by simulating increased data loads, where it maintained consistent performance with minimal degradation in classification and extraction accuracies.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	82.3	80.5	81.2	80.8
Support Vector Machine	88.5	87.1	86.8	87.0
Logistic Regression	85.6	84.0	83.7	83.8

Fig -2: Performance comparison of machine learning models

5. CONCLUSION

In this paper, a sentiment-based machine learning approach for detecting, classifying, and mapping citizen problems from social media posts has been proposed and implemented. Given the increasing reliance of citizens on social media platforms to express grievances, the need for an automated, real-time monitoring system is both evident and urgent. The proposed system efficiently captures and processes user-generated data, enabling the detection of significant public issues across various regions.

Through the integration of natural language processing techniques and supervised machine learning algorithms, the system successfully filters spam, identifies problem-related content, classifies issues into predefined categories, and analyzes the sentiment associated with each post. The additional extraction of geographic information further empowers authorities and organizations to visualize the spatial distribution of citizen concerns and prioritize areas requiring immediate intervention.

The results demonstrate that the approach is capable of providing valuable real-time insights into public grievances with high accuracy. This system can serve as a critical tool for governmental agencies, municipal authorities, and non governmental organizations to facilitate prompt, data-driven decision-making and enhance citizen engagement. Future enhancements may include extending the system to incorporate multimedia analysis, multilingual capabilities, and deeper semantic understanding for improved performance.

6. REFERENCES

[1] G. Abalı, E. Karaarslan, A. Hürriyetoğlu, and F. Dalkılıç, "Detecting citizen problems and their locations using twitter data," in Istanbul, Proc. 6th Int. Istanbul Smart Grids and Cities Congress and Fair (ICSG), Turkey, 10.1109/SGCF.2018.8408936.

[2] I. Dongo, Y. Cardinale, A. Aguilera, F. Martinez, Y. Quintero, G. Robayo, and D. Cabeza, "A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility.

[3] A. Goel, J. Gautam, and S. Kumar, "Real time sentiment analysis of tweets using Naive Bayes," in Proc. 2nd Int. Conf. Next Generation Computing Technologies (NGCT), Dehradun, India, 2016, pp. 257–261, doi: 10.1109/NGCT.2016.7877424.

[4] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in Proc. Int. Conf. ITS Telecommunication (ITST), 2011, pp. 107–112.

[5] M. Hasby and M. L. Kodra, "Optimal path finding based on traffic information extraction from Twitter social-based traffic information," in Proc. Int. Conf. ICT for Smart Society (ICISS), 2013, pp. 1–5.

[6] S. B. Marupudi, "Framework for semantic integration and scalable processing of city traffic events," M.Sc. thesis, Wright State Univ., 2016.

[7] W. Guo, N. Gupta, G. Pogrebna, and S. Jarvis, "Understanding happiness in cities using Twitter: Jobs, children and transport," in Proc. IEEE Int. Smart Cities Conf., 2016, pp. 1–7.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in Proc. Int. Conf. World Wide Web (WWW), 2010, pp. 851–860.