International Journal of Computer Applications Technology and Research
Volume 13–Issue 10, 62 – 65, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1310.1005

Comparing Political Inclination Classification on Twitter Posts using Naive Bayes, SVM, and XGBoost

Shashank Shree Neupane
Presidential Graduate School,
Westcliff University
Kathmandu, Nepal

Atish Shakya
Presidential Graduate School,
Westcliff University
Kathmandu, Nepal

Bishan Rokka
Nine Seven Seven Ventures
Pvt Ltd
Kathmandu, Nepal

Sagar Acharya
Presidential Graduate School
Westcliff University
Kathmandu, Nepal

Abstract: For centuries, ideology has been reflected in a person's expression. The expression points out the bias or support the person holds. Nowadays, expressions are well seen on social media in the form of text. X (Formerly Twitter) has become the favoured medium for these expressions. Nepal, a politically highly influenced country where political changes have been frequent in a short period, has people's thoughts expressed on social media. This paper presents a novel approach to finding a person's political inclination through their Nepali tweet using machine learning techniques. By leveraging data pre-processing and XGBoost, we achieve a promising accuracy of 73%. We also discuss potential avenues for further improving accuracy, such as expanding the dataset to include other social media platforms and enhancing data pre-processing techniques.

Keywords: Political inclinations, Twitter data analysis, Machine Learning, Natural Language Processing, Data Preprocessing

1. INTRODUCTION

With the widespread adoption of social media platforms, individuals have gained an unprecedented avenue to express their political views and engage in discussions on socially relevant topics. Analyzing user-generated content on these platforms can provide valuable insights into the public's political inclination and sentiment [1]. In Nepal, a linguistically diverse country, understanding the political leanings of social media users is crucial for policymakers, researchers, and political strategists.

People have always expressed their views to the public. People have expressed themselves in a crowd, on a stage, or in broadcasting media for centuries. With the rapid growth of users on social media like Twitter, people have found a place to lay down their opinions. Several social media have their restrictions, but Twitter has become a way to express views freely. The sentiments of Tweets now understand the sentiment of people. The tweets can also predict the electoral outcomes [2, 3]. Thus, Twitter has become a place where people put their political inclinations in the medium of tweets. Our research contributes to this understanding by providing a method to accurately identify political inclinations from Nepali tweets, which can have significant implications for political analysis and prediction.

1.1 Objective

The main objective of this research paper is to find the best classification model for determining tweets' political inclination and publish the dataset for future researchers to use.

1.2 Related Work

Amador Diaz Lopez et al. [4] analyzed that Twitter data analytics is more effective than telephoning people for election polls. Di Giovanni et al. [5] used content-based classification of Twitter users' political inclinations in their tweets. Their analysis proved that the writing patterns of political persons of the same political parties are similar.

Another study examines how social media, especially Twitter, affects political polarization. The paper by Kim & Hong [6] focuses on how politicians' extreme views attract more followers on Twitter even after considering other factors. The research involves the 111th U.S. House of Representatives members and checks their demographics and social media use. The findings show that politicians with strong left or correct views tend to have more Twitter followers than those with moderate views. This suggests that Twitter can reinforce political divisions. The study also finds that extreme politicians get much attention in traditional media [6]. Even when looking at how often politicians are mentioned in newspapers, the Twitter divide remains clear. The research shows how social media, especially Twitter, plays a big role in shaping political opinions and divisions today.

The study done by Jia et al. [7] utilizes a custom web crawler to collect Twitter data, ensuring efficient retrieval of tweets that meet specific criteria and circumventing limitations imposed by the Twitter API. The research aims to improve stakeholders' understanding of public opinion and foster increased participation in transportation management by developing a robust framework for extracting and analyzing public sentiments about transportation services from Twitter.

Though the Nepali corpus does not have many papers, many researchers are working on it, and several papers are out. Chaudhary [8] shows that TFIDF is better than the Continuous Bag of Words (CBOW) and the Skip-Gram Model. Shahi & Pant [9] demonstrated that on Naıve Bayes, Support Vector Machine (SVM), and Backpropagation Multilayer perceptron with stochastic gradient descent optimization, the SVM with

RBF kernel got the highest accuracy of 74.65 in news text classification using stop words removal and lemmatization.

2. MATERIAL AND METHODS

2.1 Data Collection

The dataset for this study comprises Twitter data obtained from the accounts of members of the House of Representatives in Nepal. Initially, the names of the 275 members and their respective party affiliations were accurately gathered from the official website of the Nepalese Parliament [10] and meticulously organized into a Google Sheet. It resulted in a total of 275 members with distribution across various political parties presented in Table 1:

Table 1. Political Parties and their MPs

Party	MPs	Twitter Account found	Twitter Account used
NC	**88**	**17**	**Yes**
CPN (UML)	**77**	**7**	**Yes**
CPN (MC)	30	2	No
RSP	**21**	**12**	**Yes**
RPP	14	1	No
PSPN	12	0	No
CPN (US)	10	3	No
JANAMAT	6	2	No
LSP-N	4	0	No
NWPP	1	0	No
RJM	1	0	No

2.1.1 Twitter Data Extraction

Websites such as Twitter usually grant programmatic access to their data through APIs. This is the primary method for collecting online data in a structured format. Associating each member with their official Twitter account involves a two-step approach. Firstly, tools such as Google Chrome extensions, Twitter Exporter [11], and Twitter Scraper [12] were utilized to extract data directly from Twitter. However, browser extensions can introduce biases due to their reliance on website structure. Therefore, as emphasized by Barchard & Verenikina [13], a subsequent manual verification was conducted for each identified profile to ensure accuracy and completeness. Due to the nature of the extraction process, not all member profiles were successfully located. The data is made public for future researchers [14].

2.1.2 Limitations of Data Extraction

Despite efforts to systematically extract Twitter data from the accounts of members of the House of Representatives in Nepal, several limitations were encountered during the data collection process. Firstly, using automated tools for data extraction posed challenges due to constraints imposed by the Twitter API. One tool restricted data retrieval to the latest 100 tweets at a time for free, while another provided a limited date range for extraction. One extension allowed the retrieval of only the last 100 tweets, while the other extension utilized a query-based search in Twitter to retrieve tweets from a specific date range.

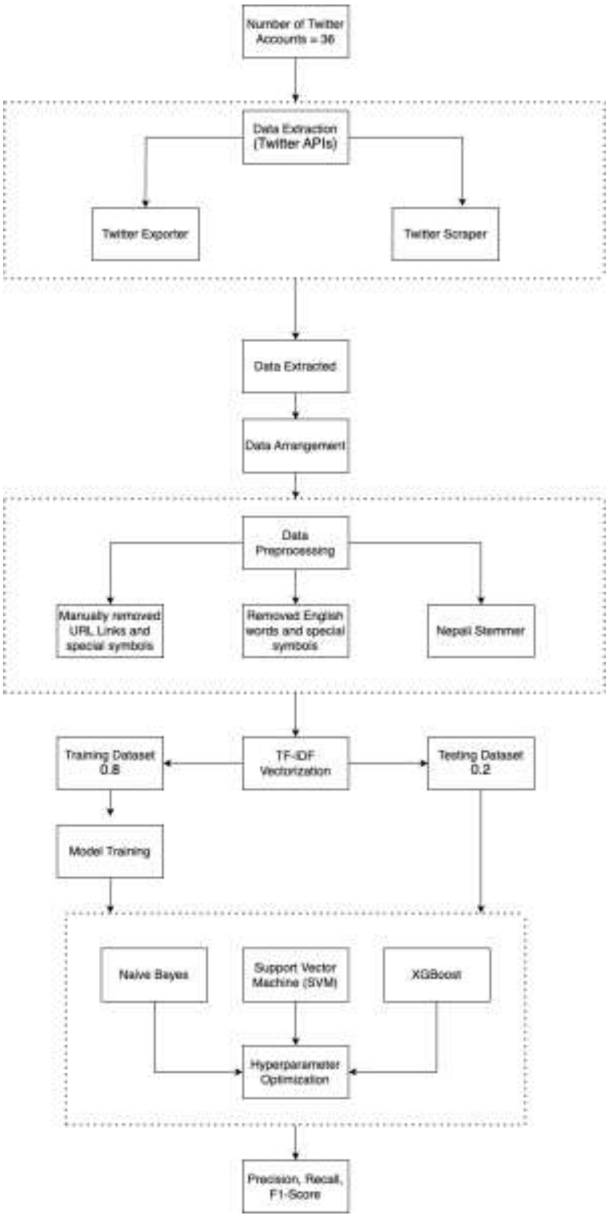
Number of Twitter Accounts = 36
Data Extraction (Twitter APIs)
Twitter Exporter
Twitter Scraper
Data Extracted
Data Arrangement
Data Preprocessing
Manually removed URL Links and special symbols
Removed English words and special symbols
Nepali Stemmer
Training Dataset 0.8
TF-IDF Vectorization
Testing Dataset 0.2
Model Training
Naive Bayes
Support Vector Machine (SVM)
XGBoost
Hyperparameter Optimization
Precision, Recall, F1-Score

Figure. 1 Methodology Chart

2.1.3 Arranging the Twitter Data

After extraction, the data was organized into folders based on party affiliation for ease of access and analysis for each identified Twitter account. However, no cleaning or preprocessing of the data was performed at this stage. Each tweet and associated metadata were saved as a separate file within its corresponding folder, with the file name being the parliament member's name. This organization facilitated subsequent analysis and processing of the Twitter data. Furthermore, Twitter's API rate limits, which restrict the number of requests per 24 hours to 300 and the volume of tweets that can be retrieved per month to 1500, posed significant constraints on data collection [15]. Additional accounts had to be created for scraping purposes. This was necessary to circumvent the imposed limits on API requests and tweet retrieval, thus ensuring a more comprehensive dataset for analysis.

2.2 Data Preprocessing

The data collected through scraping contains URL links, punctuations, retweets, image links, and text. Since the dataset was small, we manually removed the retweets. We only require text to train our model. Thus, we remove URL Links and special symbols. The writer may have also written in English, but our scope is defined only in Nepali words, so we removed the English words from the document. Also, we used Nepali Stemmer to remove the proposition to the ends so that it can be removed from the corpus and work as lemmatization.

We have chosen TF-IDF for Vectorization as it surpasses several other vectorization techniques in a less dataset environment, as shown by [8]. We define the vector's label to 1 if the account belongs to NC or CPN-UML and 0 if the account belongs to any other party. We do this to check the effect on a binary classifier, where the classifier can learn the text differences between the major party and the non-major party.

2.3 Model Training

With the processed text converted into vector representation using TF-IDF, the dataset was divided into training and testing sets using an 80-20 split. We used Machine Learning algorithms to train the model, including Naive Bayes, Support Vector Machine (SVM), and XGBoost. The XGBoost algorithm, due to its effectiveness in handling structured data and robustness against overfitting, was employed to build the final classification model. Hyperparameters for the XGBoost model were optimized through a grid search to achieve the best performance.

3. RESULTS AND DISCUSSION

The model trained on the preprocessed dataset achieved promising results, with the XGBoost classifier yielding the highest accuracy of 73.

Despite the encouraging results, the study faced several challenges. The primary challenge was the limited availability of Twitter data from the accounts of members of the House of Representatives in Nepal. Furthermore, the preprocessing steps, though effective in cleaning the data, resulted in a loss of information, which might have affected the model's performance.

Future work should focus on expanding the dataset to include tweets from other social media platforms and further refining the preprocessing techniques. Additionally, exploring the impact of incorporating more advanced Natural Language Processing (NLP) techniques, such as word embedding and deep learning models, could significantly improve the accuracy of political inclination classification.

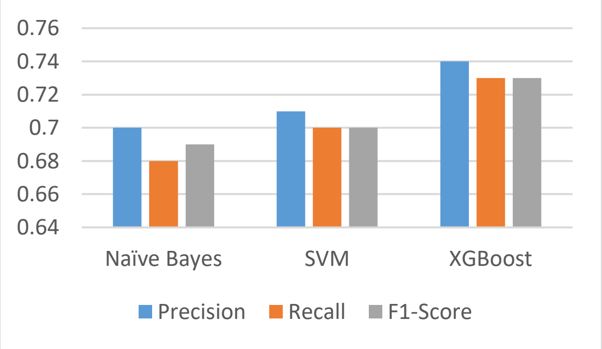

Figure. 2 Performance of Different Models

4. CONCLUSIONS

In this paper, we presented a method to classify the political inclination of individuals based on their Nepali tweets using machine learning models. The proposed approach utilized data preprocessing techniques and the XGBoost algorithm to achieve a promising accuracy of 73%. Our findings underscore the potential of using social media data for political analysis in Nepal and highlight the importance of refining data collection and preprocessing methods to enhance model performance. Future research should aim to address the limitations identified in this study and explore more advanced NLP techniques to improve classification accuracy further.

5. ACKNOWLEDGMENTS

We thank the Presidential Graduate School for providing us with the opportunities and space to conduct the research.

6. REFERENCES

[1] Ansari, F., Farooqi, M., Biyani, P., & Chourasia, S. (2020). Political inclination detection of Twitter users using tweet features. In Proceedings of the International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-5).

[2] Khatua, A., Khatua, A., & Cambria, E. (2020). A tale of two epidemics: Contextual Word2Vec for classifying Twitter streams during outbreaks. Information Processing & Management, 57(1), 102137.

[3] Rodriguez-Rodriguez, I., Marin-Caballero, A., & Garcia-Mendez, S. (2021). Sentiment analysis on Twitter for predicting political preferences. Applied Sciences, 11(4), 1854.

[4] Lopez, A. D., Sun, W., & Bajracharya, S. K. (2017). Predicting elections from Twitter: A content analysis-based approach. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data) (pp. 3513-3518).5

[5] Giovanni, M. D., Stilo, G., & Velardi, P. (2018). Content-based classification of Twitter users' political inclinations. Social Network Analysis and Mining, 8(1), 1-15.

[6] Kim, Y. M., & Hong, S. (2016). Political polarization on Twitter: Implications for the use of social media in digital governments. Government Information Quarterly, 33(3), 273-282.

[7] Jia, Y., Zhao, J., Zhou, Y., & Li, W. (2020). A framework for extracting and analyzing public opinions of transportation services from Twitter data. Transportation Research Part C: Emerging Technologies, 116, 102641.

[8] Chaudhary, A. (2019). Comparison of text classification techniques in Nepali news. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(2), 338-345.

[9] Shahi, T., & Pant, A. (2018). Classification of Nepali news text using Naïve Bayes, SVM and backpropagation multilayer perceptron with stochastic gradient descent optimization. In Proceedings of the International Conference on Computing, Communication and Automation (ICCCA) (pp. 1-6).

[10] Nepalese Parliament Official Website. (n.d.). Retrieved from https://www.parliament.gov.np/

[11] Extensionsfox. (n.d.). Twitter Exporter. Retrieved from https://www.extensionsfox.com/twitter-exporter/

[12] Nighthustle. (n.d.). Twitter Scraper. Retrieved from https://nighthustle.com/twitter-scraper/

[13] Barchard, K. A., & Verenikina, I. (2013). Improving data quality: Coding, reliability, and validity. In K. A. Barchard, I. Verenikina, & K. Weingarten (Eds.),

Research Methods: The Essential Knowledge Base (pp. 139-160). Pearson.

[14] iigal, “GitHub - iigal/political-tweets-dataset: The dataset of the twitter accounts of politically inclined people of Nepal,” *GitHub*, 2024. https://github.com/iigal/political-tweets-dataset.

[15] Platform. (n.d.). Twitter API Rate Limits. Retrieved from https:// developer.twitter.com/en/docs/twitter-api/rate-limits