JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

Pakistan's Political Sentiments Analysis based on Twitter Using Machine Learning

Arfan Ali Nagra[1], Muahmmad Abubakar[1*], Syeda Urwa Warsi[1], Saba Mohsin[1], and Hadi Abdullah[1]

[1]Department of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan.
Corresponding Author: Muhammad Abubakar. Email: mabubakarqazi@gmail.com

Received: November 18, 2023 **Accepted:** January 26, 2024 **Published:** Marcg 01, 2024

Abstract: Pakistan's Politics has always been a hot topic. Politicians wastes no time to express their sentiments and narratives through social media without knowing the whole situation. These social media posts and tweets create a situation which can lead to anarchy in the society. It is a need of hour to figure out which politician creates the more dramatic situation. It can be done through sentimental analysis of their social media account. Sentiment analysis aims to extract sentiments, opinions, and emotions from social media sites like Twitter. The conventional technique of sentiment analysis is concerned with textual data in which users post updates relating to various themes. This manuscript examines five Pakistani politicians consisting of Ex-Prime Minister Imran Khan, Vice President PML-N Marium Nawaz, Chairman PPP Bilawal Bhutto, Spoke-person PTI Shebaz Gill, and Information Minister PML-N Mariem Aurangzeb. The focus is to analyze how they have utilized Twitter to interact with their followers and, in doing so, influence the political process. Data are collected from Twitter accounts belonging to various Pakistani politicians. A comprehensive framework of pre-processing procedures for making tweets more manageable is presented. The main goal is to make sure that people get knowledge about the better direction for the society. In political sentiment analysis, each politician's tweets are classified into positive, negative, and neutral sentiments to provide a unique perspective on how hate speech is used by politicians. This is accomplished with a machine learning classifier i.e Support Vector Machine, Random Forest, Decision Tree and Logistic Regression. After comparative study of these classifiers, Random Forest achieved the highest accuracy 86% among all. Such classifier will aid organizations, political parties, analysts, and others in assessing public sentiments regarding them. As a result, the most negative tweets are used by Imran Khan.

Keywords: Sentiment Analysis; Supervised Machine Learning; Tweets; Naive Bayes Analyzer; Classification.

1. Introduction

As the world becomes increasingly connected, the way we communicate is evolving. Hate speech is a form of communication intended to injure, degrade, or intimidate an individual or group based on their membership in a certain social category. In many cases, hate speech is directed at groups that are already marginalized or disadvantaged in society [1]. Politicians often use hate speech to score points with their base or rally support for their policies. However, this can have dangerous consequences, as it can further marginalize and dehumanize the groups that are being targeted [2]. Hence, this method may be automated in order to reduce the time and effort it takes. Machine learning (ML) is a crucial element in this scenario. The method of Sentiment Analysis (SA) being classified as machine learning aids the program in determining the attitude of a particular statement. The system is based on a number of ML algorithms that may effectively identify sentiment in text or a set of text.

However, it is simply not feasible for a single individual to go through each review. It would be a waste of time. As a result, it may be automated to make the process easier. Machine learning (ML) is an

important element here. Analyzing a statement's emotional tone is referred to as "sentiment analysis" [3]. Many ML algorithms are employed to create the system, which can identify sentiment in words or a set of them. Methods of ML have won out over knowledge- and dictionary-based approaches in research to determine polarity [4].

The proposed methodology claims a technique for extracting Twitter data of different Pakistani politicians including Ex-Prime Minister Imran Khan, Vice President PML-N Marium Nawaz, Chairman PPP Bilawal Bhutto, Spoke-person PTI Shebaz Gill, and Information Minister PML-N Mariem Aurangzeb been scraped for Sentiment Analysis. That data is then stored in a data frame after being corrected and pre-processed which includes tokenization, bag of words, and word cloud etc. It is filtered and pre-processed on Twitter, one of the most popular sources with a huge amount of data. Machine learning classifiers Support vector machine, random forest, decision tree and logistic regression are used to compare analyse sentiment on which politician used more hate speech on twitter categories as positive, negative and neutral to achieve exact results in terms of high-performance evaluation parameters random forest achieved the best accuracy 86%. Our methods provide a unique perspective on how hate speech is used by politicians and can be used better to understand the impact of this type of rhetoric.

The rest of the article is organized as follows: Literature review is described in section 2. Section 3 presents the proposed methodology. Experimental analysis is made in Section 4. Finally, the Result is drawn in Section 5 and Conclusion & Future work is discussed in Section 6.

2. Literature Review

Sentiment analysis is the evaluation of sentiments and narratives that occur in the form of posts on social media. Many techniques were proposed in this regard; but with different methodologies.

Lee et al. [5] proposed a system for sentiment analysis to check instance situation when Taliban took over Afghanistan. As the result it concluded that public of Afghanistan was satisfied with current government. The system showed 97% accuracy. The paper [6] evaluated the performance of SVM and MaxEnt for analyzing hashtags used in tweets. After classification, MaxEnt showed better accuracy as compared to SVM. Lee et al. [7] suggested a system to highlight such tweets which contain racist text. It was done by hybrid model GCR-NN which showed 98% accuracy. In [8], Ishu Gupta et al. proposed LSTM for predicting future prices of stock efficiently. The accuracy of this model increased up to 5% as compared to existing systems. A system was proposed to analyze tweets to find attitude of public towards different products. It combined SVM and Universal language model fine tuning which showed 99.78% accuracy [9,10]. Chetanpal et al. [11] designed a model to analyze tweets related to COVID-19. It utilized LSTM-RNN based algorithm which showed 20% improvement in accuracy as compared to existing systems. Anisha et al. [12] demonstrated a model in which LSTM showed 98.7% accuracy for spam detection and 73% accuracy in sentiments analysis. A system was presented to determine public opinion regarding different events, companies or products. It was done by hybrid model (KNN plus naïve base) which enhanced system's accuracy as compared to current systems [13,14]. Nikhil et al. [15] proposed a model to estimate the performance of algorithms while sentiment analysis. As the result, SVC gave highest accuracy of 83.7%. Akshi et al. [16] introduced a system which utilized two different methods i.e. corpus based and dictionary based for understanding semantics of tweet data. A system was presented which practiced CA (Contextual Analysis) to identify SML model performances. As result, MNB (Multinomial Naive Bayes) and Random Forest showed highest accuracy [17]. In [18] paper, A custom RoBERTa Q&A model was proposed for sentiment analysis which showed Jaccard score of 78%. Shihab et al. [19] proposed a model which evaluate the performance of different algorithms. As the result, Decision Tree outperformed all other algorithm; giving highest accuracy of 98.1%. In [20] paper, the uncommon hyperbole features were used in proposed system for detecting sarcasm. This model gave average 75% accuracy. Andry Chowanda et al. [21] compared different algorithms; as the result Generalized Linear model outperformed all other algorithms by giving accuracy of 90.02%.

All above contributions have different merits and demerits. The main contribution of this research is to analyze the political sentiments of different Pakistani Politicians which include Imran Khan, Bilawal Bhutto, Maryam Nawaz, Maryam Aurangzeb and Shahbaz Gill.

3. Proposed Methodology

This section delves into the comprehensive workflow of the proposed system. The intension behind this work is to analyze tweets to see which politician spread more negativity among masses. Figure 1 shows proposed model diagram which includes detail flow from dataset gathering till results of the experiments.

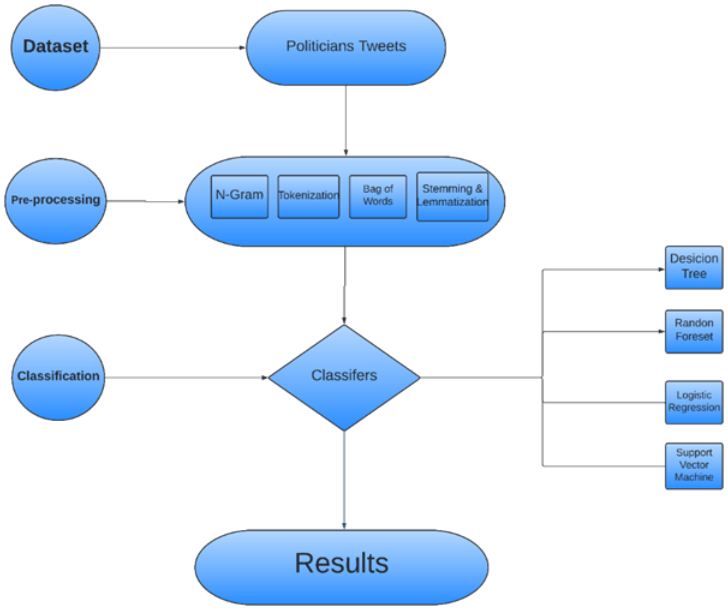

Figure 1. Proposed Model

3.1 DataSet

3.1.1. Data Gathering

In the proposed model, first step is Data gathering. Data gathering is done by analyzing five politician's Tweeter accounts; which are of Imran Khan, Maryam Nawaz, Maryam Aurangzeb, Bilawal Bhutto and Shahbaz Gill. This article focusses only those tweets which are written in English Language. This tweet dataset then divided into train and test datasets. Train dataset is used for training purpose whereas; testing dataset is used to analyze the performance of proposed model. In this way, model is capable for classification of tweets where are fetched via snscrape. Snscrape is a python library used for scraping the tweets. Figure 2 shows tweet's count, where x-axis represents Politician's names and y axis represents total count.

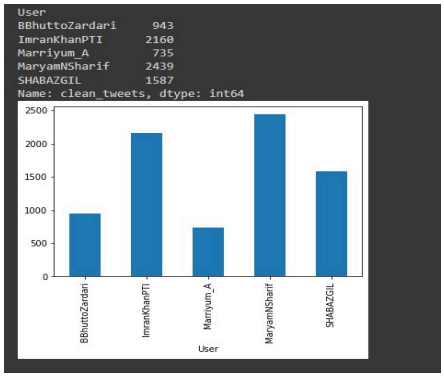

Figure 2. Total Tweets count

3.1.2. Data Preprocessing

Raw tweets which are scraped from twitter contain noise. Noisy data can affect the performance of model. So this raw dataset should be converted into standard form which can be trained using classifier. Different preprocessing steps are followed to standardize tweet dataset.

- *Remove Hashtags*

Different keywords are followed by hashtag symbol. (#) is used by people when they are concerning about something in particular. Such as a tweet about food may contain hashtags like #healthyfood, #Foodlover etc. These hashtags are removed from dataset.

- *Remove URL*

People add different hyperlinks in their tweets, which direct to different pages. URL is in particular format of (https. //www.). These URLs are removed from dataset.

- *Remove Mentions*

People make the use of mentions when they are talking about a person in particular. The format of mention is @username. The proposed model completely removes these mentions if used in tweet dataset.

- *Remove Punctuations*

Punctuation marks like (?,.”; etc) are removed from dataset in order to achieve only required content.

- *Lowercase Conversion*

Tweet dataset contains a combination of lowercase and uppercase words. This model converts whole dataset to lowercase in order to standardize it.

- *Stemming and Lemmatization*

Both stemming and lemmatization are used in proposed model for normalization of tweet data. The purpose of these algorithms is to convert multiple same context word to relevant single word.

- *Removing Stop words*

For instance, if there is a statement ‘I will do protest’; after removal of stop words the statement looks like ‘do protest’. Where I and will act as stop words.

- *Tokenization*

Tokenization is done in preprocessing for sake to break down large content into small parts which are often known as tokens.

- *Bag of words*

Bag of words represents existence of particular text within dataset. Figure 3 shows bag of negative words.

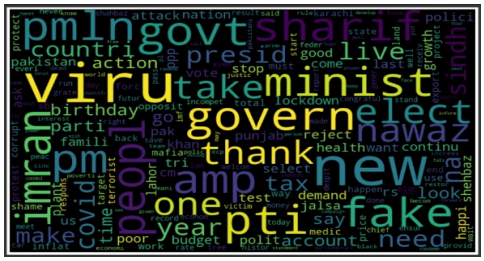

Figure 3. Bag of negative words

- *N-gram*

N-gram models are used in text mining and NLP for analysis purposes. It indicates the collection of items exists frequently in dataset. Items can be in the form of numbers, symbols, word or any sort of punctuation mark. As the dataset is in the form of tweets, here n is the number of words.

Bi-gram is a form of n-gram where n=2. It is a sequence of two successive words exist in dataset. Figure 4 shows bi-gram of used words, where x-axis represents number of occurrence and y-axis represents frequently used words.

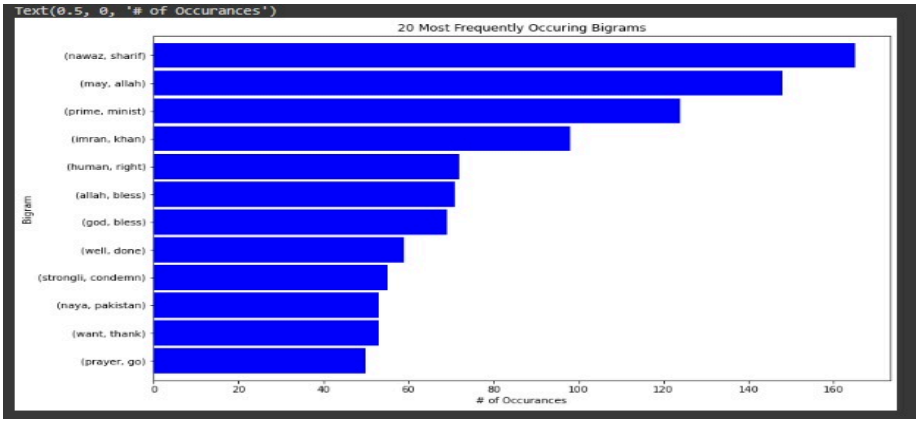

Figure 4. Bi-gram of frequently used words

Tri-gram is a form of n-gram where n=3. It is a sequence of three successive words exist in dataset. Figure 5 shows tri-gram of twenty frequently used words in dataset, where x-axis represents number of occurrence and y-axis represents frequently used words.

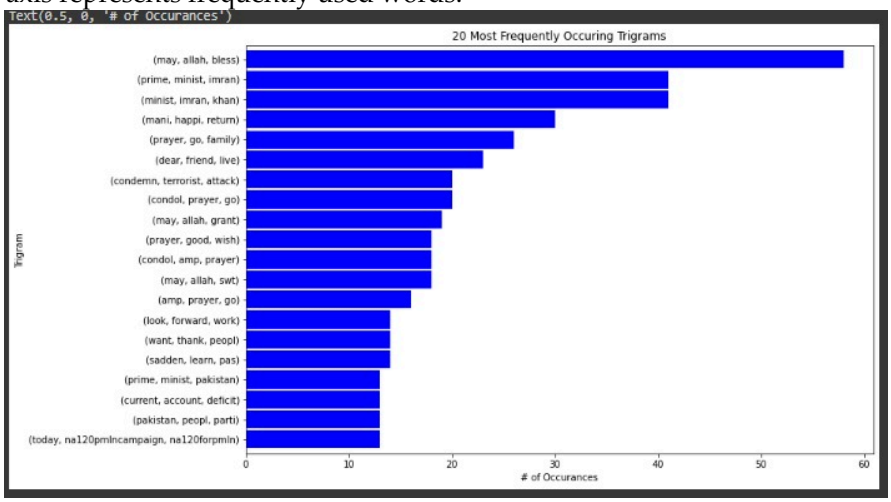

Figure 5. Tri-gram of frequently used words

3.2 Classification

For classification purpose, different classifiers are proposed such as Decision tree, Random forest, Logistic Regression and Support Vector Machine.

3.2.1. Logistic Regression:

Logistic Regression is also known as logit model. It is a particular case of linear regression which evaluates the probability. The output of logistic regression is probability which is in range of 0 to 1. It manufactures sigmoid (a S-shaped curve).

$$Hypothesis: Z = WX + B \tag{1}$$

$$h\,\theta\,(x) = sigmoid(Z) \tag{2}$$

The target variable in logistic regression is of binary nature. It makes the use of cross- validation estimator.

3.2.2. Support Vector Machine:

SVM is mostly known as support vector machine. It is not a probabilistic based classifier. When the space for features is massive, this algorithm works well for content classification. The goal is to find maximum-margin hyperplane which separates the set of points (xi, yi); where yi is the target class and xi is the feature vector. The hyperplane equation is as follows:

$$w.x - b = 0 \tag{3}$$

Here, w is form of vector which acts normal to hyperplane and b is bais. The Sklearn package provides SVM support. It works with different n-grams like unigram and bi-gram.

3.2.3. Decision Tree

It is a classification model in which each node tends to test the attributes used in dataset and the offspring are concerned to outcome of classifier. The last class of dataset is represented by leaf node. It belongs to supervised learning classifier which makes the use of labeled data for preparing decision tree. The decision must be taken for each node, according to P. For ideal split selection, use GINI factor. Where (|) denotes the general recurrence (Relative Frequency) of class j at node t for a given hub (node).

$$GINI(t) = 1 - \sum [\,P\,(\,j \mid t\,)] \tag{4}$$

The scikit-learn provides sklearn.tree packages.

3.2.4. Random Forest:

Random forest is also known as random decision forest. It is ensembling based learning model which is used for both classification and regression purposes. In case of classification, the outcome is the class which is frequently selected. For a huge number of tweets with the specific assessment marks (sentiment labels) x_1, x_2,... x_n, loading y_1, y_2,... y_n continuously chooses a random instance (X_b, Y_b) with replacements. Every configuration tree fb is set to use a different random sample (X_b, Y_b), where b is a value between 0 and 1 and B. Lastly, a decisive vote has been cast upon those B trees' forecasts. Applying Scikit-Random Forest Classifier Learn's of sklearn.ensemble, one can construct a random forest method.

3.3. Results

The outcome of classifiers is following three classes:

- Positive class
- Negative Class
- Neutral Class

The performance of classifiers is evaluated in terms of classification. As this manuscript is concerned with negative statements given by politicians, it helps in analyzing which politician use more negative words.

4. Experimental Analysis

System specification includes Windows 10 of 64-bits, Processor model is Intel(R) Core (TM) i5-6200U CPU @ 2.30GHz 2.40 GHz. RAM used for implementing proposed model is 8GB. Google Colab is the tool which is used for performing the experiment. The performance of different classifiers is evaluated on basis of confusion matrices.

4.1. Decision Tree

Decision Tree is deployed on dataset for classification purposes. Figure 6 shows confusion matrix of decision tree.

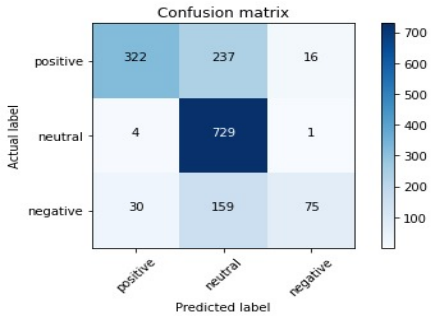

Figure 6. Confusion Matrix of Decision Tree

4.2. Logistic Regression

Classification of tweet dataset is done using Logistic Regression as well. Figure 7 represents confusion matrix of Logistic Regression.

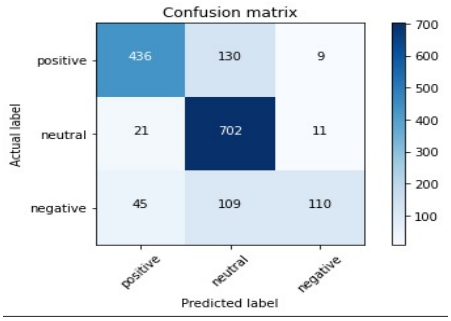

Figure 7. Confusion Matrix of Logistic Regression

4.3. SVM

Support Vector Machine is used to classify dataset into three classes. In Figure 8, confusion matrix indicates performance of SVM.

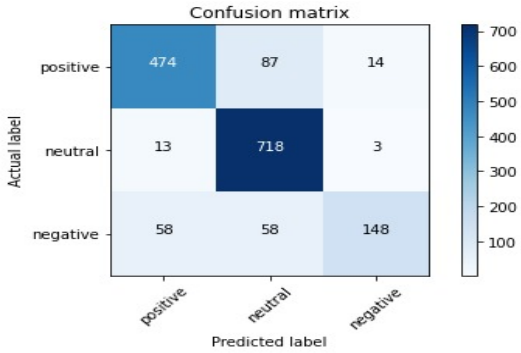

Figure 8. Confusion Matrix of SVM

4.4. Random Forest

Random Forest classifies dataset into positive, negative and neutral classes. The performance of classifier is shown in Figure 9.

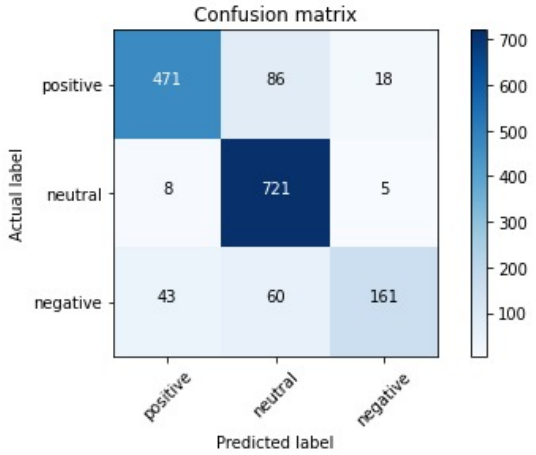

Figure 9. Confusion Matrix of Random Forest

The evaluation parameters are Precision, Recall, Accuracy and F1 Score. The formula for respective parameters are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (8)$$

Where, TP, FP, TN and FN indicate the true positive, false positive, true negative and false negative values, respectively.

The summary of classification report of four classifiers are shown in Table 1.

Table 1. Classification Report.

Algorithm	Class	Precision	Recall	F1-Score	Support
SVM	Negative	0.90	0.56	0.69	264
	Neutral	0.83	0.98	0.90	734
	Positive	0.87	0.82	0.85	575
Random Forest	Negative	0.88	0.61	0.72	264
	Neutral	0.83	0.98	0.90	734
	Positive	0.90	0.82	0.86	575
Decision Tree	Negative	0.82	0.28	0.42	264
	Neutral	0.65	0.99	0.78	734
	Positive	0.90	0.56	0.69	575
Linear Regression	Negative	0.85	0.42	0.56	264
	Neutral	0.75	0.96	0.84	734
	Positive	0.87	0.76	0.81	575

5. Results

After analyzing results of different classifiers, it is to conclude that Random Forest performed best among all classifiers giving 86.01% accuracy. Figure 10 shows final results, where x-axis represents classifiers and y-axis represents their accuracy.

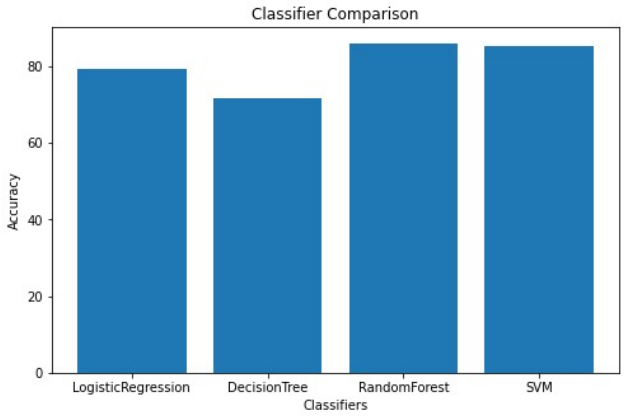

Figure 10. Accuracy of classifiers

The main crux of this manuscript is to analyze which politician among five politicians spread more negative impact through their statements. After analyzing the ratio of negative words used, it came across that Imran khan used to make more intense tweets as compared to others. Figure 11 shows the result after sentiment analysis performed on tweet dataset, where x-axis represents Politician's names and y-axis represents negative words's count.

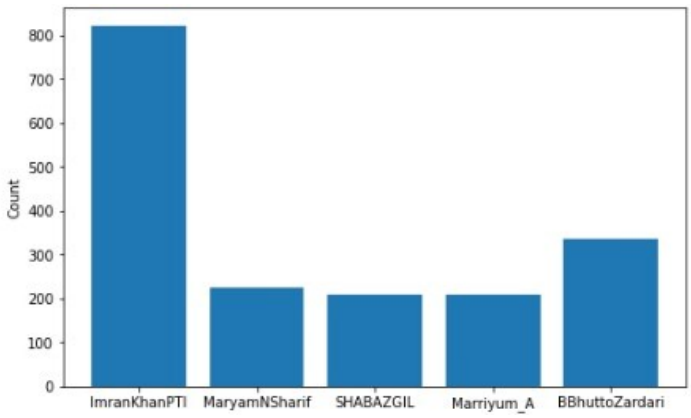

Figure 11. Political Sentiment Analysis

5.1 Comparative Analysis

The following table compares existing research with proposed methodology. The suggested methodology outperformed the research with 86% accuracy.

Table 2. Comparative Table

Paper	Dataset	Model	Results
Nikhil et al. [15]	Sentiment140	Linear SVM	83.71% Accuracy
Kabir et al. [18]	manually labeled tweets dataset on COVID-19 emotional responses	roBERTa model	89.5%
Govindan [20]	536,719 tweets were collected using Streaming Twitter API	RF + Bagging	78.74%
Proposed Methodology	7,864 tweets were collected based on important Pakistani Political Personalities	RF, SVM, DT, LR	86.01%, 85.2%, 71.5%, 79.3%

6. Conclusion & Future Work

This research is about classifying a very large number of tweets into three categories of sentiments: neutral, negative, and positive. The performance of different classifiers were evaluated and as the result, Random Forest outperformed all other classifiers with 86% accuracy. This method is deployed for anybody interested in social justice, as it provides an insight into how hate speech is employed to exacerbate prejudice and division in society. Future studies will include analysis not just in the English language but also other regional languages. It will also analyze complex emotions like sarcasm and combine them with a hybrid classifier to achieve the greatest accuracy.

Funding: This research received no external funding.

References

1. del Valle, E., & de la Fuente, L. (2023). Sentiment analysis methods for politics and hate speech contents in Spanish language: a systematic review. *IEEE Latin America Transactions*, *21*(3), 408-418.
2. Park, S., Strover, S., Choi, J., & Schnell, M. (2023). Mind games: A temporal sentiment analysis of the political messages of the Internet Research Agency on Facebook and Twitter. *New Media & Society*, *25*(3), 463-484.
3. Antypas, D., Preece, A., & Camacho-Collados, J. (2023). Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. *Online Social Networks and Media*, *33*, 100242.
4. Bestvater, S. E., & Monroe, B. L. (2023). Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, *31*(2), 235-256.
5. Lee, E., Rustam, F., Ashraf, I., Washington, P. B., Narra, M., & Shafique, R. (2022). Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis. IEEE Access, 10, 10333-10348.
6. Shinde, G. K., Lokhande, V. N., Kalyane, R. T., Gore, V. B., & Raut, U. M. Sentiment Analysis on Twitter Hashtag Datasets.
7. Lee, E., Rustam, F., Washington, P. B., El Barakaz, F., Aljedaani, W., & Ashraf, I. (2022). Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets using Stacked Ensemble GCR-NN Model. IEEE Access.
8. Gupta, I., Madan, T. K., Singh, S., & Singh, A. K. (2022). HiSA-SMFM: Historical and Sentiment Analysis based Stock Market Forecasting Model. arXiv preprint arXiv:2203.08143.
9. AlBadani, B., Shi, R., & Dong, J. (2022). A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. Applied System Innovation, 5(1), 13.
10. Singh, C., Imam, T., Wibowo, S., & Grandhi, S. (2022). A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. Applied Sciences, 12(8), 3709.
11. Rodrigues, A. P., Fernandes, R., Shetty, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Computational Intelligence and Neuroscience, 2022.
12. Yadav, N., Kudale, O., Rao, A., Gupta, S., & Shitole, A. (2021). Twitter sentiment analysis using supervised machine learning. In Intelligent Data Communication Technologies and Internet of Things (pp. 631-642). Springer, Singapore.
13. Tyagi, P., Chakraborty, S., Tripathi, R. C., & Choudhury, T. (2019, March). Literature review of sentiment analysis techniques for microblogging site. In International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India.
14. Tyagi, P., & Tripathi, R. C. (2019, February). A review towards the sentiment analysis techniques for the analysis of twitter data. In Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE).
15. Yadav, N., Kudale, O., Rao, A., Gupta, S., & Shitole, A. (2021). Twitter sentiment analysis using supervised machine learning. In Intelligent Data Communication Technologies and Internet of Things (pp. 631-642). Springer, Singapore.
16. Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. Concurrency and Computation: Practice and Experience, 32(1), e5107.
17. Aziz, A. A., & Starkey, A. (2019). Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. IEEE Access, 8, 17722-17733.
18. Kabir, M. Y., & Madria, S. (2021). EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets. Online Social Networks and Media, 23, 100135.
19. Saad, S. E., & Yang, J. (2019). Twitter sentiment analysis based on ordinal regression. IEEE Access, 7, 163677-163685.
20. Govindan, V., & Balakrishnan, V. (2022). A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection. Journal of King Saud University-Computer and Information Sciences.
21. Chowanda, A., Sutoyo, R., & Tanachutiwat, S. (2021). Exploring text-based emotions recognition machine learning techniques on social media conversation. Procedia Computer Science, 179, 821-828.