PollCast India 2024: Harnessing Data for Accurate Election Predictions

Soham Bhole
Department of Computer Engineering
(Vivekanand Education Society's
Institute of Technology) Mumbai, India
2020.soham.bhole@ves.ac.in

Deanna Fernandes
Department of Computer Engineering
(Vivekanand Education Society's
Institute of Technology) Mumbai, India
2020.deanna.fernandes@ves.ac.in

Prerna Bajaj
Department of Computer Engineering
(Vivekanand Education Society's
Institute of Technology) Mumbai, India
2020.prerna.bajaj@ves.ac.in

Nishtha Batra*
Department of Computer Engineering (Vivekanand Education
Society's Institute of Technology) Mumbai, India
2020.nishtha.batra@ves.ac.in

Mrs. Abha Tiwari
Department of Computer Engineering(Vivekanand Education
Society's Institute of Technology) Mumbai, India
abha.tewari@ves.ac.in

Abstract:
Social media platforms like Facebook play a pivotal role in facilitating mass communication, opinion expression, and information sharing, generating extensive unstructured data. Researchers are increasingly focusing on analyzing sentiments within this vast user-generated text. This study utilizes machine learning algorithms to explore the complexities of scrutinizing voter sentiments on social media, providing insights into evolving public opinion dynamics. It covers sentiment analysis (SA) methods, standard preprocessing techniques, diverse word embeddings, popular benchmark datasets, evaluation metrics, and publicly available resources for SA tasks. The implemented system is evaluated using Naïve Bayes, SVM, and a modified BERT approach. The survey also explores practical applications of SA tasks and concludes by identifying research challenges and suggesting future directions for investigation.

Keywords:
Benchmark Datasets, BERT Approach, Evaluation Metrics, Machine Learning Algorithms, Naïve Bayes, Sentiment Analysis (SA), SVM (Support Vector Machine), User-generated Text, Word Embeddings.

1. Introduction

In the evolving landscape of democratic processes, technology plays a pivotal role in predicting and understanding electoral outcomes, a significance particularly pronounced in the lead-up to the 2024 elections. This study aims to pioneer a novel approach to election prediction by leveraging modern data analytics techniques, with a notable emphasis on sentiment analysis. Going beyond conventional polling methods, sentiment analysis, powered by natural language processing and machine learning, enables the extraction of nuanced public sentiments from extensive datasets sourced from social media, news articles, and online forums [1]. The approach seeks to provide sophisticated insights into the electorate's mood by analyzing prevailing sentiments around political candidates and key issues. Incorporating machine learning ensures real-time adaptability to evolving language patterns and emerging sentiments, ensuring the model's ongoing relevance and accuracy. This research extends beyond traditional election prediction methodologies to offer a comprehensive platform, not only forecasting electoral outcomes but also serving as a valuable tool for campaign strategists and politicians [2]. By exploring real-time sentiment trends, identifying influencers, and measuring the impact of events on public perception, stakeholders gain crucial insights for informed decision-making in the dynamic political landscape [3]. The rest of this article is organized in the following manner. Section II offers an overview of the current body of literature, followed by Section III which elaborates on the different data sources accessible and the construction of the machine learning model. Section IV provides a summary of diverse results, evaluation metrics, and tools employed. Moving forward, Section V discusses several research gaps and potential future directions. Finally, Section VI concludes the study. We then spotlight the referenced research papers. The research hypothesis underlying this study lies in the effectiveness of leveraging sentiment analysis, coupled with machine learning techniques, to predict electoral outcomes accurately. The hypothesis is validated by comparing the predictions generated by the proposed model with the actual election results. Additionally, assessing the model's performance metrics, such as accuracy, precision, and recall, against established benchmarks and traditional polling methods would further validate its efficacy.

1.1 Motivation :

The motivation behind exploring election prediction for the 2024 elections using sentiment analysis stems from the transformative potential this approach holds in reshaping the way we understand and engage with the democratic process. Traditional polling methods often face challenges in capturing the complexity and dynamism of public opinion, especially in an era where information is rapidly disseminated through various digital channels. The motivation arises from the recognition that sentiment analysis, powered by advancements in natural language processing and machine learning, provides an unprecedented opportunity to tap into the pulse of the electorate. By analyzing the sentiments expressed in social media discussions, news articles, and online forums, we can gain real-time insights into the evolving opinions and attitudes of voters.

2. Literature Survey

In later a long time, there has been a surge of intrigued in utilizing machine learning methods for estimation examination errands. Striking calculations such as Back Vector Machines (SVM), Gullible Bayes (NB), and Calculated Relapse (LR) have earned consideration for their adequacy, particularly when coupled with appropriate highlight extraction strategies [1].

One ponder [4] digs into the progressions in fine-grained opinion examination utilizing both machine learning (ML) and profound learning (DL) techniques. Conventional ML calculations such as SVM and choice trees, nearby DL methods like Convolutional Neural Systems (CNN) and Repetitive Neural Systems (RNN), have appeared guarantee in precisely categorizing opinions inside printed information. In spite of these progressions, challenges hold on, counting the shortage of labeled datasets and the complexity of taking care of wonders like mockery. The paper moreover diagrams future investigate roads, such as investigating self-supervised learning, outfit strategies, and novel building plans, to advance progress in fine-grained assumption examination frameworks.

Another investigative endeavor [5] underscores the developing significance of Profound Learning (DL) strategies in Twitter Assumption Investigation (SA). It gives an outline of later DL progressions, categorizes existing writing, talks about challenges particular to Twitter information, such as its inborn complexities and estimation awkwardness, and proposes coordination SA with other Normal Dialect Preparing (NLP) strategies for upgraded execution. In spite of the guarantee of DL, determined challenges recommend progressing openings for refinement in DL-based SA on Twitter. In a partitioned examination [6], analysts utilized the WEKA apparatus to assess different classifiers on a dataset related to portable workstations. Classifiers surveyed included Gullible Bayes, Multilayer Perceptron, J48, Arbitrary Timberland, and OneR, among others. Interestingly, J48 showed the foremost favorable execution generally, outflanking other classifiers in spite of Credulous Bayes' advantage in speed of development. This perception proposes that diverse classification calculations may exceed expectations based on the characteristics of the dataset being analyzed.

Moreover, a distinctive thing about [7] presents an inventive approach to social image–text assumption classification, outperforming existing methodologies. It utilizes a half breed cross-modal consistency demonstrated together with advantages including refining to exchange visual information to content opinion classification, coming about in an outstanding precision pick up of 3.3% and a weighted F1 pick up of 4.8%. The paper insights at future applications, proposing the expansion of this information refining approach to other assumption classification errands past social media image-text investigation.

In outline, these thoughts collectively outline the advancing scene of opinion examination, highlighting the adequacy of different machine learning and profound learning methods whereas moreover underscoring tireless challenges and future investigate bearings within the field.

Table 1- Metrics Used

Metrics	Publication
Precision	[1], [4], [5], [9], [11], [18]
Recall	[3], [6], [9], [12], [16], [19], [20]
F1 score	[3], [7], [8], [10], [15], [19]
Accuracy	[2], [13], [14], [17], [18], [20]
ROC Area	[1], [12], [16], [17]

Table 2-Twitter Based analysis

Publication	Key Takeaway	Limitations
[8]	The methodology analyzed 796 messages using Naïve Bayes, SVM, NLTK, TextBlob, and an RNTN model for sentiment analysis. The weighted score predicted outcomes for the 2018 US midterm elections.	Manually selecting 796 messages is labor-intensive and impractical for handling a large volume of messages.
[9]	Two major parties, BJP and Congress, were analyzed using Tweepy for tweet collection, TextBlob for sentiment scores, and bidirectional RNN models; visualizations included bar graphs and Word Cloud to identify the party with the most positive tweets as the potential election winner.	Emoticons and sub-regional specificity weren't considered; the analysis relied on a small dataset using only Twitter.
[10]	This paper utilizes sentiment	Sentiment analysis

	analysis on tweets related to the 2019 Indian Lok Sabha elections, employing machine learning models and feature extraction methods to predict election outcomes based on positive or negative scores assigned to each tweet.	challenges encompass language complexity, biases in social media data, limited English tweet coverage, difficulty capturing rapid sentiment changes, and the challenge of generalizing findings to diverse elections.
[11]	The paper introduces a sentiment-based methodology, utilizing a sentiment dictionary to calculate candidate sentiment scores as indicators for polls, with the primary goal of enhancing election prediction accuracy by integrating collective wisdom from social media data. The approach innovatively combines volume of mentions, sentiment analysis, and public acceptance scores to achieve this objective.	The dataset does not include non-author users, and although users can publish multiple articles during the collection period, they are limited to casting only one vote in the election.

3. Methodology

3.1. *Data Collection:*

Our data gathering approach runs from 1962 to 2019, with Kaggle datasets used to track state-level election results and determine which political party won. This historical information serves as the foundation for predictive analysis of the 2024 elections. To measure current emotions, we scraped the official pages of political parties and extracted comments from the previous five years. RAPIDAPI was used to retrieve comments in real time from Twitter and YouTube. Additionally, the number of followers and followers in real time for every party was taken from every social media platform. Each party's real-time news piece was also gathered. This real-time dataset provides important insights on changing public engagement. Furthermore, Wikipedia data on the number of seats won by each party improves our study by providing a broader context and historical trends that help to make informed predictions for the forthcoming elections.

3.2. *Data Preprocessing:*

During the data preprocessing step, Facepager [12] was utilized to collect information from Facebook [13] using political party official Facebook IDs. We improved the extraction approach by employing date filters to gather only comments made during the specified window. This concentrated approach ensured that the dataset reflected current and relevant public sentiments. We streamlined the data after extraction by converting it to an Excel sheet, which allowed for further analysis and integration with other datasets. Following that, rigorous preprocessing is performed, which includes noise removal, format standardization, and the appropriate handling of missing values via imputation or deletion.[14]

Table 3- Pre processing Steps

Steps	Publications
Lowercasing	[12], [15], [18], [19]
Removing punctuation and special characters	[1], [2], [3], [7], [9], [11], [12], [14]
Removing stop words	[4], [5], [6], [7], [10], [15], [16], [20]
Tokenization	[11], [13], [14], [19], [20]
Text Vectorization	[3], [8], 15] ,[16], [17], [18]

3.3. *Feature Engineering:*

During our feature engineering process, we selected and refined key features for predicting the 2024 elections, including constituency details (name and number), election type, state, candidate name, party affiliation, number of electors, total votes, turnout percentage, margin of victory, margin percentage, and election year [15]. These features serve as a solid foundation for creating a predictive model that integrates historical patterns as well as contextual factors important for forecasting future political results.

3.4. *Predictive Modeling:*

In our 2024 election prediction model, we used Random Forest Regression [16]. This approach was chosen for its ability to handle complex datasets while capturing subtle data patterns. Using this method, we intend to predict the winning political party based on engineered characteristics, historical trends, and a range of other factors. The Random Forest Regression model enables a complete evaluation of the expected outcome of the upcoming elections.

3.5. *Random Forest Training:*

```
Input: Training dataset with features (X) and class labels (Y).
Output: A collection of decision trees (T1, T2, ..., Tn).
Procedure:
For i = 1 to n:
Randomly select a subset of the training data with replacement (bootstrap sample): Di.
Train a decision tree Ti using Di by:
Selecting a random subset of features at each split.
Splitting nodes based on the best feature (e.g., Gini impurity or information gain).
Output the collection of decision trees: {T1, T2, ..., Tn}.
```

3.6. *Random Forest Prediction:*

Input: A new input sample Xnew.

```
Output: Predicted class label (Ŷ).
Procedure:
For each decision tree Ti in the forest:
Pass Xnew through Ti.
Collect the predicted class label Ŷi.
Determine the majority class label as the final prediction:
Ŷ = mode(Ŷ1, Ŷ2, ..., Ŷn)
where n is the number of trees in the forest.
```

3.7. *Sentiment Analysis from Facebook Comments:*

Supplementary sentiment analysis [17] on Facebook comments categorizes sentiments as positive, negative, or neutral, providing additional insights.

3.8. *State-level Prediction:*

In our state-level prediction research, we aggregated the data by state and political party to calculate the average margin percentage. We then determined which party won in each state and year. By resetting the index, we acquired a structured DataFrame including information on the winning party, state, and year. To calculate the overall winner, we counted the number of winning parties in each state-year combination. This approach enabled us to generate informed forecasts about the winning party in each state, providing a more comprehensive picture of the probable national outcomes for the 2024 elections.

3.9. *Visualization:*

In order to visualize the distribution of positive, negative, and neutral remarks for our Facebook study, we created a bar chart [18]. This allowed us to quickly summarize the sentiments of people on social media. Additionally, a line graph that showed the evolution of comment sentiments over time for each political party was included to shed light on sentiment trends and advancements. The winning party in each state was visually represented on a colored map that made understanding local political dynamics quick and easy. By enabling a detailed analysis of public viewpoints, temporal trends, and geographic patterns within the context of political discourse on Facebook, these visualizations enhance the interpretability of our data.

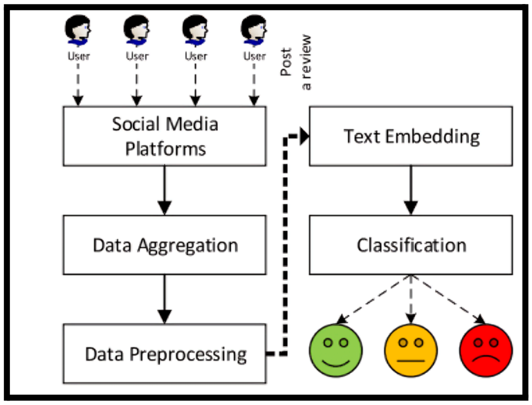

Fig 1- Modular Workflow for Sentiment Analysis and Election Prediction

3.10. *Integration with Predictive Modeling:*

A comprehensive prediction for the overall political winner was generated by integrating our machine learning models with forecasts from past winner party models and Facebook sentiment analysis [19]. Easefully incorporated into a front-end dashboard, this unified result merged data from previous electoral outcomes and social media sentiment. A comprehensive and informed perspective on the upcoming elections is provided to stakeholders by the dashboard, an intuitive tool that gives them a real-time, data-driven image of the political landscape.

3.11. *Visualization Enhancement:*

The dashboard provides a complete snapshot of the political situation through the use of graphs, numerical summaries, and state-by-state winners. It is designed to facilitate extensive visualization. While sentiment analysis results offer a snapshot of public sentiment, dynamic line graphs show patterns over time. Analytical graphs allow for comparison studies between the parties, and the addition of a map representation gives a geographical understanding of state-by-state winners. By attempting to condense complicated political data into a clear and useful interface, this multimodal approach hopes to empower all stakeholders to make informed choices.[19]

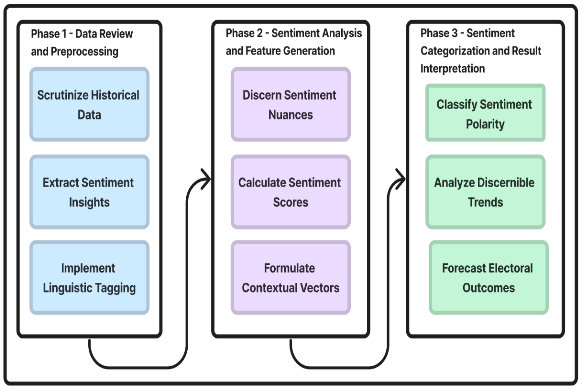
Phase 1 - Data Review and Preprocessing
Scrutinize Historical Data
Extract Sentiment Insights
Implement Linguistic Tagging
Phase 2 - Sentiment Analysis and Feature Generation
Discern Sentiment Nuances
Calculate Sentiment Scores
Formulate Contextual Vectors
Phase 3 - Sentiment Categorization and Result Interpretation
Classify Sentiment Polarity
Analyze Discernible Trends
Forecast Electoral Outcomes

Fig 2- Experimental data flow diagram

This project involves a thorough analysis of historical data, with a focus on sentiment insights extraction from social media sites like Facebook. Linguistic tagging is used in the process to improve understanding. The second phase involves calculating sentiment scores and creating contextual vectors to capture the subtleties of language and expression in order to identify complex sentiments. Sentiment polarity is methodically categorized in the last stage, which makes trend analysis and election outcome prediction possible. The main objective is to use sentiment analysis tools to offer insightful analysis of the impending elections.

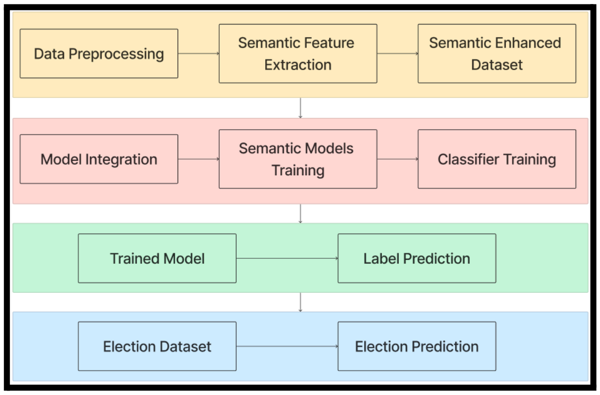

Fig 3- Modular Workflow for Sentiment Analysis and Election Prediction

This modular workflow consists of preprocessing the data, extracting semantic features, integrating machine learning models (Naive Bayes, SVM), and training semantic models and classifiers [20]. The trained model makes sentiment label predictions while preprocessing real-time election data for the election prediction step. Sentiment-driven election forecasts have a strong foundation thanks to the modular design, which guarantees scalability and ease of integration.

4. Results

We retrieved vital insights from real-time data by using sentiment analysis to identify prevailing thoughts and attitudes in political conversation. This advanced technique aided in the detection of developing trends and the accurate assessment of public attitude [20]. The result of this strategy was a visually appealing map, color-coded to represent political affiliations and opinions at the state level. This map is an important tool for not just illustrating party distribution, but also exposing sentiments that influence political dynamics. Machine learning algorithms based on historical data anticipate political trends, which are refined in real time using sentiment analysis. The state-by-state distribution map visualizes the political landscape based on expected outcomes, providing a clear portrayal of the current political reality.

To improve the interpretability of our findings, we used word cloud visualization on the extracted dataset, which shows a compact depiction of frequently recurring phrases and feelings. In addition, a graph was created to show the distribution of sentiments, which were classified as neutral, positive, or negative. This graphical form provides a concise overview of sentiment trends in the dataset, allowing for a more complete understanding of political discourse dynamics. Our analysis uses these visualizations to not only delve into the intricate components of sentiment, but also to give accessible and informative tools for evaluating the dataset.

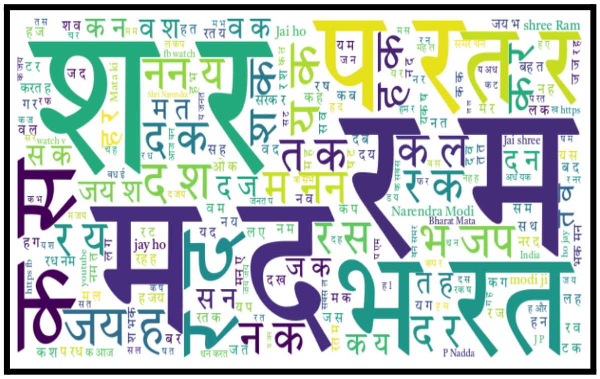

Fig 4- Word Cloud

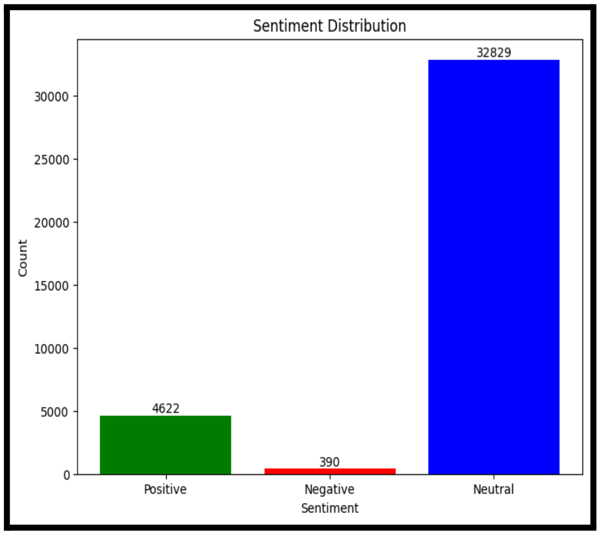

Fig 5- Sentiment Distribution of facebook comments

Furthermore, our in-depth analysis uncovered temporal trends in political preferences, emphasizing the pivotal role of historical data in recognizing patterns and shaping predictions. The geographical variations highlighted in the map provide additional insights into regional dynamics, shedding light on the diverse political landscape [18]. In essence, this comprehensive analysis combines historical and real-time data, offering nuanced insights into the evolving political landscape by integrating predictive modeling, sentiment analysis, and geographical considerations.

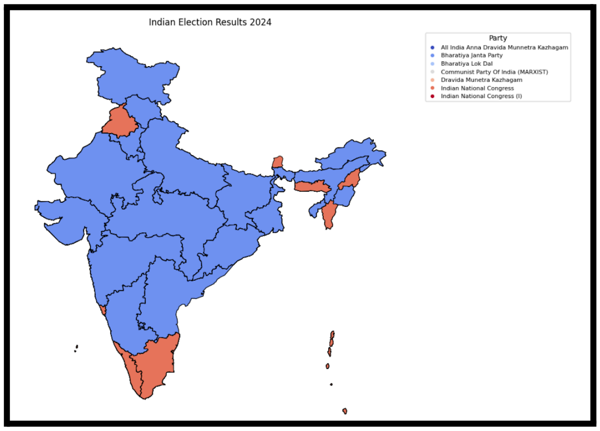
Indian Election Results 2024
Party
● All India Anna Dravida Munnetra Kazhagam
● Bharatiya Janta Party
● Bharatiya Lok Dal
● Communist Party Of India (MARXIST)
● Dravida Munetra Kazhagam
● Indian National Congress
● Indian National Congress (I)

Fig 6- The representation of predicted results based on location, displayed on the political map of India.

5. Future Work

The strategic roadmap for Project II uses machine learning and a wide range of data to provide exact election forecasts. Feature engineering combines sentiment analysis, past election data, and voter demographics. The user-centric dashboard, created with Figma, displays real-time forecasts and sentiment analysis results via intuitive charts and maps. Rigorous testing, user feedback, and ethical concerns ensure accuracy, usability, and privacy. The dashboard is distributed with thorough documentation, stressing continual monitoring and adaptation for important insights in the face of dynamic data trends.

6. Limitations of proposed study

The existing system for election prediction in 2023, utilizing sentiment analysis from social media and online databases, faces several significant drawbacks. These include the potential for bias and inaccuracy stemming from selection bias, where certain demographic groups are overrepresented while others are underrepresented, leading to skewed views of public sentiment. Moreover, the prevalence of misinformation and manipulation on social media platforms poses a challenge, as sentiment analysis may inadvertently capture sentiments influenced by false information or orchestrated disinformation campaigns, thus resulting in flawed predictions. Privacy concerns also loom large, as the collection of data from these sources raises ethical questions regarding consent and data privacy. Additionally, the dynamic nature of social media, characterized by rapidly changing trends and sentiments, poses a challenge for the existing system, potentially leading to outdated or irrelevant predictions. Finally, sentiment analysis's reliance on textual data presents limitations, as nuances such as sarcasm, irony, and subtle emotions may be challenging to accurately interpret, leading to misclassification issues.

6. Conclusion

The Election Prediction System for the 2024 elections employs sentiment analysis and a modular design, providing a comprehensive and adaptable approach to assessing public opinion. The solution provides real-time insights thanks to seamless workflow integration from data collection to preprocessing, as well as advanced NLP and machine learning capabilities. The Historical Analysis Module adds nuance by evaluating long-term sentiment trends, whereas the Decision Support Module provides useful information for political stakeholders. The User Interface Module improves accessibility, while the Security and Privacy Module promotes ethical data practices and protects sensitive information. Overall, this comprehensive methodology offers a reliable and ethical method of assessing public mood in a volatile political setting.

7. References

[1] Wani, G., & Alone, N. (2014). *A survey on impact of social media on election system*. International journal of computer science and information technologies, 5(6), 7363-7366.

[2] Nugroho, D. K. (2021). *US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis. 2021* 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[3] Khurana Batra, P., Saxena, A., Shruti, & Goel, C. (2020). *Election Result Prediction Using Twitter Sentiments Analysis. 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)..*

[4] Yadav, Nikhil & Kudale, Omkar & Rao, Aditi & Gupta, Srishti & Shitole, Ajitkumar. (2021). *Twitter Sentiment Analysis Using Supervised Machine Learning.* 10.1007/978-981-15-9509-7_51.

[5] Chaudhary, L., Girdhar, N., Sharma, D., Andreu-Perez, J., Doucet, A., & Renz, M. (2023). *A Review of Deep Learning Models for Twitter Sentiment Analysis: Challenges and Opportunities.* IEEE Transactions on Computational Social Systems.

[6] Singh, N., & Jaiswal, U. C. *Prediction of Student Score Performance of Sentiment Analysis Using Hybrid Cross Validation Machine Learning Techniques.*

[7] Liu, H., Li, K., Fan, J., Yan, C., Qin, T., & Zheng, Q. (2022). *Social Image-text Sentiment Classification With Cross-Modal Consistency and Knowledge Distillation.* IEEE Transactions on Affective Computing.

[8] Tsai, M. H., Wang, Y., Kwak, M., & Rigole, N. (2019, December). *A machine learning based strategy for election result prediction.* In 2019 international conference on computational science and computational intelligence (CSCI) (pp. 1408-1410). IEEE.

[9] Kumar, R., Kumar, S., & Soni, A. (2021, March). *Election prediction using twitter and GIS*. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 306-311). IEEE.

[10] Batra, P. K., Saxena, A., & Goel, C. (2020, November). *Election result prediction using twitter sentiments analysis*. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 182-185). IEEE.

[11] Firmansyah, F., Zulfikar, W. B., Maylawati, D. S., Arianti, N. D., Muliawaty, L., Septiadi, M. A., & Ramdhani, M. A. (2020). *Comparing Sentiment Analysis of Indonesian Presidential Election 2019 with Support Vector Machine and K-Nearest Neighbor Algorithm.* 2020 6th International Conference on Computing Engineering and Design (ICCED).

[12] https://github.com/strohne/Facepager

[13] https://www.facebook.com/

[14] Singh, L. G., Anil, A., & Singh, S. R. (2020). *SHE: Sentiment hashtag embedding through multitask learning*. IEEE Transactions on Computational Social Systems, 7(2), 417-424.

[15] Wang, L., Niu, J., & Yu, S. (2019). *SentiDiff: combining textual information and sentiment diffusion patterns for Twitter sentiment analysis.* IEEE Transactions on Knowledge and Data Engineering, 32(10), 2026-2039.

[16] Tsai, M. H., Wang, Y., Kwak, M., & Rigole, N. (2019, December). *A machine learning based strategy for election result prediction.* In 2019 international conference on computational science and computational intelligence (CSCI) (pp. 1408-1410). IEEE.

[17] Bayrak, C., & Kutlu, M. (2022). *Predicting Election Results via Social Media: A Case Study for 2018 Turkish Presidential Election.* IEEE Transactions on Computational Social Systems.

[18] Nazir, A., Rao, Y., Wu, L., & Sun, L. (2020). *Issues and challenges of aspect-based sentiment analysis: A comprehensive survey.* IEEE Transactions on Affective Computing, 13(2), 845-863.

[19] Zeng, J., Zhou, J., & Huang, C. (2023). *Exploring Semantic Relations for Social Media Sentiment Analysis.* IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[20] Alagi, S., Kolambe, M. T., Bibe, M. V., Gite, M. K., & Chachar, M. S. *A MACHINE LEARNING APPROACH FOR PREDICTION OF ELECTION INFLUENCE USING SOCIAL MEADIA.*