Election Prediction: Collecting And Analyzing User Sentiments Through Public Polls, YouTube Comments of Campaign Speeches, and Manifesto

Posa Hemanth Kumar, K. Yashwanth, Manubothula Namrata
National Institute of Technology Puducherry.

Abstract

This paper proposes a novel approach to election result prediction by analyzing YouTube comments on party manifestos, campaign speeches, and public opinion polls. YouTube became a prominent channel for expressing the opinions and feelings of the general public. Statistical and sentimental analysis are the methods that are used to know the general public opinion. In this respect, statistical and sentimental analysis can be used in predicting the outcomes of upcoming elections by assessing the sentiments of the public through YouTube and public media reports. This paper outlines the evaluation of statistical and sentimental analysis to predict election results. This paper totally concentrates on Andhra Pradesh state by taking two major parties Party Y and Party T which are in this state.

Keywords: Election prediction, YouTube comments, opinion polls, manifesto, campaign speeches, machine learning, Neural Network, sentimental analysis.

1. Introduction

1.1 Background

Elections are a major part of democratic societies, as citizens have control over their country by selecting the leader who can fulfill their needs and make the future bright. This state situated in the southeastern part of India often has highly contested elections every time a major characteristic of the world's largest democracy. In this diverse and culturally rich state, two major political parties, Party Y and Party T have constantly fought for power in this state. The accurate prediction of election outcomes is necessary for the democratic process, which affects the sentiments of voters on a party. The challenge of election prediction is more difficult in this state, due to the influence of various cultures and situations.

1.2 Research Domain Introduction

Election prediction, a multidisciplinary field at the intersection of political science, data science, and social media analysis is highly complex task. While traditional methods like opinion polls and surveys have provided valuable insights it is not always efficient due to the mass number of people and time this is where social media comes in. Social media platforms have loads of data for understanding voter sentiment and behavior.

The present paper focuses on using two significant domains: politics in this state opinion polls and the social media platform YouTube, a platform that highly contains political updates. YouTube's humongous collection of campaign speeches, manifestos, and user opinions expressed through comments form the basis of our research, which serves as data for predicting the trend of elections in this state.

1.3 Research Gaps in the Existing Field

While the research scope in election prediction has grown significantly, many research papers are based on the sentimental analysis of tweets from Twitter. While Twitter's role in politics cannot be understated, YouTube, has huge database and videos where people can view and react real-time mass of users prefer YouTube over Twitter for political reference so many research papers are not totally dependent. So, this paper aims to tackle this problem by using YouTube comments as a rich source of election sentiment data.

1.4 Challenges and Missing Parameters

The prediction of election outcomes is highly complex task due to the effect of more factors that influence voter decisions. In the collection of YouTube comments, unique challenges persist. Sentiment analysis in this context is complicated by the different mix of languages, regional dialects, and subtexts found in the comments. Furthermore, user bias, noise, and misleading information pose significant challenges in interpreting the data. Existing research models does not provide a way to merge opinion polls, campaign speeches and manifestos into one unified predictive framework. This paper aims to solve these challenges.

1.5 Limitations of Previous Studies

Past research has specifically concentrated on using Twitter data for election prediction, employing techniques such as sentiment analysis and social network analysis. Our research extends the scope to include YouTube comments, offering a more comprehensive view of voter sentiment and behavior in this state.

1.6 The Role of Ensemble Learning

In this research we use ensemble learning to predict the best output in manifestos and campaign speeches we used several machine learning algorithm approaches to maximize predictive accuracy. By combining different machine learning techniques and algorithms, we aim to create a model with the top three accuracy models. Our approach will consider methods such as decision trees, support vector machines, and neural networks, among others. The ensemble learning technique used to select the best frequented output of that comment which was predicted top three models. This method will significantly enhance the reliability and accuracy of the model.

1.7 Contributions

The contribution of this project to the field of election prediction:

- The development of election prediction model that uses YouTube comments as a primary data source.
- A new and enhanced approach that uses ensemble learning to select the best prediction outcome for election prediction.

2. Related work

Many research articles exist which helped in predicting election outcomes using social media data and public polls. This section lists few of such articles and their findings.

2.1 Influence of social media on elections.

Based on the election prediction methods as researched by this paper [7], it has been found that many of the studies used Twitter as a main corpus to analyze the results and they used social media APIs for collecting such data and as a major method used sentimental analysis from the collected data to predict winners and for this analysis many used lexicon-based approaches but only few used deep learning [2] techniques to predict the outcome.

The paper [5] took a different approach he found how pre poll projections affects the user sentiments also a major flaw which is how false twitter data affects overall prediction accuracy in the KNN algorithm he used another major disadvantage identified is the voters swing and real time political events and their influence.
Many papers[1] also tend to prove that hybrid approach for sentimental analysis i.e. combining machine learning and lexicon based approaches increases accuracy and is highly accurate as compared to traditional machine learning algorithms.

The paper [9] on his work while predicting German elections (2009) found that many insights can be drawn from just the sentimental analysis of tweets by politicians and their party offices. He also found that as low as just number of tweets on a party is enough to predict election results.

The paper [8] studied the election results of European union and analyzed the tweets corresponding to 24 politicians spanned across 3 countries and correctly predicted the outcome. This study proved that this tweet analysis is highly accurate way to predict the elections outcome.

But Jaidka et al.[1] used combination of three methods volumetric, sentiment, and social network analysis, for predicting election results of Malaysia (2013), Pakistan (2013), and India (2014). It proved to be correct for India and Pakistan but not for Malaysia. Findings include using recent twitter tweets and combination of different methods increase accuracy compared to one.

Chung and Mustafaraj [3] combined two methods lexicon-based implemented in two papers Tumasjan et al. [6] and O'Connor et al. [6] and applied to US election and found that just number of mentions a politician gets is not enough to predict the election results. Hence why their accuracy was very low 47.19%.

Ceron-Guzman and Leon-Guzman[2] studied the 2014 Colombia election using volumetric approach. They combined this with polls and used linear regression (Ordinary Least Square, Ridge, Lasso, and Support Vector Regression). Their results were very inaccurate hence showing how twitter data might not always be correct.

Anjaria et al. [1] (Anjaria & Guddeti, 2014) used SVM, Naïve Bayes, maximum entropy, and MLP-BP for sentiment analysis to predict 2012 U.S elections and the 2013 Karnataka elections. But it was not accurate and largely differed from actual vote share.

3. Proposed Framework

In this section, the idea behind this project is explained by illustration the Fig. 1.

Proposed framework in this project is hybrid approach, where we considered statistical approach and sentimental analysis approach. In statistical approach, public polls before elections are taken as opinion polls, and after elections taken as exit polls. In sentimental analysis approach, YouTube comments which are given by public. Their opinion on campaign speeches and manifestos are taken in sentimental approach. This hybrid approach is illustrated in Fig. 1.

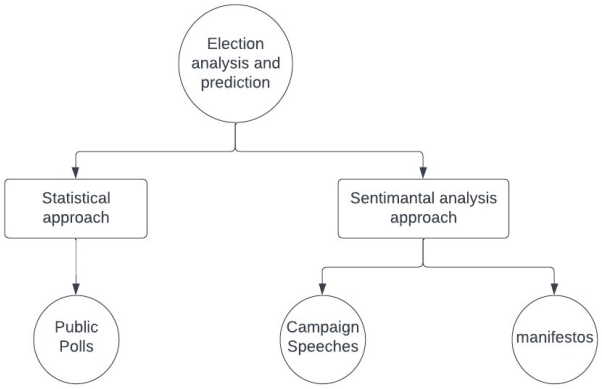

Fig. 1 Election analysis and prediction approach

4. Statistical and Sentimental analysis process

Statistical and Sentimental analysis can be viewed as disciplines such as information extraction in data collection, natural language processing in data preprocessing, statistical and machine learning for model training, and predicting the results. The step-wise step process of statistical and sentimental analysis is depicted in Fig. 2.

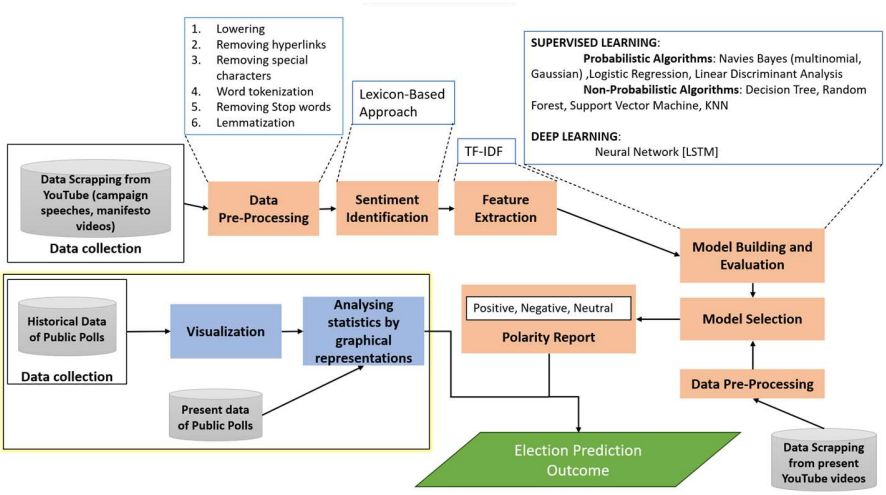

Fig. 2 Statistical and sentimental analysis process

4.1 Data collection

In a statistical approach, data is collected from public media reports about public polls. Reports are generated monthly and taken as opinion polls, while reports after elections are taken as exit polls. 2019 data was taken as a train data set where opinion polls and exit polls of 2019 were statistically studied by comparing them with the results of 2019 elections. Up to 2024 data taken as a test data set where only opinion polls are available and predicting the 2024 results based on 2019 available data. For the statistical study, Tableau software was used to analyze the trends.

In sentimental analysis approach, campaign speeches and manifestos data were taken from YouTube comments, which were given by the public on their opinion. 2019 Campaign speeches and manifestos are considered as train data sets. And then up to 2024 data was taken as test data sets.

4.2 Data Pre-processing

Pre-processing involves filtering and cleaning the data. If preprocessing is not done correctly then result will be incorrect in classification. Comments include much noisy text like misspelled words, emojis chat, leading and trailing special characters which leads to degrading the accuracy of a model. Data filtering and cleaning contains various steps such as removing hyperlinks in comments, lowering the text, removing emojis, white spaces, stop words, and punctuations.

4.3 Sentimental identification

Most of the comments are having only subjective text. Subjective text alone cannot give satisfactory results for sentiment analysis, including objective gives the best sentiment for given comment. For example, comment "*Jai party Y*" or "*Jai party T*" does not contain any sentiment which is neutral but it is positive for supported party and negative for opposite party. Sentiment identification done by using VADER (Valence Aware Dictionary and sEntiment Reasoner) which is lexicon-based technique as shown in Fig. 2.

4.4 Feature extraction

Raw data which are preprocessed comments are in objects. Models cannot be trained by objects or strings; model can understand only numerical data (may be integers or floats). There are many techniques that transforms preprocessed comments to numerical like Term Frequency-Inverse Document Frequency (TF-IDF), CountVectorizer, Bag of words, Pos-tagging, etc., Here we have used TF-IDF using unigrams.

4.5 Model selection

For model selection, we have used an ensemble approach for predicting the most accurate output. In supervised machine learning there are various probabilistic classifications (Multinomial Navies Bayes (MNB), Gaussian Navies Bayes (GNB), Logistic Regression (LoR), Linear Discriminant Analysis (LDA)) are used whereas non-probabilistic classifications (Decision Tree, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN)) are used. In deep learning, Long Short-Term

Memory (LSTM) model is used. Top three models are taken based on accuracy and predicts the output on most frequent result by three models, which known as ensemble approach.

4.6 Polarity report

Polarity report provides outcomes of public polls 2024 prediction along with 2019 and campaign speeches and manifestoes, 2024 prediction along with 2019 for both parties. These three parts (public polls, campaign speeches and manifestoes) of our project are compared with each other with their respective parties and finally normalized to 100 percent. Which party has greater percentage when compared to other party, that party is considered as a winner in 2024 upcoming elections on our prediction.

5. Findings

In this section, the outputs of every section are shown in pictorial representation to get an overview of whole project process and analysis. The study has analyzed some trends followed by the researchers in predicting election results. A graphical and pictorial representation gives the whole idea and approach of prediction on results. Firstly, analyzing the public polls. In public polls about statistical seat share of parties of opinion polls of 2019 is plotted as a bar graph used as test data and illustrated in Figure 3. Blue color bar in representation talks about seat share of Party Y, yellow for color Party T, and pink color for other Party seat share. Similarly, in public polls about seat share of parties of exit polls of 2019 is illustrated in Figure 4.

Figure 5 is about final seat share after elections 2019 results. In 2019, Party T got 3 seat share, Party Y got 22 seat shares, and others got 0 seats in 2019 elections. Figure 6 illustrates about predicted seat share for 2024 elections on the statical study of 2019 elections.
Figure 6 illustrates about sentiment distribution in percentage of Party Y Campaign Speeches in 2019 in the form of a pie chart with emotions as positive negative and neutral. Figure 7 illustrates about sentiment distribution of Party T Campaign Speeches in 2019.

Figures 8 and 9 talk about sentiment distribution of Party Y and Party T manifestos in 2019 respectively. Figures 10 and 11 talk about sentiment distribution of Party Y and Party T Campaign Speeches for 2024 elections respectively. Figures 12 and 13 talk about sentiment distribution of Party Y and Party T manifestos for 2024 elections respectively.

Table 1 describes about predicted sentiment distribution of both the parties as positive percent, negative percent, and neutral percent of Campaign Speeches for 2024. Similarly, Table 2 describes about predicted sentiment distribution of Manifestoes for 2024. Table 3 describes about supported percent for respective parties in all public polls, campaign speeches and manifestos. This table shows overall conclusions derived from this project.

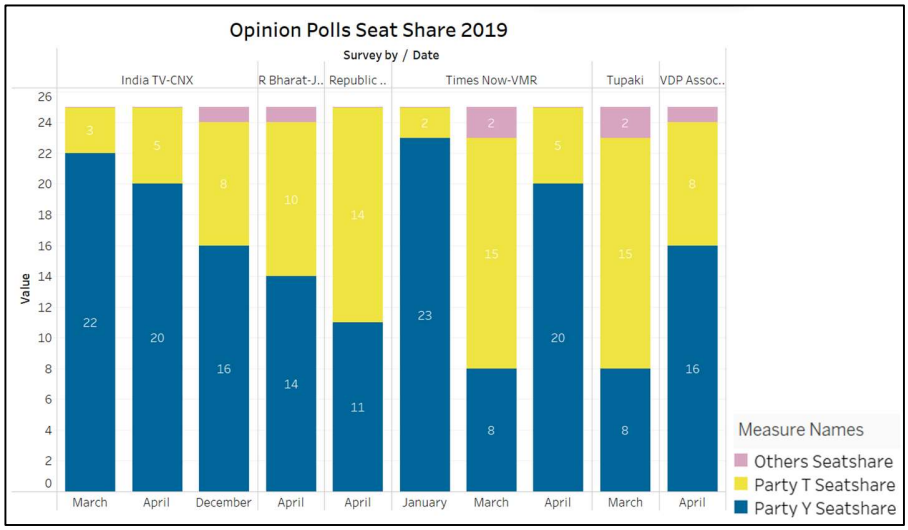

Fig. 3 Seat share on opinion polls for 2019 elections

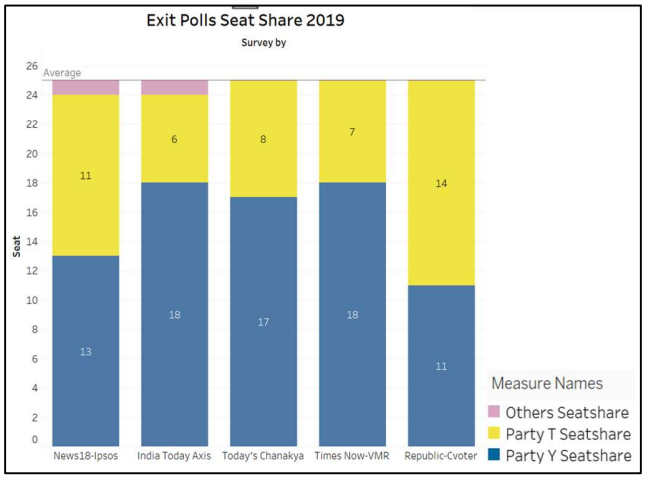

Fig. 4 Seat share on exit polls for 2019 elections

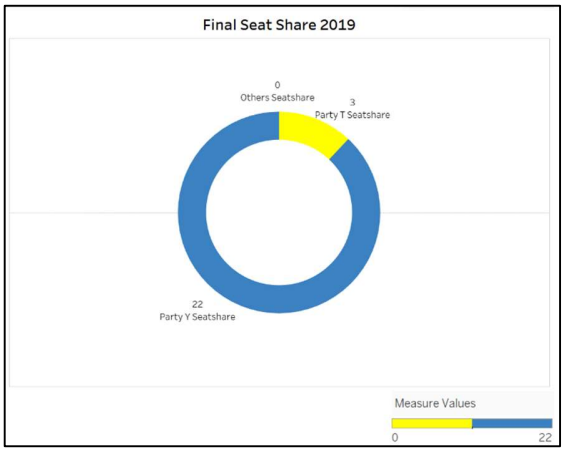

Fig. 5 Original seat share in 2019 elections

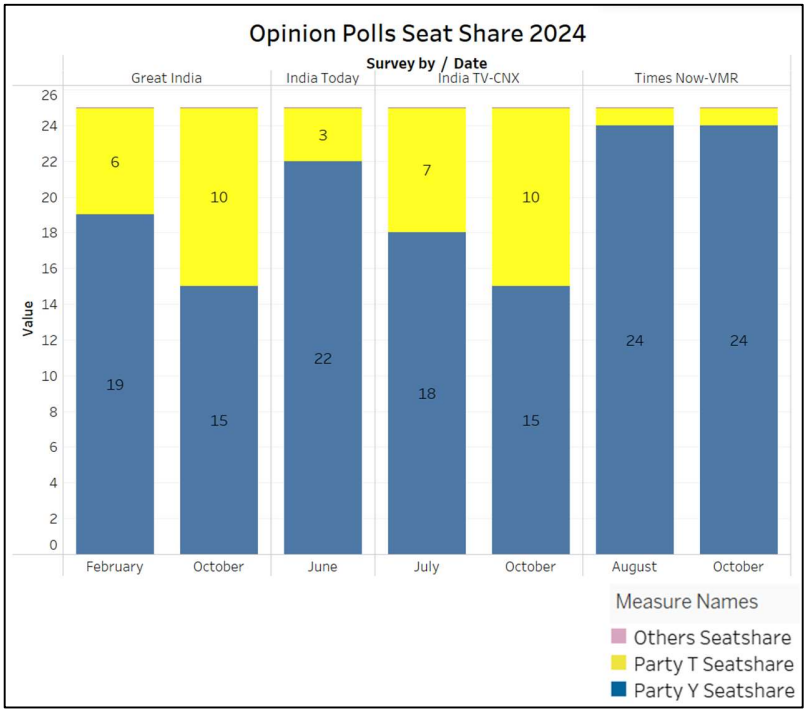

Fig. 6 Predicted seat share by opinion polls for 2024 elections

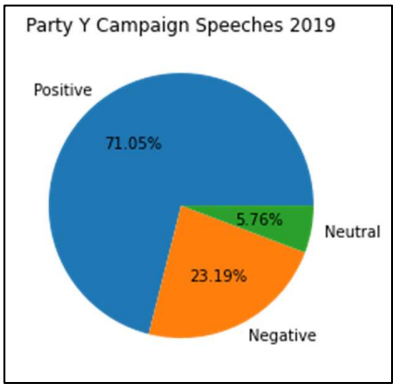

Fig. 7: Sentiment distribution of Party Y Campaign Speeches in 2019

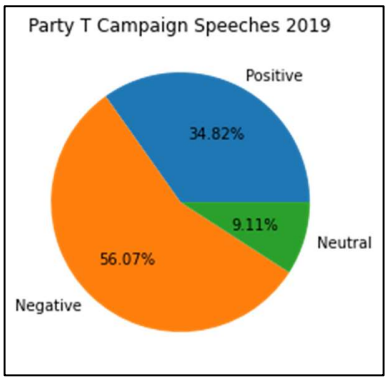

Fig. 8: Sentiment distribution of Party T Campaign Speeches in 2019

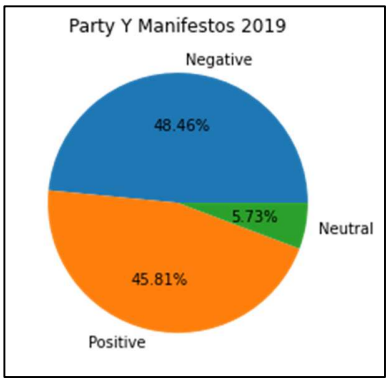

Fig. 9: Sentiment distribution of Party Y Manifestos in 2019

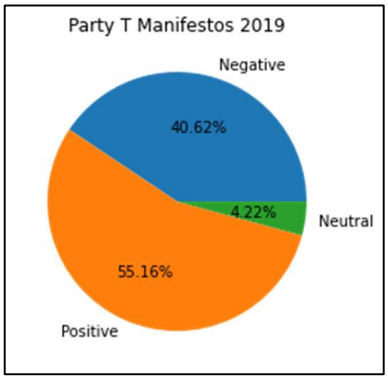

Fig. 10: Sentiment distribution of Party T manifestos in 2019

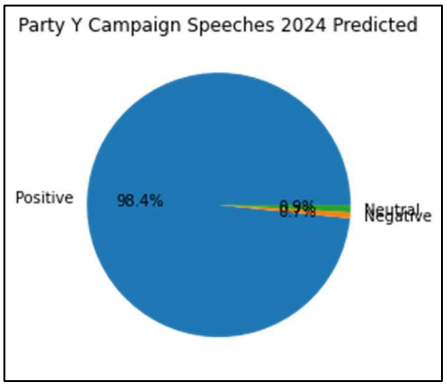

Fig. 11: Sentiment distribution of Party Y Campaign Speeches (2024)

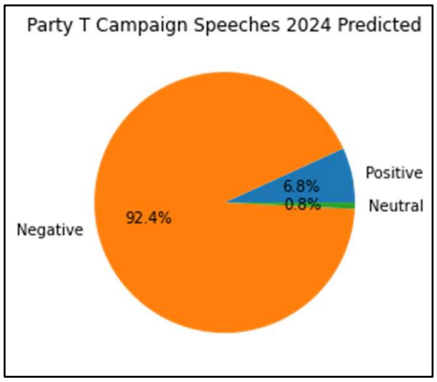

Fig. 12: Sentiment distribution of Party T Campaign Speeches (2024)

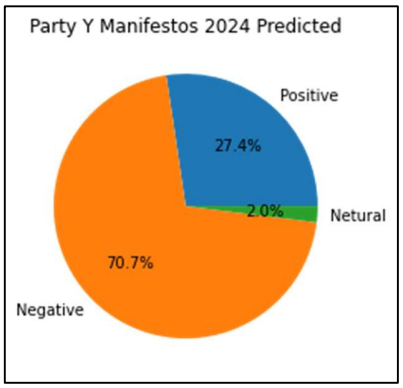

Fig. 13: Sentiment distribution of Party Y Manifestos 2024

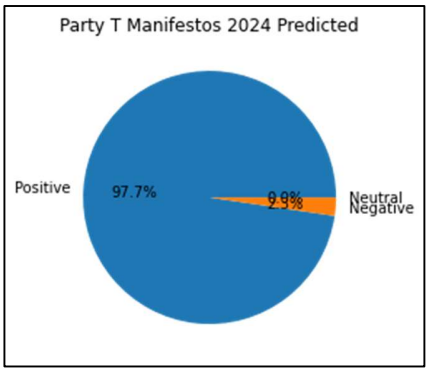

Fig. 14: Sentiment distribution of Party T manifestos 2024

	Positive	**Negative**	**Neutral**
Party Y Campaign Speeches	98.38%	0.72%	0.90%
Party T Campaign Speeches	6.81%	92.41%	0.79%

Table 1: Campaign Speeches Sentiment distribution for 2024

	Positive	**Negative**	**Neutral**
Party Y Manifestos	27.36%	70.68%	1.95%
Party T Manifestos	97.66%	2.34%	0.00%

Table 2: Manifestos Sentiment distribution for 2024

	Party Y	**Party T**
PUBLIC POLLS	78.285714%	21.714286%
CAMPAIGN SPEECHES	95.432598%	4.567402%
MANIFESTOS	15.673048%	84.326952%

Table 3: Prediction for 2024

6. Research challenges

- Irrelevant data, spam comments and social media bots can manipulate the comments will degrades the accuracy of prediction.
- The mood of the voters will change from time to time (known as swing voters). Monitoring the voters in every instance is very hard and challenging task.

7. Conclusion

Predicting election have many traditional ways but in old-fashioned. Nowadays technology has changed a lot of data is transferring through social media in no time. For political, YouTube became one of the common social media. Here people see live political issues with their opinions by giving comments in comment section. This paper aims to provide an overview of combining quantitative data from opinion polls, party manifestos, and campaign speeches with qualitative insights derived from YouTube comments. This way the ever-changing voter swings can be analyzed by collecting periodic data from public polls and can be used for predicting the election results. But still keeping track of swing voters becoming a difficult task. So, predicting the results of elections always becomes challenge and seems to be unexplored.

8. Future Scope

This section provides suggestions for researchers and practitioners to accurately predict election results through statistical and sentimental analysis. In our knowledge, predicting the election results is difficult. Various other sources of data can be used for user comments like twitter, Instagram to get more accurate predictions and also include news reports and current political scenario in state which changes user sentiments. On future study, one can integrate YouTube comments along with twitter tweets for better prediction in election results.

9. References

1. Ardabili, Sina, Amir Mosavi, and Annamária R. Várkonyi-Kóczy. "Advances in machine learning modeling reviewing hybrid and ensemble methods." *International conference on global research and education*. Cham: Springer International Publishing, 2019.

2. Cerón-Guzmán, Jhon Adrián, and Elizabeth León-Guzmán. "A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election." *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (socialcom), sustainable computing and communications (sustaincom)(BDCloud-socialcom-sustaincom)*. IEEE, 2016.

3. Chung, Jessica, and Eni Mustafaraj. "Can collective sentiment expressed on twitter predict political elections?." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 25. No. 1. 2011.

4. Jaidka, Kokil, et al. "Predicting elections from social media: a three-country, three-method comparative study." *Asian Journal of Communication* 29.3 (2019): 252-273.

5. Myilvahanan, Karthick, et al. "A Study on Election Prediction using Machine Learning Techniques." *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. IEEE, 2023.

6. O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Proceedings of the international AAAI conference on web and social media*. Vol. 4. No. 1. 2010.

7. Chauhan, Priyavrat, Nonita Sharma, and Geeta Sikka. "The emergence of social media data and sentiment analysis in election prediction." *Journal of Ambient Intelligence and Humanized Computing* 12 (2021): 2601-2627.

8. Tsakalidis, Adam, et al. "Predicting elections for multiple countries using Twitter and polls." *IEEE Intelligent Systems* 30.2 (2015): 10-17.

9. Tumasjan, Andranik, et al. "Election forecasts with Twitter: How 140 characters reflect the political landscape." *Social science computer review* 29.4 (2011): 402-418.

10. Zhang, Lei, Shuai Wang, and Bing Liu. "Deep learning for sentiment analysis: A survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018): e1253.

11. Zhang, Lei, et al. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." *HP Laboratories, Technical Report HPL-2011* 89 (2011): 1-8.