International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.088

Social Media Analytics For Political Domain

Narayana Darapaneni
Director - AIML
Northwestern university/Great Learning
Bangalore, India
darapaneni@gmail.com

Anwesh Reddy Paduri
Research Assistant - AIML
Great Learning
Bangalore, India
anwesh@greatlearning.in

Sri HarshaUppaluri
Student - AIML
Great Learning
Bangalore, India
sriharsha.caspl@gmail.com

Mukesh S
Student - AIML
Great Learning
Bangalore, India
Smukeshmech@gmail.com

Balasubramanian Petharaj
Student - AIML
Great Learning
Bangalore, India
harish1986@gmail.com

RekhaSuresha
Student - AIML
Great Learning
Bangalore, India
rekhasureshakulal@gmail.com

Manish Chakravarty
Student - AIML
Great Learning
Bangalore, India
manish@fastmail.fm

Pallichadayath Harisankar Nair
Student - AIML
Great Learning
Bangalore, India
harisankar41@gmail.com

***Abstract*—This paper aims to study influence of social media in Indian elections to forecast the mood of the public. Taking the Gujarat election of 2022 as a case study, the paper attempts to gauge, what is the public sentiment in social media (twitter in this case). It attempts to understand public opinion and sentiment towards political figures, parties, policies, and events. It attempts compare the actual election result output with the model prediction. During the study Twitter data was extracted for certain hash tags to extract domain data. The different process like, natural language processing (NLP) and sentiment analysis are used to determine the emotional tone of any tweet relates to a particular hashtag. The goal of sentiment analysis was to classify tweets into different categories such as positive (In favor of ruling party), negative (not in favor of ruling party), or neutral, based on the text of the tweet. Attempts were made to implement a model using Transformer based model for sentiment analysis. The steps for the same is described in this paper.**

***Keywords*—Gujarat Elections 2022, Forecasting, Bagging, Boosting, AFINN, Sentiment Analysis, Natural Language Processing.**

I. INTRODUCTION

Sentiment analysis is the method of identifying the mood or sentiment of a the text content. It classifies the words or combination of words as positive, negative or neutral sentiment in text. It's regularly utilized by businesses to come across sentiment in social facts, gauge emblem recognition, and apprehend customers. It makes a specialty of the polarity of a textual content (advantageous, neutral, bad) but it additionally goes beyond polarity to discover specific feelings and feelings (irritated, glad, unhappy, and so forth), urgency or even intentions[1]. Depending on how we need to interpret customer comments and queries, we can outline and tailor your classes to meet our sentiment evaluation desires.

Social media has become a powerful tool in the political domain, allowing political candidates and parties to connect with voters and promote their agendas. Social media analytics, the practice of analyzing social media data[3] to extract meaningful insights, has become an essential tool for political campaigns. It offers a way to measure sentiment, identify trends, and gain a deeper understanding of voters' preferences and behaviors.

One of the primary applications of social media analytics in the political domain is sentiment analysis. By analyzing social media content, such as tweets, posts, and comments, sentiment analysis can provide insight into how voters feel about a particular candidate or issue. This allows political campaigns to adjust their messaging and outreach strategies accordingly.

Another application is social network analysis, which can identify influential users, communities, and networks. By understanding the structure and dynamics of social networks, political campaigns can target their messaging to key influencers and amplifiers, who can help spread their message to a broader audience.

Social media analytics can also provide insights into voter demographics and behaviors. By analyzing social media data, political campaigns can identify patterns in how voters engage with different types of content, such as videos, images, and articles. This information can be used to optimize content and messaging for maximum impact.

In present times, humans express their feelings and thoughts more freely than ever. As a result, sentiment analysis has become an essential tool for monitoring and identifying sentiments across various forms of data. By automatically analyzing consumer feedback, such as survey responses and social media conversations, businesses can determine what satisfies or frustrates their customers[2]. This facilitates the customization of products and services to better meet customer needs.In this paper, we would be focusing on the sentiment analysis of twitter data on the backdrop of certain Assembly elections in India. We would explore whether the data conveys positive, negative or neutral sentiments towards a political party prior to elections. We would be exploring Machine Learning techniques to understand how to utilize twitter data to analyze the sentiment and give insights for political parties on how to improve their public appeal.

The future scope for social media analytics in the political domain is immense. As social media continues to play an increasingly significant role in politics, the need for advanced analytics tools and techniques will only grow. Here are some potential areas for future development:

DOI: rp-9788770040723.088

Predictive Analytics: Predictive analytics involves using machine learning algorithms to forecast outcomes based on past data. In the political domain, predictive analytics could be used to forecast election results, predict voter behavior, and identify emerging trends.

Real-time Analytics: Real-time analytics allows for immediate insights into social media conversations. In the political domain, real-time analytics could be used to monitor political events, track sentiment, and identify emerging issues.

Personalization: Personalization involves tailoring content and messaging to specific individuals or groups. In the political domain, personalization could be used to target specific demographics, such as age, gender, or location, with customized messaging and content.

Ethical Considerations: As social media analytics becomes more advanced, there will be increasing ethical considerations around the use of personal data[4] and privacy. Political campaigns will need to be transparent about how they are using social media analytics and ensure that they are not violating individuals' rights.

In conclusion, the future scope for social media analytics in the political domain is promising. As technology continues to evolve, new tools and techniques will emerge, providing political campaigns with even more advanced insights into voter sentiment, behavior, and preferences. However, ethical considerations will need to be at the forefront of development to ensure that social media analytics is used responsibly.

II. Literature Review

A. *Ensemble Learning*

Ensemble Learning is a technique based on combining multiple classifier algorithms to produce an output that has greater accuracy in the text classification in comparison to using these algorithms individually for the same task. Ensemble approach is found to improve the accuracy of individual classifiers using methods like Bagging, Boosting, Stacking and Voting.

Bagging: It is short for Bootstrap Aggregating, is a commonly used ensemble method in machine learning that enhances the accuracy of a single model by merging the predictions of multiple models. The Bagging technique works by generating multiple sub-samples of the training data, obtained by randomly selecting instances from the data with replacement, and then training a distinct model on each sub-sample.

Boosting: It is an ensemble learning technique in which multiple weak learners are combined to create a single strong learner. Unlike bagging, which creates independent weak learners, boosting uses a sequential approach in which each weak learner is trained to focus on the mistakes of the previous weak learners. The final prediction is made by combining the predictions of all the weak learners, with more weight given to the predictions of the stronger learners.

Stacking: It is an ensemble learning technique that combines the predictions of multiple base models using a meta-model, or "stacking model". The basic idea behind stacking is to train several diverse base models that have different strengths and weaknesses, and then use a meta-model to combine their predictions. The stacking process involves splitting the training data into two or more sets, where the base models are trained on one subset of the data and the meta-model is trained on another subset. The base models can be of different types and can use different learning algorithms, while the meta-model is typically a simple model, such as linear regression or logistic regression, that is trained on the predictions of the base models.

Voting: It is an ensemble learning technique that combines the predictions of multiple base models by taking a majority vote. The idea behind voting is to train multiple diverse base models, each of which makes predictions on a given set of inputs, and then use a voting scheme to combine their predictions into a final prediction. In binary classification problems, voting can be done by taking a simple majority vote, where the class that receives[5] the most votes is chosen as the final prediction. In multi-class classification problems, voting can be done by taking a plurality vote, where the class with the highest number of votes is chosen as the final prediction.

B. *Deep Learning*

Deep learning is a subset of machine learning that is characterized the use of neural networks with multiple layers. In traditional machine learning approaches, feature engineering is often a critical step, where features are manually defined and extracted from the input data. However, in deep learning models, features are automatically learned and extracted from the data during the training process. This is achieved by passing the data through multiple layers of nonlinear transformations, which allows the model to learn more complex and abstract representations of the input data.

Deep Neural Networks (DNN): This is a type of artificial neural network (ANN) that incorporates multiple hidden layers between the input[6] and output layers. These hidden layers enable the network to learn intricate and abstract representations of the input data. A typical DNN architecture includes an input layer, one or more hidden layers, and an output layer. Each layer comprises a group of interconnected neurons that are linked to the neurons in the adjacent layers. The neurons perform nonlinear transformations on their inputs using activation functions, generating outputs that are transmitted to the next layer.

Convolutional Neural Networks (CNN): This is type of deep neural network that is particularly suited for video and image recognition tasks. It is designed to automatically learn spatial feature layers from raw inputs such as pixels in images. CNNs are designed to automatically learn and extract features from raw image data without manual feature engineering. A key feature of CNNs is that they use convolutional layers. A convolutional layer consists of a set of filters[7], each filter being a small matrix of weights. These filters are applied to the input image in sliding window mode to compute the dot product between the filter and each local input region. The result of the convolution operation is a series of feature maps representing the presence of different patterns or features in the input image.

DOI: rp-9788770040723.088

Recurrent Neural Networks (RNN): Recurrent Neural Networks (RNNs) are a specialized type of deep neural network that excel at processing sequential data, such as time series or natural language text. Unlike feedforward neural networks that treat each input independently, RNNs maintain an internal state that enables them to capture the temporal dependencies of the input sequence. The distinctive characteristic of RNNs is their utilization of recurrent connections, which enable information to be propagated from one time step to the next. At every time step, the input is combined with the prior hidden state to produce a new hidden state. This operation is executed for each time step in the sequence, and the final hidden state is utilized to generate a prediction. This concept was introduced by Hochreiter and Schmidhuber in 1997.

C. Transformer Architecture

Transformer-based models have revolutionized the field of natural language processing (NLP) by outperforming previous state-of-the-art models on a wide range of NLP tasks, including sentiment analysis. In a transformer-based model, the input text is first transformed into a sequence of embeddings, which capture the meaning of each word in the text. The embeddings are then processed by a series of transformer layers, which learn to model the relationships between the words and their context.

During training, the model is fed a large dataset of text that is labeled with their corresponding sentiment. The model learns to predict the sentiment of the text by adjusting the weights of its parameters to minimize a loss function. Once the model is trained, it can be used to predict the sentiment of new text by feeding it through the same process of embedding and processing. The output of the model is a probability distribution over the possible sentiments, which can be used to determine the most likely sentiment. Overall, transformer-based models have shown to be highly effective in sentiment analysis and have achieved state-of-the-art results on several benchmark datasets.

A pre-trained model named "distilbert-base-uncased-finetuned-sst-2-english" was used.DistilBERT-base-uncased-finetuned-SST-2-English is a pre-trained language model developed by Hugging Face that is based on the BERT architecture. It has been fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset for English language text classification. The model uses a smaller and more efficient architecture compared to the original BERT model, which makes it faster and easier to deploy on devices with limited computational resources.

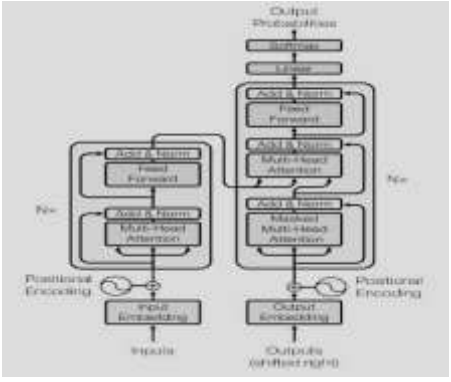

Fig1. Steps in Transformer Architecture

Advantages of Transformer Architecture

Large Vocabulary: Transformer models can learn to understand a much larger vocabulary, which allows them to capture the sentiment of text that contains words that are not included in a pre-defined word list.

Context Awareness: Transformers can consider the context in which words are used, which can lead to more accurate sentiment analysis.

Cultural Relevance: Transformer models can be trained on text from multiple cultures and languages, making them more suitable for sentiment analysis in a global context.

Up-to-date: As transformer models can be re-trained on new data, they can adapt to changes in language usage and sentiment over time.

Sophisticated Methodology: Transformers are a complex type of deep learning architecture that have demonstrated exceptional performance in numerous Natural Language Processing (NLP) tasks, such as sentiment analysis, surpassing previous state-of-the-art results.

III. Materials and Methods

A. Dataset

With the advent of a wide range of social media applications, there are millions of data generated which involves a wide range of topics across the globe. It is observed that people are more vocal about their opinions on the social media platform. Twitter is one such platform which serves as bridge between common people and the authorities involved in the governing body.

Recently we have seen surge in the usage of twitter by political parties to orchestrate their political campaigns and to reach out to people to every corner in the country. People are becoming more and more vocal about their political opinions which gives us an immense opportunity to mine these data points and try to arrive at some conclusion based on people sentiments.

This research involves the study of Twitter data associated with Indian State elections conducted during 2023.The study aims to analyze and understand the sentiments and opinions of people through their twitter comments.

B. Model Building : Steps

1. Data extraction using Twitters APIs
2. Data understanding through data dictionary
3. Tokenization
4. Data preprocessing
5. Data Encoding
6. Sentiment Analysis using Transformer Architecture (DistilBERT base uncased finetuned SST-2)

C. Sample Dataset

As for initial model, sample dataset containing 1212 tweets were taken which are associated with Gujarat State election. The dataset used is extracted from Twitter using the hashtag #gujaratelections2022

DOI: rp-9788770040723.088

Fig. 2. Sample Tweets

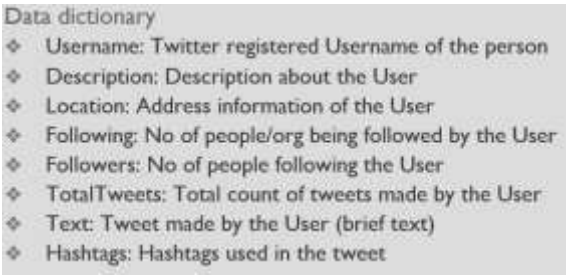

Fig. 3. Standard Data Dictionary for Twitter Extract

D. Data Pre-processing

The twitter data is unstructured and it contains a lot of noise and unwanted information. Data pre-processing is required to filter out such noise and unwanted information. Further, it simplifies the vocabulary used for analyzing sentiments from the tweets. The model will use different data prepossessing techniques such as Tokenization, Lower case conversion, Stop words removal, Punctuation removal,Stemming,Lemmatization and Parts of Speech Tagging.

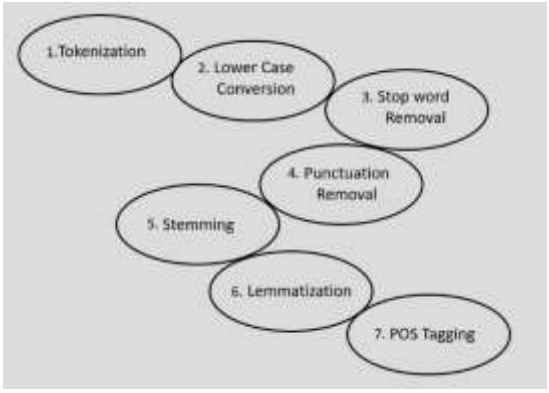

Fig. 4. Data Pre-Processing Steps

Encoding Techniques: Encoding techniques refer to methods used to represent data in a format that can be easily processed by a computer system. Techniques like Bag of Words, Binary Bag of Words, Bigram, Ngram, TF-IDF(Term Frequency-Inverse Document Frequency) are being utilized in this study.

Fig. 5 Encoding Steps Used

IV. RESULTS

This model reaches an accuracy of 91.3 percent on the development corpus from which it was trained. This reduces our training time drastically instead of training the model fresh from start using our data. The model output shows that the sentiment of the people is positive towards the ruling party in the State of Gujarat during 2022 India State Elections.

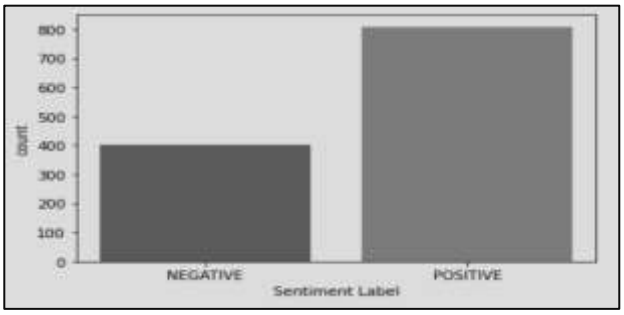

Fig. 6. Sentimental Labelling

The bar graph shows the distribution of the tweets after applying the sentiment analysis using Transformer model (distilbert-base-uncased-finetuned-sst-2-english). Majority of the comments in this dataset convey positive sentiment towards the ruling dispensation.

On analyzing the most frequently occurring words in the set of positive tweets, we found that there is a generic positive sentiment towards ‘amitshah’ (as shown in word clouds below).

Fig. 7.Positive (In favor of ruling Party) tweets WC

The above word cloud captures the basic sentiments of the Tweets, that were identified as supporting the Ruling dispensation prior to the 2022 election. The words like Choose and Amit Shah were repeated multiple times.

DOI: rp-9788770040723.088

Fig. 8. Negative (Against ruling Party) tweets WC

The above word cloud captures the basic sentiments of the Tweets, that were identified as opposing the Ruling dispensation prior to the 2022 election. The words like change and Shah ignored were repeated multiple times here. However, the word cloud for Negative comments also have used words like Amit Shah Magic and Shah for Gujarat, which in a way also conveys positive sentiment. Hence Word Cloud is not a apt representation for Negative Sentiments as such.

V. Discussions and Conclusions

A. Conclusion

Through this study it is concluded that Transformers models have been shown to be very effective in performing sentiment analysis of social media data. Social media platforms like Twitter, Facebook, and Instagram generate huge volumes of text data that contain a wide range of sentiments, opinions, and emotions. Transformer models excel in handling this type of unstructured and noisy text data because they can capture complex relationships between words and their context, as well as handle variations in sentence structure and grammar.

The results indicate high positive sentiment towards ruling political party.The influence of Third-Party Social Media Agencies and paid tweets has the potential of introducing a bias in the final outcomes. Techniques and Methods to identify and eliminate this bias is an opportunity, that can be considered as a future scope.

References

[1] Troussas, Christos, Akrivi Krouska, and Maria Virvou, "Evaluation of ensemble-based sentiment classifiers for Twitter data," 2016 7th international conference on information, intelligence, systems & applications (IISA). IEEE, 2016.

[2] Dang, Nhan Cach, María N. Moreno-García, and Fernando De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics, vol. 9.3, p. 483, 2020.

[3] Jose, Rincy, and Varghese S. Chooralil. "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," 2016 international conference on data mining and advanced computing (SAPIENCE). IEEE, 2016.

[4] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ... & Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis. Computers and Electrical Engineering, 106, 108556.

[5] Athanasiou, Vasileios, and Manolis Maragoudakis. "A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek." Algorithms, vol. 10.1,p. 34, 2017.

[6] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., & Manoharan, R. (2021). Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[7] Asghar, and Muhammad Zubair, et al. "A review of feature extraction in sentiment analysis," Journal of Basic and Applied Scientific Research , vol.4.3, pp. 181-186, 2014.