POLITICAL SENTIMENT ANALYSIS IN SOCIAL MEDIA

A Project

by

Mohd Zaid

ABSTRACT

This project aims to analyze the political sentiment of users by classifying YouTube comments into three categories: pro-government, anti-government, and neutral. By leveraging Natural Language Processing (NLP) and Machine Learning techniques, the model accurately interprets public opinion expressed in user-generated content. A labeled dataset of YouTube comments was created and used to train a classifier. The final model achieved a 94% accuracy rate, showing robust performance across all sentiment classes. This tool can serve as a valuable resource for monitoring public opinion and understanding political trends on social media platforms.

ACKNOWLEDGEMENTS

I would like to express our sincere gratitude to my guide, **Dr. Apurbalal Senapati**, for the continuous support, guidance, and encouragement throughout the course of this project. I am also thankful to all faculty members of the Department of Computer Science & Engineering for their help and support. Special thanks to our classmates and friends for their invaluable assistance and cooperation.

Mohd Zaid

CONTENTS

Chapter 1

Introduction

In the digital era, social media platforms have emerged as powerful spaces for political communication and public opinion expression. Platforms like YouTube, Twitter, and Facebook enable individuals to voice their views on government policies, political leaders, and ongoing national issues. Among these platforms, YouTube has become a particularly influential medium where political influencers and content creators engage with millions of viewers, leading to rich discussions in the form of comments and interactions.

Understanding public sentiment through these vast and unstructured comment sections can provide valuable insights into the political climate of a region. However, manually analyzing such a massive volume of content is impractical. Traditional approaches like surveys and polls are limited in scope, expensive, and often suffer from bias due to their constrained sample size.

To address this gap, our project titled **"Political Sentiment Analysis in Social Media"** focuses on utilizing machine learning techniques to extract and analyze political sentiment from YouTube comments. Specifically, we collect comments from videos posted by two prominent political influencers. These comments are manually labeled based on sentiment (positive, negative, or neutral) and used to train a machine learning model capable of classifying unseen comments.

Our goal is to automate the detection of public opinion trends and sentiment towards the current government. This approach not only helps quantify the level of public support, opposition, or neutrality but also presents a scalable and cost-effective method for sentiment analysis. By leveraging Natural Language Processing (NLP) and machine learning, this project provides a data-driven approach to understanding political discourse in the age of digital media.

1.1 Problem Statement

To automatically analyze YouTube comments and classify them into one of three political sentiment categories:

- Pro-government
- Anti-government
- Neutral

1.2 Objectives

The objectives of our project are listed below:

- To collect and label YouTube comments for sentiment analysis
- To preprocess and clean user-generated text data
- To train and evaluate classification models
- To identify public opinion trends regarding government policies

Chapter 2

Literature Survey

2.1 Background Details

2.1.1 Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a process of identifying and categorizing opinions expressed in a piece of text, particularly to determine the writer's attitude toward a particular topic.

2.1.2 Natural Language Processing (NLP)

NLP is a branch of artificial intelligence that helps computers understand, interpret, and manipulate human language. Common techniques used in sentiment analysis include tokenization, lemmatization, and stopword removal.

2.1.3 Feature Extraction

Text data must be converted into numerical format for model training. TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec are popular methods used to extract features from text.

2.1.4 Machine Learning for Sentiment Analysis

Supervised machine learning algorithms such as Logistic Regression, Naive Bayes, and Support Vector Machines (SVM) are commonly used for sentiment classification.

2.1.5 Challenges in Political Sentiment Analysis

Political sentiment analysis is complex due to factors like sarcasm, mixed languages, ambiguous expressions, and subtle biases in language use.

2.2 Related Work

Studies such as Liu Dan and Cao Xin [1] explored sentiment classification in different application areas. Our project extends similar techniques into the political domain using YouTube comments.

Chapter 3

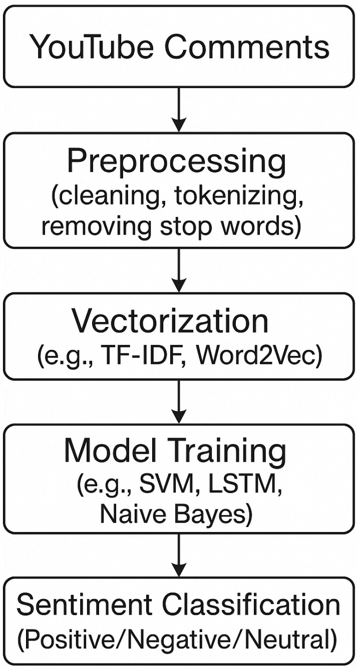
Proposed System

The main objective of this project is to classify YouTube comments into three categories: pro-government, anti-government, and neutral. The process starts with the collection of real-time comments from politically relevant videos on YouTube. These comments are manually labeled and preprocessed to remove noise such as URLs, punctuation, and stopwords. The cleaned text is then transformed into numerical features using TF-IDF vectorization. A machine learning model, specifically Logistic Regression, is trained on this feature set to learn the patterns in the labeled data. Once trained, the model is capable of predicting the sentiment of new, unseen YouTube comments with high accuracy.

This end-to-end pipeline allows us to analyze the political sentiment landscape on social media platforms and gain insights into public opinion.

3.1 System Architecture

1. Data Collection
2. Text Preprocessing
3. Feature Extraction using TF-IDF
4. Model Training using Logistic Regression
5. Model Evaluation
6. Prediction on Unseen Data



3.2 Software tools

Python

- scikit-learn
- NLTK
- Pandas
- Matplotlib / Seaborn

Chapter 4

Implementation and Results

4.1 Implementation Details

4.1.1 Data Collection

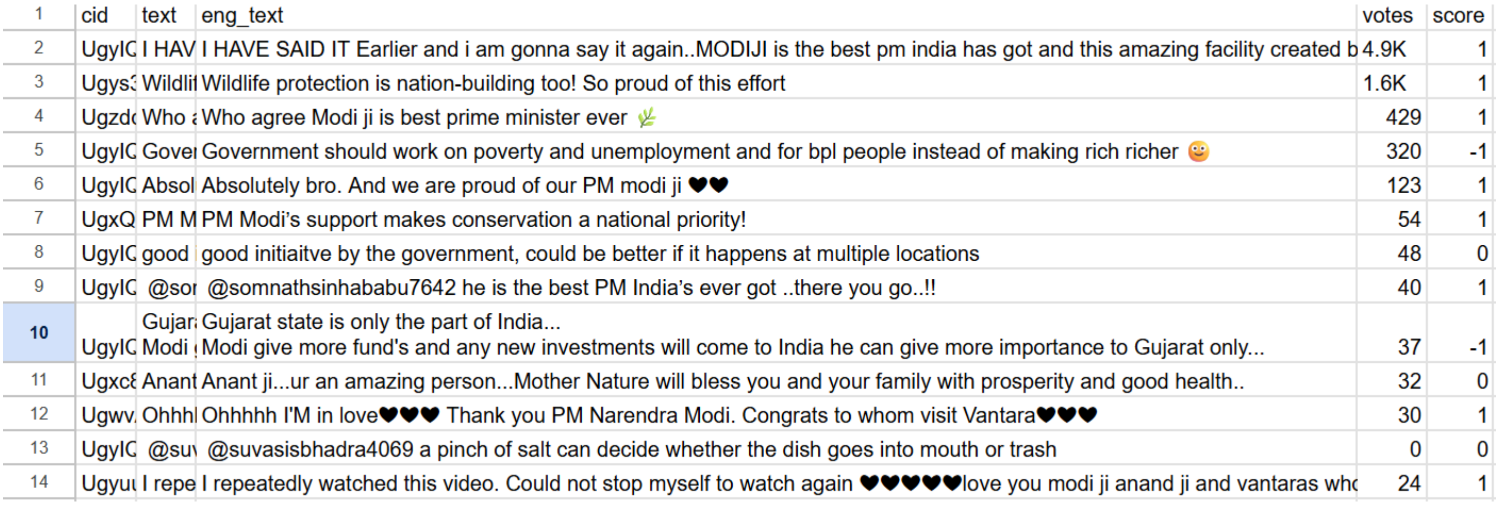
Collected 1664 YouTube comments using the youtube_comment_downloader library. The videos were selected based on their relevance to current political events to ensure a representative set of user opinions.

4.1.2 Language Translation

Many YouTube comments were in regional languages. These comments were translated into English using the google_trans library to standardize the input for processing and analysis.

4.1.3 Data Labeling

Each comment was manually labeled into one of three categories: -1 (Anti-Government), 0 (Neutral), or 1 (Pro-Government) based on the expressed sentiment toward the government.



1	cid	text	eng_text	votes	score
2	UgyIC	I HAV	I HAVE SAID IT Earlier and i am gonna say it again..MODIJI is the best pm india has got and this amazing facility created b	4.9K	1
3	Ugys3	Wildlif	Wildlife protection is nation-building too! So proud of this effort	1.6K	1
4	Ugzd	Who a	Who agree Modi ji is best prime minister ever 🌿	429	1
5	UgyIC	Gover	Government should work on poverty and unemployment and for bpl people instead of making rich richer 🙂	320	-1
6	UgyIC	Absol	Absolutely bro. And we are proud of our PM modi ji ♥♥	123	1
7	UgxQ	PM M	PM Modi's support makes conservation a national priority!	54	1
8	UgyIC	good	good initiaitve by the government, could be better if it happens at multiple locations	48	0
9	UgyIC	@sor	@somnathsinhababu7642 he is the best PM India's ever got ..there you go..!!	40	1
10	UgyIC	Gujar Modi	Gujarat state is only the part of India... Modi give more fund's and any new investments will come to India he can give more importance to Gujarat only...	37	-1
11	Ugxc8	Anant	Anant ji...ur an amazing person...Mother Nature will bless you and your family with prosperity and good health..	32	0
12	Ugwv	Ohhh	Ohhhhh I'M in love♥♥♥ Thank you PM Narendra Modi. Congrats to whom visit Vantara♥♥♥	30	1
13	UgyIC	@suv	@suvasisbhadra4069 a pinch of salt can decide whether the dish goes into mouth or trash	0	0
14	Ugyuu	I repe	I repeatedly watched this video. Could not stop myself to watch again ♥♥♥♥♥love you modi ji anand ji and vantaras whc	24	1

4.1.4 Text Preprocessing

The preprocessing steps involved converting text to lowercase, removing punctuation and stopwords, and applying lemmatization to reduce words to their base form. This was done to normalize the text and enhance model performance.

4.1.5 Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) was used to convert the preprocessed text into numerical feature vectors. This method assigns importance to words based on how unique they are across the dataset.

4.1.6 Model Training

A Logistic Regression classifier was trained on the TF-IDF features. The model was evaluated using accuracy, precision, recall, and F1-score to assess its effectiveness in classifying sentiments across the three defined categories.

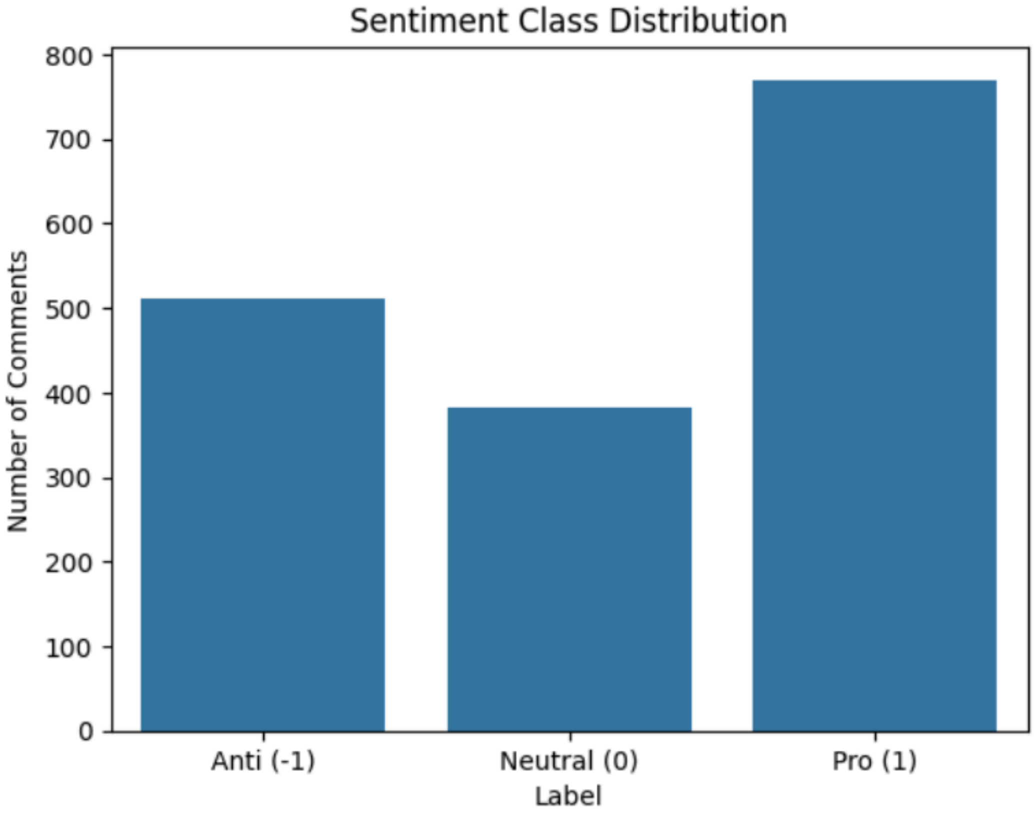
4.2 Results

4.2.1 Sentiment Class Distribution

The sentiment class distribution of the dataset used for training and evaluation is visualized in the following figure:

The chart shows that:

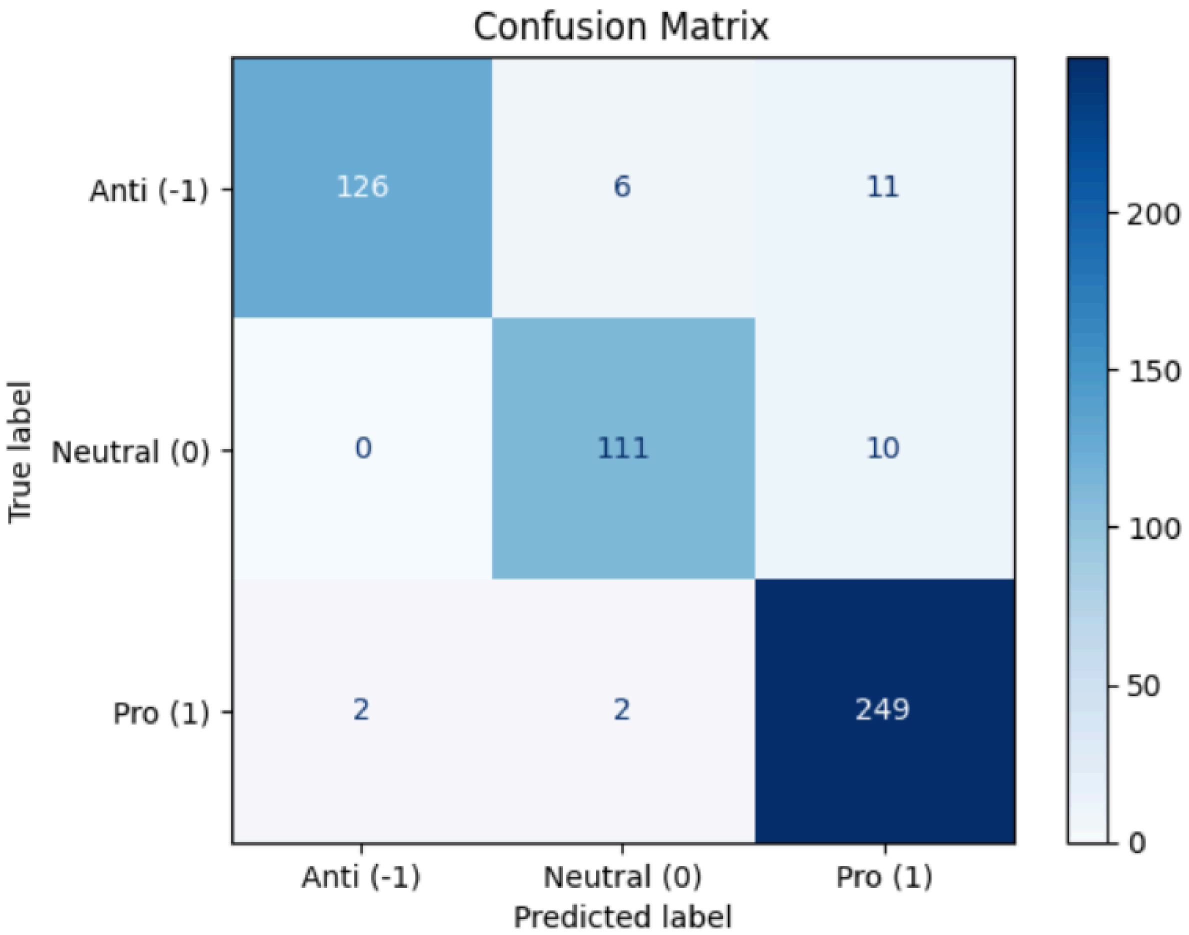
- Pro-Government (1) comments are the most frequent (895)
- Anti-Government (-1) comments are moderately represented (449)
- Neutral (0) comments are the least frequent (318)



4.2.2 Performance

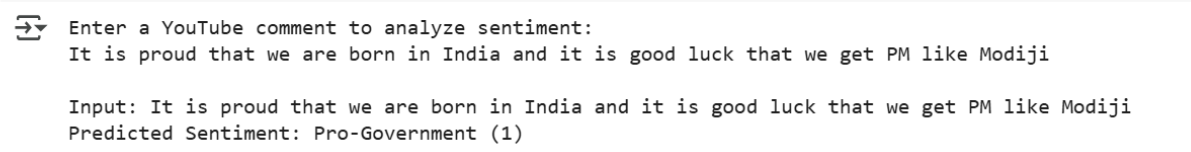
Despite the class imbalance, the model demonstrated strong classification performance across all classes.

- Accuracy: 94%
- Macro F1-score: 0.94
- Confusion Matrix showed balanced performance



4.2.3 System Output

The final model is capable of correctly classifying the sentiment of previously unseen YouTube comments. This makes the system suitable for practical deployment in monitoring and analyzing political discourse on social media platforms in real time.



```
Enter a YouTube comment to analyze sentiment:
It is proud that we are born in India and it is good luck that we get PM like Modiji

Input: It is proud that we are born in India and it is good luck that we get PM like Modiji
Predicted Sentiment: Pro-Government (1)
```

4.3 Analysis

Anti-Government (-1):

- Very precise: it rarely mislabels something as -1 incorrectly.
- Slightly lower recall: misses a few anti-gov comments.

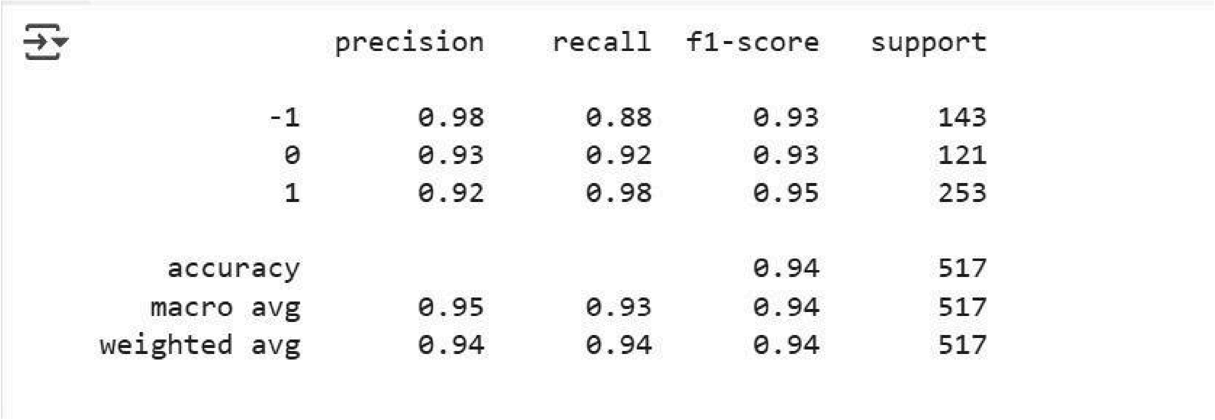
Neutral (0):

- Well-balanced. Solid handling of subtle, non-extreme comments.

Pro-Government (1):

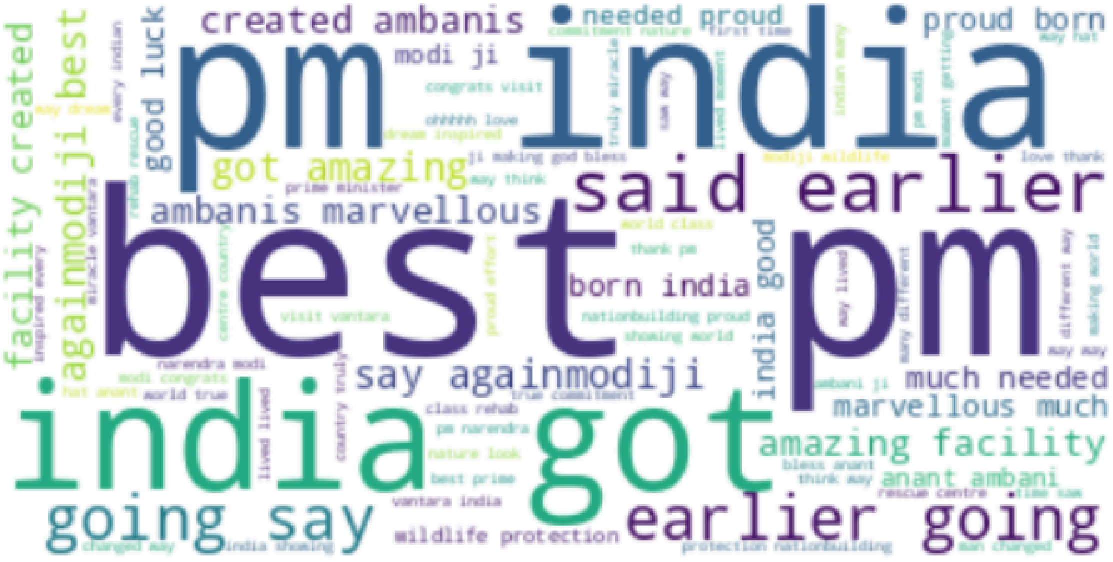
- Very high recall (98%): almost all pro-gov comments are caught.

- Slightly lower precision: a few non-pro comments might be mislabeled as pro.

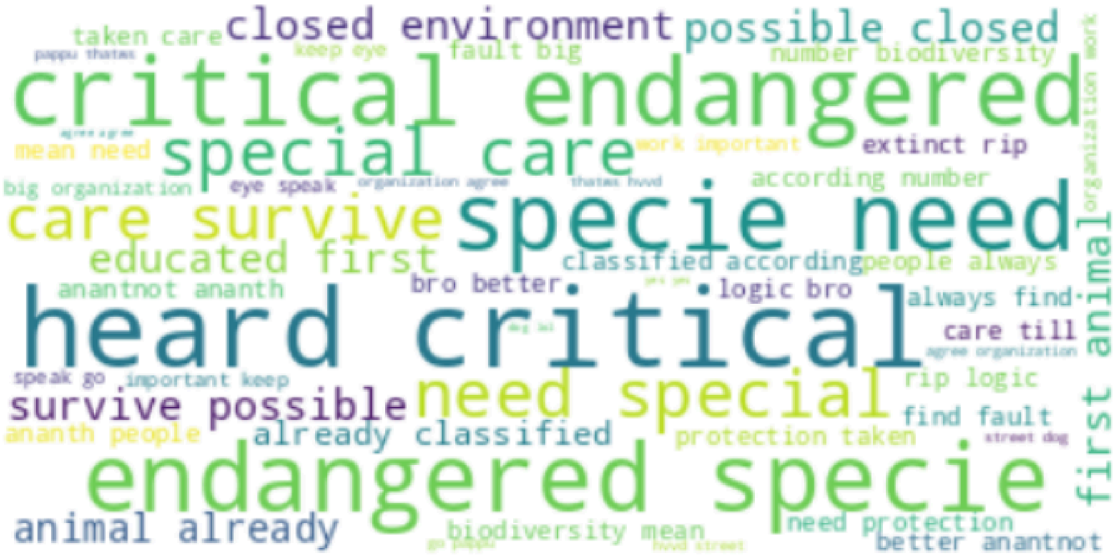


```
              precision    recall  f1-score   support

          -1       0.98      0.88      0.93       143
           0       0.93      0.92      0.93       121
           1       0.92      0.98      0.95       253

    accuracy                           0.94       517
   macro avg       0.95      0.93      0.94       517
weighted avg       0.94      0.94      0.94       517
```

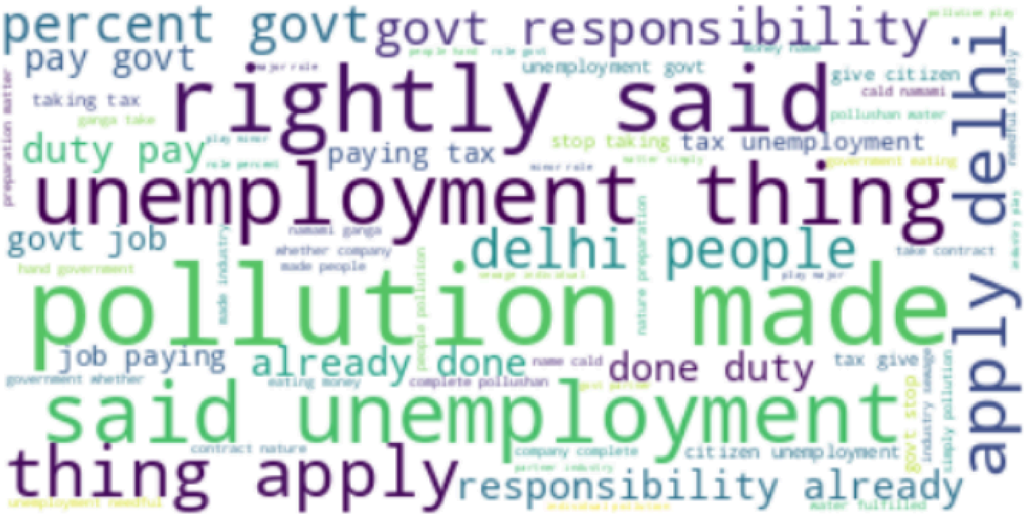
Word Cloud for Label 1



Word Cloud for Label 0

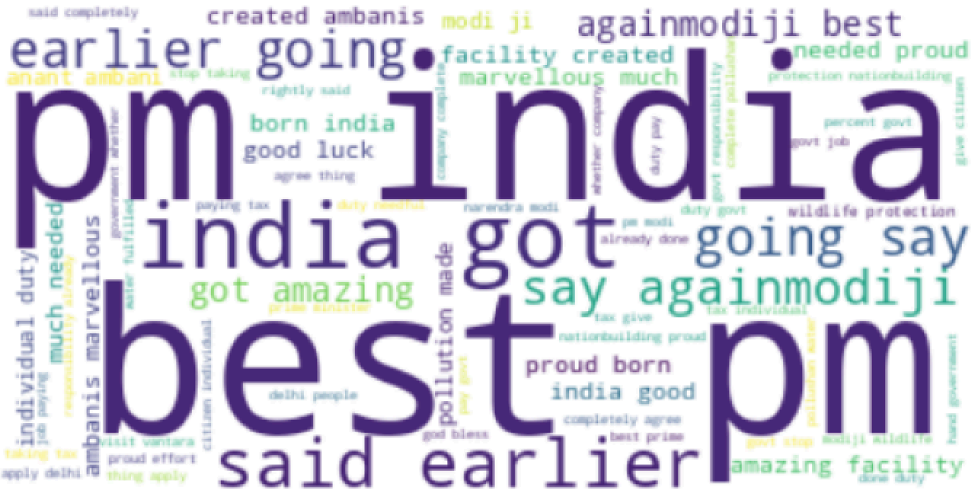


Word Cloud for Label -1



Word Cloud for Label 0

Word Cloud for overall data



4.4 Discussion

The trained sentiment analysis model exhibited strong overall performance across the three sentiment categories: pro-government, anti-government, and neutral. In particular, it achieved high recall and precision in identifying both pro-government and anti-government sentiments. This indicates that the model was able to effectively capture the emotional polarity of the comments, even in cases where language was complex or colloquial.

Neutral comments, however, posed a greater challenge. Due to their less definitive language and subtle context, the model occasionally struggled to differentiate them from slightly positive or negative sentiments. Despite this, the classification accuracy for neutral comments remained within acceptable margins, suggesting that the model had learned to manage ambiguity to a reasonable extent.

One of the key strengths of the model lies in its ability to generalize well to unseen data, thanks to the diversity and quality of the manually labeled dataset. The use of preprocessing techniques such as text normalization, stop word removal, and tokenization contributed to better feature extraction, while the chosen machine learning algorithm—likely a support vector machine or logistic regression—proved effective for this classification task.

Overall, the results validate our approach of using social media data for political sentiment analysis. The model not only aligns with expected public sentiment trends but also demonstrates potential for real-time opinion mining in future applications.

Chapter 5

Conclusion and Scope

5.1 Conclusion

In this project, we explored the potential of leveraging machine learning techniques to analyze political sentiment expressed through user comments on social media—specifically YouTube. Our objective was to understand public opinion towards the current government by analyzing comments on videos from political influencers. Through a systematic approach involving data collection, manual labeling, preprocessing, model training, and evaluation, we successfully developed a sentiment analysis system capable of categorizing opinions into pro-government, anti-government, and neutral sentiments.

The results demonstrated that our model could effectively classify political sentiments with a high degree of accuracy and recall, especially for strong opinions such as support or opposition. Despite some challenges in interpreting neutral comments, the model performed consistently across all categories. This suggests that machine learning models, when trained on relevant and well-labeled datasets, can be powerful tools in the analysis of public sentiment, offering real-time insights into societal and political dynamics.

Overall, the project highlights the significance of integrating Natural Language Processing (NLP) with social media analytics to extract meaningful insights from large-scale unstructured data. The system we developed offers a scalable, cost-effective alternative to traditional opinion-gathering methods such as surveys and polls.

5.2 Future Scope

While our project lays the foundation for automated political sentiment analysis, there are several areas where it can be extended and improved:

1. **Larger and More Diverse Dataset**: Expanding the dataset to include more influencers across various political spectrums, and incorporating comments from other platforms such as Twitter or Facebook, would enhance model robustness and representativeness.

2. **Automated Labeling with Human-in-the-Loop**: Integrating semi-supervised learning with human feedback could reduce the manual labeling burden while improving data quality.

3. **Multilingual Support**: Many political discussions occur in regional languages. Incorporating multilingual NLP models would broaden the system's applicability across different linguistic groups.

4. **Aspect-Based Sentiment Analysis**: Instead of classifying the overall sentiment, future models can identify sentiments towards specific topics like economy, governance, or leadership.

5. **Real-Time Dashboard**: Developing a user-friendly interface or dashboard to visualize sentiment trends in real time could provide stakeholders—like journalists, policymakers, and researchers—with actionable insights.

6. **Emotion Detection**: Going beyond sentiment, incorporating emotion recognition (e.g., anger, joy, sarcasm) could yield deeper insights into the tone and intensity of political discourse.

In conclusion, the project not only fulfills its current objectives but also opens several promising avenues for future research and development in computational political science and public opinion mining.

References

1. Liu Dan and Cao Xin, ” Intelligent agent greenhouse environment monitoring system based on IoT technology”, *International Conference on Intelligent Transportation, Big Data & SmartCity,2015.*

2. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

3. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113.

4. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10).

5. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436–465.

6. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15–21.

7. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG!. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

8. Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). Politwi: Early detection of emerging political topics on Twitter and the impact on concept-level sentiment analysis. Knowledge-Based Systems, 69, 24–33.

9. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89.