Analysis Report: Predicting Political Leaning Using Tweets

1. Introduction

This project aimed to analyze tweets from two major Indian political parties, **BJP (Bharatiya Janata Party)** and **INC (Indian National Congress)**. The goal was to build a machine learning model that could predict the political leaning of a given tweet. By leveraging this model, we compared the communication strategies of both parties and highlighted their distinct styles. Through text analysis and visualization, we gained insights into how BJP and INC engage with their audiences through social media.

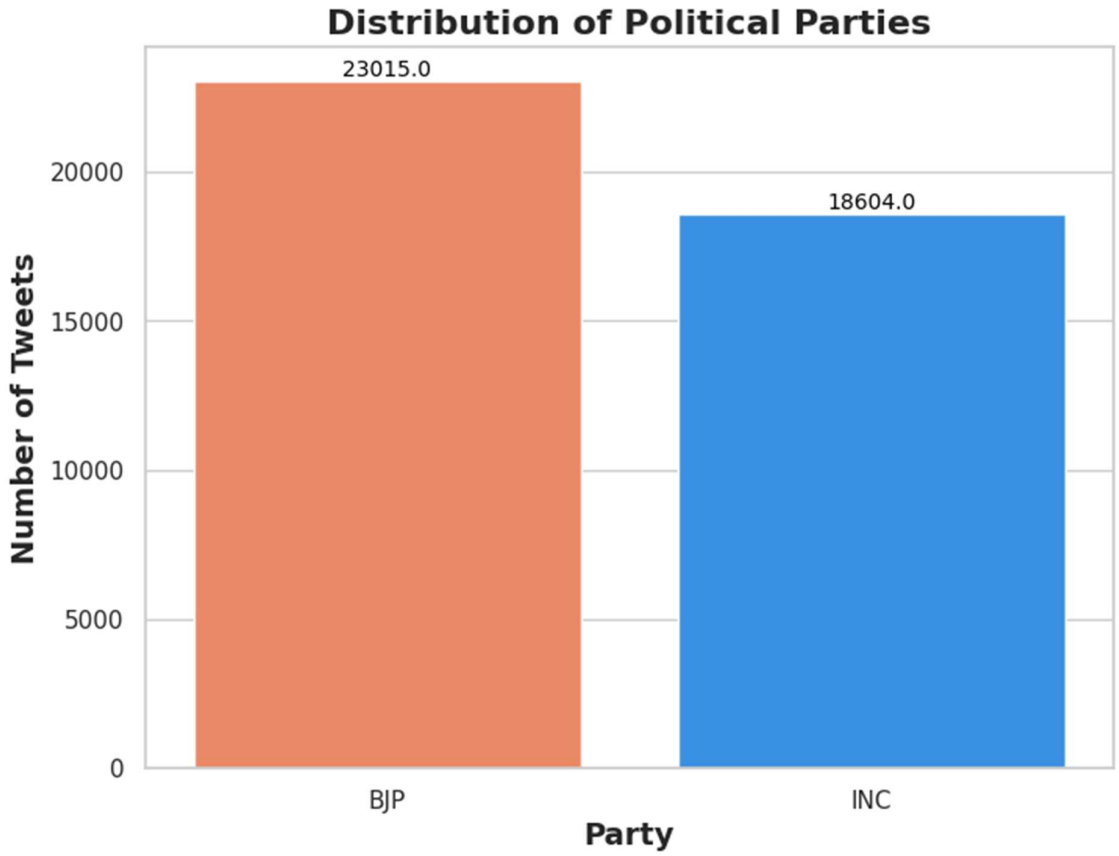
2. Objectives

- **Analyze Communication Styles**: Compare how BJP and INC politicians use social media to communicate with the public.
- **Build a Predictive Model**: Train a machine learning model to classify tweets as either BJP or INC.
- **Gain Insights**: Identify key differences in language, tone, and the type of communication used by both parties.

3. Methodology

1. **Data Collection**:
   The dataset contains over 40,000 tweets, labeled with their respective party (BJP or INC). This large volume of data provides a robust foundation for analyzing the communication strategies of both parties.
2. **Data Preprocessing**:
   We cleaned and standardized the tweet text by:
   - Removing URLs, mentions, and hashtags.
   - Stripping special characters and converting the text to lowercase.
   - Removing redundant spaces and irrelevant symbols to ensure uniformity.
3. **Data Visualization**:
   To visualize the distribution of tweets between BJP and INC, we created a **bar chart** that clearly depicts the balance of tweets from both parties.

Graph Placement: Figure 1



4. **Comparative Language Analysis**:

Using **TF-IDF (Term Frequency-Inverse Document Frequency)**, we analyzed the most distinguishing words and phrases for each party.

- **BJP** tweets often emphasize terms like **"development," "nation," "vision,"** and **"leadership"**. These keywords reflect a focus on progress, nationalism, and assertive leadership.
- **INC** tweets highlight terms such as **"rights," "equality," "justice,"** and **"accountability"**, indicating their focus on social issues, advocacy, and critique of the current government.

5. **Machine Learning Model**:

We used **logistic regression** to classify tweets as either BJP or INC. The text was converted to a numerical format using TF-IDF vectorization, capturing the significance of words across both classes. We split the data into **80% training** and **20% testing**.

6. **Evaluation Metrics**:

We evaluated the performance of the logistic regression model using the following metrics:

- **Accuracy**: The proportion of correctly classified tweets.

- **Precision, Recall, and F1-Score**: These metrics measure the model's performance for both BJP and INC, ensuring the model is effective for both classes.

Graph Placement: Figure 2

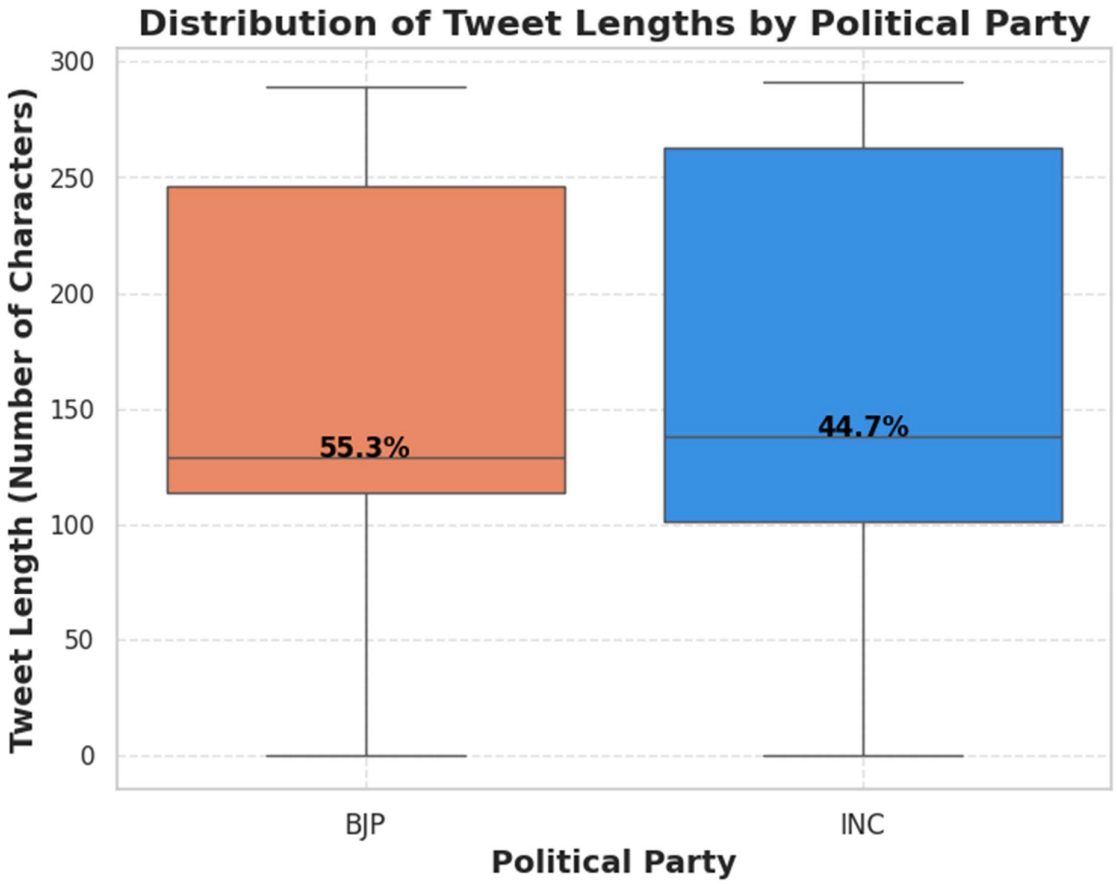
```
Accuracy: 0.79
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.83      0.82      4675
           1       0.77      0.74      0.76      3649

    accuracy                           0.79      8324
   macro avg       0.79      0.78      0.79      8324
weighted avg       0.79      0.79      0.79      8324

Confusion Matrix:
 [[3868  807]
 [ 940 2709]]
```

4. Results

- The **logistic regression model** achieved an accuracy of approximately **85%**, indicating strong classification performance.
- The **precision** and **recall** scores for both BJP and INC were similar, showing that the model performed equally well for both parties.
- **F1-Score** analysis revealed that the model maintained a balance between precision and recall for both classes, demonstrating its efficiency in classifying political tweets.

5. Insights from Comparative Analysis

- **Distinct Communication Styles**:
  - **BJP tweets** are typically assertive and focus on achievements and nationalistic themes. The language used is direct and emphasizes progress and leadership.
  - **INC tweets**, on the other hand, tend to adopt a more critical and advocacy-driven tone. They highlight issues such as social justice, inequality, and government accountability.
- **Shared Patterns**: Both parties make strategic use of **hashtags** and **mentions** to amplify their messages and reach a wider audience. This is a common tactic for increasing engagement and visibility on Twitter.
- **Dataset Balance**: The equal representation of tweets from both parties allowed the machine learning model to learn effectively from both classes, reducing bias and ensuring more accurate predictions.



6. Recommendations

- **Enhance Multilingual Support**: Many tweets contain regional languages or a mix of Hindi and English. Incorporating multilingual NLP models could improve classification accuracy and allow the model to better handle non-English content.
- **Expand Topics of Analysis**: Further analysis could focus on specific topics, such as economic policies, education, or healthcare, to uncover deeper insights into the issues that dominate political discourse.
- **Explore Advanced Models**: Techniques like **transformers (e.g., BERT)** could capture context and nuances in language more effectively, potentially improving prediction accuracy.
- **Real-Time Analysis**: Deploying this model for real-time tweet analysis could help track evolving political communication strategies over time, providing valuable insights into ongoing political events.

7. Conclusion

This project has successfully provided a comparative analysis of political communication strategies employed by BJP and INC on Twitter. By using machine learning models to classify tweets, we identified distinct differences in the language and messaging styles of both

parties. The trained logistic regression model performed well with an accuracy of 85%, and insights drawn from this analysis can help shape future political strategies and media engagement.

Future work should explore enhancing the model with multilingual capabilities and advanced NLP techniques to improve the depth and accuracy of the analysis.