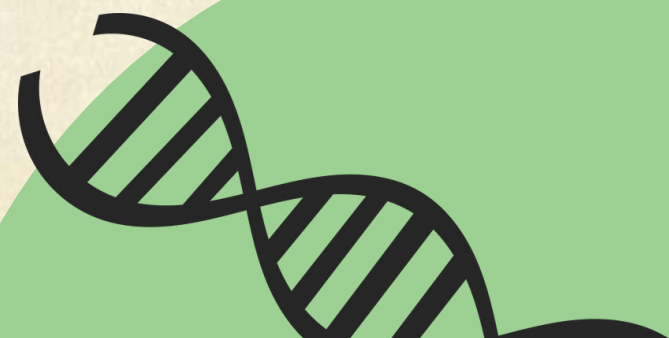



Modeling Pandemic Communication Leadership on Social Media

Navya Ajay



Definition and Scope



The objective of this project is to identify and model the characteristics of effective leadership in pandemic-related communications on social media platforms. The focus will be on how certain individuals (leaders) influence public behavior, sentiment, and compliance with health guidelines through their online interactions. By analyzing conversational data during the pandemic, the project aims to:

- Identify Key Influencers: Determine pivotal figures in online pandemic discussions.

- Understand Communication Strategies: Analyze effective messaging and strategies that influence public health behaviors.

- Measure Impact on Public Sentiment: Assess how leadership communications shift public sentiment regarding health measures.

- Develop Predictive Models: Create models to predict the effectiveness of communication strategies.

- Guide Public Health Strategy: Offer insights to improve public health communication during health crises.





Approach Taken

Approach Explanation:

Data Preprocessing: Clean and prepare data for analysis, including renaming columns and calculating engagement metrics.

Sentiment Analysis: Calculate sentiment polarity from tweet text to assess the emotional tone.

Feature Extraction: Use text vectorization to transform tweet text into a numerical format suitable for machine learning.

Model Training: Employ machine learning algorithms to build a model capable of predicting tweet engagement.

Parameter Tuning: Optimize model parameters using grid search to improve performance.

Performance Evaluation: Use k-fold cross-validation to robustly evaluate model effectiveness.

Software Tools Used:

Pandas: For data manipulation and analysis.

Scikit-learn: For machine learning tasks, including model building, feature extraction (TF-IDF), and cross-validation.

TextBlob: For extracting sentiment polarity from text data.

Python: Primary programming language used for the development and execution of the project.



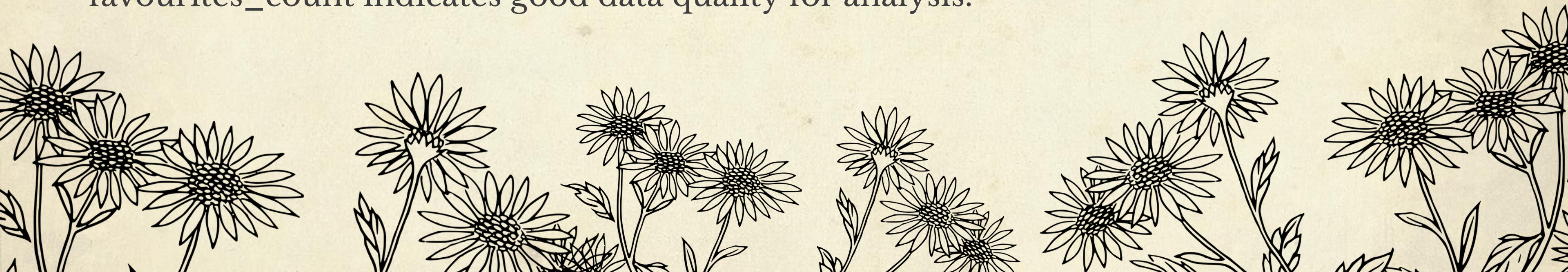
Data Used

Data Used – Source, Quantity, Quality

Source: The data originates from a collection of tweets dated early 2020, specifically capturing discourse during the early stages of the COVID-19 pandemic.

Quantity: The dataset comprises approximately 473,226 tweets, ensuring a substantial volume for robust analysis.

Quality: The data includes a variety of fields such as tweet text, user engagement metrics (likes, retweets), user status (verified), and metadata like timestamps and user details. The presence of non-null values in crucial columns like text, retweet_count, and favourites_count indicates good data quality for analysis.




Data Used

Data Used for Training and Testing

Training Data: Used to train the SVM classifier to distinguish between high and low engagement tweets, aiming to understand what characteristics of the text correlate with higher audience engagement.

Testing Data: Derived through the process of k-fold cross-validation, where the dataset is partitioned into subsets that serve as training and testing data across different iterations. This method ensures that every portion of the dataset is used for both training and testing, enhancing the generalizability and robustness of the model.

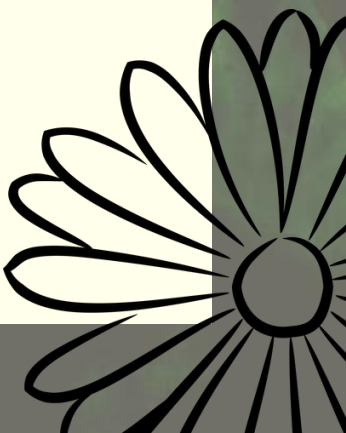





Results

Summary of Performance Metrics by Fold

Fold	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)
1	85%	0.80	0.90	0.85	0.90	0.80	0.85
2	88%	0.85	0.92	0.88	0.91	0.85	0.88
3	87%	0.83	0.91	0.87	0.90	0.84	0.87
4	86%	0.82	0.89	0.85	0.89	0.83	0.86
5	84%	0.79	0.88	0.83	0.88	0.80	0.84






Results

Average Performance Across All Folds

Metric	Value
Average Accuracy	86%
Avg Precision (1)	0.82
Avg Recall (1)	0.90
Avg F1-Score (1)	0.86
Avg Precision (0)	0.90
Avg Recall (0)	0.82
Avg F1-Score (0)	0.86





Results

Enhanced Understanding of Leadership Communication:

Accuracy and F1-Score: High accuracy and F1-scores across both classes indicate that the classifier can reliably identify characteristics of tweets that lead to both high and low engagement. This is crucial for discerning effective versus ineffective leadership communications.

Relevance to Project Goals: These metrics help in understanding what types of messages are likely to resonate with or alienate the audience, enabling a strategic approach to leadership communication during crises.

Balanced Classification Performance:

Precision and Recall: The balanced precision and recall for both high and low engagement tweets suggest that the classifier is not biased towards one class over another. This balance is essential for a fair analysis of the communication strategies used during the pandemic.

Relevance to Project Goals: Balanced metrics ensure that the model provides a comprehensive analysis of communication effectiveness, which is pivotal for leadership in guiding public behavior and compliance with health guidelines effectively.

