TWITTER SENTIMENT ANALYSIS

HARISH RAGURU   
Data Science  
University of MassachussetsDartmouth, USA  
hraguru@umassd.edu

BANOTH JEEVAN KUMAR  
Data Science  
University of MassachussetsDartmouth, USA  
jbanoth@umassd.edu

KONDURI HANUMA VENKATA VIJAY KAMAL  
Data Science  
University of MassachussetsDartmouth, USA  
hkonduri@umassd.edu

**Abstract— This study leverages Twitter data to perform sentiment analysis, aiming to understand public opinion and sentiment across various branches and countries. We employ supervised learning techniques, including Multinomial Logistic Regression and Support Vector Machines, to predict sentiment polarity and conduct extensive preprocessing and exploratory data analysis to examine tweet characteristics and sentiment distribution.**

***Keywords: Sentiment Analysis, Twitter, Data Preprocessing, Machine Learning, Supervised Learning***

I. INTRODUCTION

Social media platforms, especially Twitter, have transformed how we access and analyze public sentiment worldwide. This project leverages this extensive data pool to dissect and interpret sentiments across various regions and sectors, classifying tweets into categories such as Positive, Negative, Neutral, and Irrelevant to better understand public discourse dynamics. Using advanced machine learning techniques and thorough preprocessing including cleaning and tokenization, we analyze trends and potential predictors that impact social behavior and public reactions. Our exploratory data analysis delves into tweet characteristics and sentiment distribution, providing valuable insights for policymakers and businesses while monitoring evolving trends. Ultimately, this study aims to refine sentiment analysis tools to predict and influence socio-political events, enhancing our grasp of digital communication's role in shaping societal dynamics.

II. METHODOLOGY

Our methodology begins with an extensive preprocessing of the Twitter dataset, where we convert all text to lowercase, remove punctuation, and filter out stop words, then apply stemming to condense words to their base forms, enhancing the reliability of our data. Each tweet is then categorized into one of four sentiment classes—Positive, Negative, Neutral, or Irrelevant—based on its content and context. To analyze the processed data, we employ exploratory data analysis techniques such as word clouds, which highlight the most common terms within each sentiment category, and distribution plots, which reveal patterns in tweet lengths and sentiment distribution across different topics. This layered approach not only elucidates general sentiment trends but also provides insights into their implications for public opinion and social dynamics, offering a nuanced understanding of complex interactions on Twitter.

III. SENTIMENT ANALYSIS

A. *Dataset*

Our study leverages the Twitter Sentiment Analysis Dataset from Kaggle, designed to explore sentiment expression across social media. This diverse dataset contains a vast array of tweets each labeled with sentiments such as Positive, Negative, Neutral, or Irrelevant, reflecting a wide range of emotions and topics. Pre-processing steps have been applied to remove sensitive information and identifiers, ensuring user privacy and data integrity, which allows us to conduct a thorough analysis of sentiment distribution across various demographics. By utilizing this dataset, we aim to identify patterns and trends in public opinion, thereby enhancing our understanding of the dynamics of digital communication.

B. *Supervised learning methods*

Supervised learning is a type of machine learning where models are trained using labeled data, where each input is paired with the correct output, allowing the model to learn the mapping between them during training. This method is commonly employed in classification and regression tasks, enabling models to learn by comparing their predictions to the actual outcomes and making necessary adjustments.

1) Support Vector Machine (SVM):

Support Vector Machine (SVM) is a robust, non-probabilistic classifier that identifies the optimal hyperplane to separate different classes in a feature space and can extend to multi-class classification through techniques like one-vs-all. For datasets that are not linearly separable, SVM employs kernel functions to project the data into higher dimensional spaces where separation is feasible, enhancing its capability in handling complex, non-linear datasets such as those in text classification and sentiment analysis. Central to SVM is the maximization of the margin between class data points, which bolsters decision confidence and underpins its high accuracy and effectiveness, especially in high-dimensional feature spaces typical of text data.

2) Random forest:

Random Forest is an ensemble learning method that builds multiple decision trees during training and uses a majority voting system for classification, enhancing accuracy and reducing overfitting compared to single decision trees. It effectively handles missing data and diverse data types (binary, categorical, or numerical) by constructing each tree from a random subset of the data and features, which promotes model diversity and improves generalization. The method is notably robust to outliers and noise, making it a reliable option for dealing with complex and imperfect real-world data.

C. *Analysis*

1) Preprocessing

Converting all text data to lowercase is a standard preprocessing step to normalize the dataset, ensuring that the model does not treat words differently based on case. Removing punctuation helps in cleaning the text, eliminating potential noise that could affect text analysis and sentiment classification. Stop word removal is crucial as it excludes commonly used words (such as "and", "the", "a") that do not contribute significantly to the sentiment of the text, allowing the model to focus on more meaningful words. Stemming reduces words to their root form, which consolidates variations of the same word, thus reducing the complexity of the data and improving the efficiency of the analysis. These preprocessing steps collectively streamline the text data, enhancing the performance of machine learning models by focusing on the most informative aspects of the text.

2) Sentiment labelling:

Sentiment labeling involves categorizing text data into predefined sentiment classes such as Positive, Negative, Neutral, or Irrelevant, based on the expressed emotions and opinions. This process can be automated using predefined algorithms that analyze the words and context within the text to determine the appropriate sentiment. Manual labeling might also be employed, where human annotators review and assign sentiments, which can help in training and refining automated systems. Accurate sentiment labeling is essential for training supervised learning models that rely on labeled data to learn how to predict the sentiment of new, unseen texts. The effectiveness of sentiment analysis models heavily depends on the quality and consistency of the sentiment labeling, making this a critical step in the sentiment analysis pipeline.

IV. RESULTS

1) Positive sentiment word cloud:

The Positive Sentiment Word Cloud visualization highlights the most common words found in tweets with positive sentiment, with larger word sizes indicating higher frequencies. This display offers insights into the prevalent themes and topics associated with positive sentiments in tweets.



Fig. 1. Positive sentiment word cloud

2) Negative sentiment word cloud:

The Negative Sentiment Word Cloud visualization reveals the most frequently used words in tweets labeled with negative sentiment, with larger words signifying greater frequency. This helps illustrate the dominant themes and topics associated with negative sentiments.

A close up of words

Description automatically generated

Fig. 2. Negative sentiment word cloud

3) Irrelevant sentiment word cloud:

The Irrelevant Sentiment Word Cloud visualization displays the most used words in tweets classified with irrelevant sentiment, highlighting the frequency of each. This aids in identifying words that contribute to noise or irrelevant content within the dataset, offering insights into data that may not be useful for sentiment analysis.

A close up of words

Description automatically generated

Fig. 3. Irrelevant sentiment word cloud

4) Neutral sentiment word cloud:

The Neutral Sentiment Word Cloud showcases the most frequently used words in tweets labeled with neutral sentiment, with larger words representing higher occurrences. This visualization helps identify prevalent themes and topics within neutral sentiment tweets, illustrating commonalities in expressions that lack clear emotional tone.

A close up of words

Description automatically generated

Fig. 4. Neutral sentiment word cloud

5) Distribution of tweet length by sentiments:

This histogram illustrates the distribution of tweet lengths across different sentiment categories—positive, negative, neutral, and irrelevant—with each section of the plot representing a specific sentiment. By binning tweet lengths and counting the frequency within each bin, it provides insights into the typical tweet lengths for each sentiment category, facilitating comparisons across them.

A graph of different sizes of a graph

Description automatically generated with medium confidence

Fig. 5. Distribution of tweet length

6) Distribution of sentiment labels:

This bar plot displays the distribution of sentiment labels—positive, negative, neutral, irrelevant—across all tweets in the dataset, with each bar indicating the number of tweets associated with each sentiment. The visualization aids in understanding the overall prevalence of each sentiment category within the Twitter data, offering insights into the dominant sentiments expressed.

A graph of different colored rectangular shapes

Description automatically generated

Fig. 6. Distribution of sentiment labels

7) ROC and AUC curves:

The ROC (Receiver Operating Characteristic) curves for two machine learning models, SVM (Support Vector Machine) and Random Forest, are used for classification tasks. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, with sensitivity plotted against 1-specificity. In this chart, both models appear to perform similarly, as indicated by the curves closely aligning along the diagonal from the bottom left to the top right, which represents a random guess. The Area Under the Curve (AUC) value is 0.502, suggesting that the classification performance of both models is barely above random chance, indicating a need for model improvement or parameter tuning.

A graph showing a curve

Description automatically generated

Fig. 7. ROC & AUC curve

V. CONCLUSION AND DISCUSSION

The application of sentiment analysis techniques to Twitter data has proven highly effective for gauging public opinion and sentiment trends. By analyzing the sentiment expressed across various branches and geographic regions, we have gained valuable insights into how populations feel about different topics and events. This analysis has helped uncover patterns in sentiment that are pivotal for understanding broader social and cultural dynamics. Overall, the study demonstrates the powerful potential of leveraging social media data to capture the pulse of the global community.

One limitation of our study is the reliance on manual sentiment labeling, which introduces the potential for bias and errors in data classification. The scope of the dataset also presents constraints, as it may not fully represent the wide diversity of global Twitter users. On the positive side, our project's strength lies in its comprehensive preprocessing and the development of insightful visualizations that effectively communicate the results. These efforts have significantly contributed to the clarity and depth of our findings, making the data more accessible and understandable.

VI. REFERENCES