Project: Leveraging Social Media Data for Political Insight

Introduction

In the digital age, social media platforms have become prolific sources of data, offering a wealth of information on public opinion, trends, and societal dynamics. This project emerges at the intersection of data science and social media analysis, aiming to harness advanced computational techniques to extract meaningful insights from vast, unstructured textual datasets. The project aims to explore and analyze social media data using advanced data processing and machine learning techniques. The focus is on sentiment analysis, topic modeling, and classification to derive meaningful insights from textual data. By doing so that the stakeholders can know what kind of topics have the most impact on twitter. We received 10 months of twitter datasets particularly focused on the 2016 election year. Each dataset individual contains around 74,231,650 rows with six attributes. The attributes are: -

- Text (object): The core of our analysis, this field contains the textual content of each post. It is a treasure trove of unstructured data, rich with colloquial language, emojis, and internet slang.
- TweetID (object): A unique identifier for each post, enabling us to track and reference individual entries efficiently.
- UserID (object): The identifier of the user who posted the message. This allows for analysis of user-specific patterns or behaviors.
- Date (object): The timestamp of each post. This temporal data is crucial for understanding trends over time, including reactions to events or changes in public sentiment.
- RT_ID (object): The identifier for retweets, linking posts to their original sources and illustrating the spread of information and opinions.
- RT_UserID (object): The identifier of the user who is retweeted, providing insights into the influence and reach of specific users or messages.

This project aims to navigate the complexities of this large and intricate dataset. The primary objectives are threefold: to combine and cleanse the data to form a coherent and analyzable corpus; to conduct sentiment analysis for understanding public opinion and emotional responses; and to perform topic modeling and classification, with a particular emphasis on data related to Donald Trump.

Methodology

1.Data Collection and Preparation (from "AML_project_step1_combining&cleanignalldatasets.ipynb"

- Data Loading&Data Merging: Combining multiple datasets into a single DataFrame, ensuring consistency and proper alignment of data from the sources.
- Null Value Analysis: Inspection of null values within the dataset, particularly in columns related to retweets.
- Retweet Analysis: Counting tweet IDs in retweet-related columns and mapping these counts to specific columns in the dataset. So that we can filter all the retweets.
- Data Imputation: imputed values in specific columns rather than removing rows with null values, thereby preserving the dataset's integrity for further analysis.
- Preprocessing steps included tokenization, lemmatization, and handling of emojis of the text.

2.Sentiment Analysis (from "aml_project_step_2_sentiment_analysis.ipynb")

- Tweet Categorization: Developing a method to categorize tweets based on mentions, with a focus on Donald Trump and Hillary Clinton.
- Sentiment Analysis Attempt: Initial trials with BERT (Bidirectional Encoder Representations from Transformers) for sentiment analysis, with notes on challenges faced due to the dataset's size we made choice to use the Vader sentiment analysis on the cleaned text we got from the step: - 1. we classified the text in to two categories positive or negative by the vader score.

3.Topic Modeling and Classification (from "aml_project_step3_topic_model_+randomforest_for_trump.ipynb")

- Topic Modeling Preparation: Setting up the environment and data for topic modeling, including potential challenges with computational resources. Focus on Trump's Tweets: Narrowing the dataset to focus specifically on tweets related to Donald Trump, likely due to resource constraints.
- Data Splitting: Division of the dataset into training and test sets, a crucial step for any machine learning analysis.
- Top2Vec for Topic Modeling: Application of Top2Vec for topic discovery in the dataset, along with notes on the decision-making process due to resource limitations.
- Random Forest algorithm was applied for Regression and classification, particularly focusing on data related to Donald Trump.

Results

Notebook 1 Results:

We got a dataset by combining all the 10 months of the and doing pre data processing like imputation, dropping the duplicates and cleaning the text.

Notebook 2 Results:

Once we applied the sentiment analysis to the cleaned the text, we got to know what kind of sentiments have more retweets than the others.

Notebook 3 Results:

We have to go with the smaller datasets, so we filtered the tweets of the Trump from the dataset and used this to perform the top2vec model. The model retrieved 4 topics, and these were assigned to each tweet according to its highest topic score. By doing all of we got to know which topic of the trump has highest number of tweets.

Later we performed or predictive analysis we done three models

1. Regression Model:

Purpose: To predict the retweet count.

Features Used: A set of features was experimented with, including dominant topic and sentiment score. However, to achieve the best R² value, a refined set of features was selected.

Performance: we got the 18% r2 because we don't have the features that explain a lot of the target variable which retweet count. We would have gotten a better result if we had features like number of likes, unlike, followers and comments. even though we got the less r2 , the error percentage for the retweent count is 22% which means we got a better results form this model.

2. Classifier Model 1:

Purpose: To predict the dominant topic.

Features Used: 'day_of_week', 'month', 'day', 'hour', 'minute', 'second', 'year', 'tweet_length', 'sentiment', 'retweet_count'.

Performance: This model achieved an accuracy of 92% and an F1 score of 89%, indicating its effectiveness in classifying the dominant topic based on the given features.

3. Classifier Model 2:

Purpose: Also, to predict the dominant topic.

Features Used: 'vader_sentiment', 'day_of_week', 'month', 'tweet_length', 'retweet_count'.

Performance: This model showed an accuracy of 93% and an F1 score of 90.10%, demonstrating a slightly improved performance compared to Classifier Model 1. Even though the clasifier still gave the best results it would have given much better results if we applied for the whole dataset.

Conclusions/Recommendations

Direct Analysis: By examining the retweet count vs. dominant topic bar graph, stakeholders can identify which topics garner the most retweets and tailor their content accordingly.

Predictive Approach:

We can predict the dominant topic in two ways by the available features.

1. Time and Sentiment-Based Prediction: If only time data and sentiment (positive or negative) are available, stakeholders can first use the regression model to predict the retweet count [because it is part of the classifier model features], and then apply Classifier Model 1 to predict the most effective topic.

2. Text, Sentiment Score, and Time-Based Prediction: With access to the text, sentiment score, and time details, stakeholders can use the regression model to forecast retweet counts followed by Classifier Model 2 to determine the topic. This approach is suitable for stakeholders interested in both the potential reach (retweets) and the topic relevance of their content.

3. Direct Topic Prediction: Alternatively, stakeholders can directly use either Classifier Model 1 to predict the most suitable topic based on their requirements such as time and desired number of retweets (which they can assign them self's), without incorporating the regression model in the process.

Each approach offers different insights and applications, allowing stakeholders to choose the one that best fits their specific needs and available data. For future we could do the work on the whole dataset without any filtration which will gives us more accurate predictive models. Also, we could gather more information about the user so that we can build around the prediction retweets or likes. In the coming future, social media will hold a powerful place in the elections camping's so if we have better predictive models how to reach more people through the social media maybe there will be chance of running a successful camping's.

Acknowledgements: -

This project would not have been possible without the invaluable guidance and support of Professor Joshua Introne by not only provided us with the extensive dataset that formed the foundation of our analysis but also offered critical insights that enhanced our understanding of data science principles and methodologies

The completion of this project was facilitated through the utilization of various online resources and the assistance of ChatGPT