Name: Adrian Zevenster Position: ML Engineer

`Sentiment Analysis Repository [GitHub]`

1 Introduction

Twitter (X) posts around airline sentiment made available from Kaggle. Sentiments are categorized into three groups: neutral, positive and negative. Each airline has a confidence score correlated to sentiment classifications. We have chosen to examine whether the confidence score accurately reflects what the data conveys by determining whether scores might be biased or skewed towards particular sentiments.

Logic and data preparation are broken down into, followed by results analysis and a brief conclusion:

Step 0 : Database Creation

Step 1 : Data Preprocessing and Cleaning

Step 2 : Data exploration

Step 3 : Model Creation

Step 5 : Conclusion

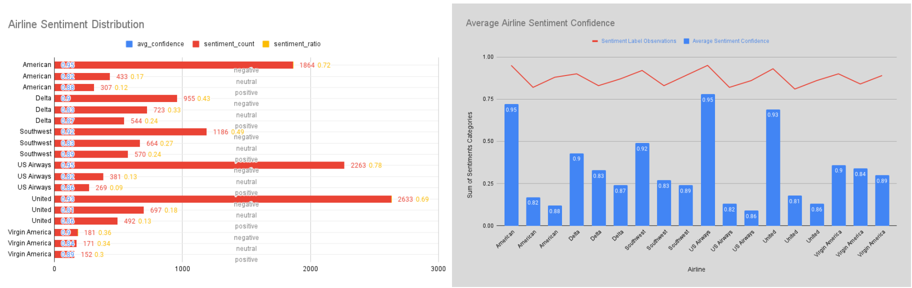
Step 6 : Exploration and Suggestions

2 Methods and Discussion

This section provides an overview of the methodologies implemented in the *'Praelexis.ipynb'* Jupyter Notebook. We discuss steps taken during data analysis with logical flow and report findings. *pymc* libraries are used for modelling, pandas for querying SQLite database a discussion of the critical files within the repository, emphasizing their relevance to our objectives. Additionally, the section delves into the rationale behind the chosen strategies, offering insights into the decision-making process that shaped our approach."

2.1 Step 0: Database Creation

The initial step of the analysis involves creating a structured database using SQLite storing airline tweet sentiment data, a lightweight and versatile relational database management system. Leveraging Jupyter Notebook and Python, data importation of tweets datasets into the SQLite schema is facilitated, enabling seamless integration of code, documentation, and visualization. This step lays the groundwork for efficient data management and retrieval, setting the stage for subsequent analysis.



(a) Airline$_sentiment$

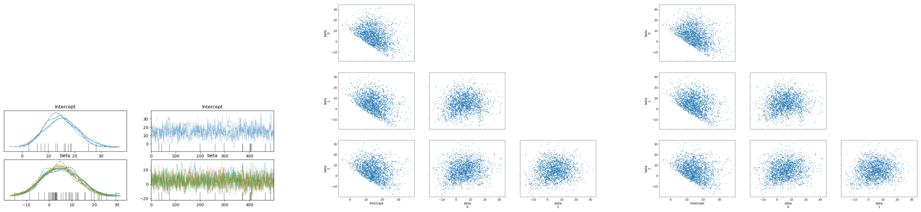
(b) Av3

Figure 1: Sentiment Prediction Analysis

2.2 Step 1: Preprocessing and Cleaning

Data preprocessing and cleaning are paramount to ensure the quality and reliability of the dataset. In this phase, various operations are performed, including handling missing values, standardizing and normalizing data, feature engineering, outlier detection, and data validation. Through meticulous preprocessing, the dataset is refined and prepared for in-depth analysis, mitigating potential biases and inaccuracies.

2.3 Step 2: Model Creation



(a) Beta-Intercept

(b) Pairplot

(c) Posterior Trance

Figure 2: pymc Markov Analysis

2.4 Step 3: Performance Evaluation

3 Conclusion and Suggestions