**Text Analysis**

**Abstract:** This project analyzes sentiment using the **Sentiment140 dataset**. It preprocesses the text, applies vectorization, and splits the data into training and testing sets to train a model. The goal is to accurately classify sentiments and gain insights.

**Data Understanding:**

The dataset used for this analysis is a collection of tweets labeled with sentiment values, specifically indicating whether the sentiment of each tweet is positive or negative. The data, originally sourced from a CSV file, includes 1,600,000 processed tweets. Each entry contains a sentiment label (0 for negative, 4 for positive), a unique identifier, the date of the tweet, the querying user, and the tweet's text. A screenshot of a computer code

Description automatically generated

For this analysis, we focus on the sentiment and text columns, as they are critical for understanding the emotional tone of the tweets. After dropping unnecessary columns and duplicates, we transformed the sentiment labels from numeric values to categorical terms—'negative' for 0 and 'positive' for 4. This structured dataset provides a rich basis for performing sentiment analysis, allowing us to explore patterns in user emotions expressed through social media.

A screenshot of a computer program

Description automatically generated

**Data Processing Steps:**

1. **Data Cleaning:** Remove duplicates and handle missing values.
2. **Text Preprocessing:** Convert text to lowercase, remove punctuation and stopwords.
3. **Tokenization:** Split text into individual words.
4. **Vectorization:** Convert tokens into numerical format using Count Vectorization or TF-IDF.
5. **Data Splitting:** Split into training and testing sets.

**Model Choice: Logistic Regression with BoW and TF-IDF**

we chose Logistic Regression for sentiment analysis due to its effectiveness in binary classification.

**Bag of Words (BoW)** captures word frequency, while **TF-IDF** highlights important words by reducing the influence of common terms.

Together, they provide a strong feature set for accurate sentiment classification.

**Performance Evaluation:**

We chose accuracy to evaluate the sentiment analysis model, as it measures the proportion of correctly classified instances, providing a clear assessment of overall effectiveness in this binary classification task.

**Insights Gained:**

The Logistic Regression model achieved an accuracy of **0.75635** with Bag of Words (BoW) and **0.7604** with TF-IDF. This slight improvement with TF-IDF suggests that considering word importance enhances sentiment classification. The results highlight the effectiveness of feature representation in text analysis, demonstrating that both methods yield reliable performance, with TF-IDF providing a marginal advantage. Overall, these insights emphasize the importance of selecting informative features to improve model accuracy.