**Data Mining & Text Analytics**

**MIS-5613-270**

**Submitted by:**

Larissa Honore

Kusum Ranabhat

Jonas Owusu-Ansah

Anthony Okpara

Ifeanyichukwu Ojuh

**Analyzing Public Sentiments on AstraZeneca’s Covid-19 Vaccine: A Twitter Based Study**

Our project on Analyzing Public Sentiments on AstraZeneca’s Covid-19 vaccine: A Twitter based study, aims to determine whether the content of a tweet is positive, negative, or neutral based on Natural Language Processing (NLP) and Machine Learning (ML). In this project, we consider a dataset of tweets related to AstraZeneca's vaccine and implement a sentiment classification pipeline. We'll load and preprocess the data, perform exploratory analysis (including plotting the distribution of sentiment scores and common words), then train a Multinomial Naive Bayes model to predict tweet sentiment. Lastly, we evaluate the model and discuss potential improvements.

Source: <https://www.kaggle.com/datasets/gallo33henrique/twitter-astrazeneca-anticovid>

Topic: Twitter AstraZeneca AntiCovid

Number of records: 1553

**Data Loading and Preprocessing**

We begin by loading the CSV dataset using pandas and preprocessing the text of the tweets for higher quality analysis. This includes lowercasing, removing punctuation, URLs, and stop words, and lemmatizing each word. These are standard preprocessing methods that normalize and reduce noise in text data.

**Exploratory Data Analysis (EDA)**

Then we perform some descriptive analysis of the dataset. We examine the number of tweets and balance of sentiment labels. We also examine the polarity and subjectivity scores provided in the dataset. Polarity generally has a range of -1 to 1, indicating how negative or positive the sentiment in the text is, while subjectivity has a range of 0 to 1 indicating how subjective/opinionated the text is.

We examined the distribution of tweets by sentiment and the outputs are:

● Positive: 739

● Neutral: 677

● Negative: 136

Polarity and subjectivity scores are:

**Polarity:**

● Mean: 0.124

● Min: -0.75

● Max: 1.00

**Subjectivity:**

● Mean: 0.351

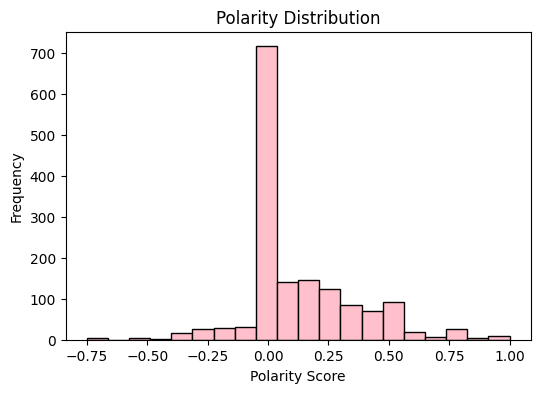
● Min: 0.00

● Max: 1.00

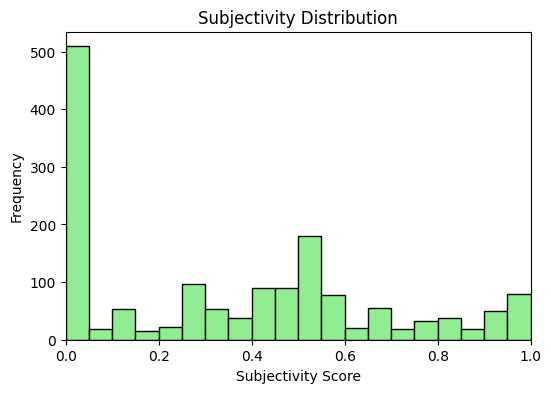
Out of 1,553 tweets, most are positive or neutral, with relatively few negative tweets. The average polarity score of 0.12 shows a slightly positive trend, and the subjectivity mean of 0.35 suggests many tweets are factual.

**Distribution of Polarity and Subjectivity**

To have an idea of the distribution of polarity and subjectivity scores, we plot histograms. It helped us to learn about how sentiment is distributed across tweets. For example: Either if most tweets are neutral (polarity near 0) or if there are many highly positive/negative tweets.



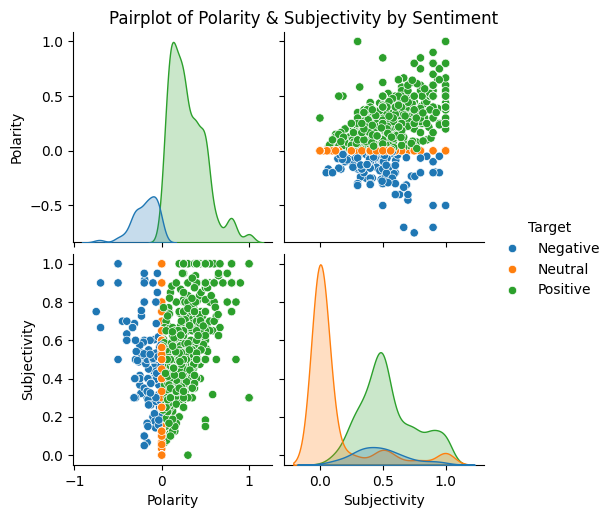
As we can observe from the above histogram, most tweets have a polarity score close to zero, and the most frequent is a neutral sentiment. We can observe fewer tweets with very positive or very negative sentiments, with a bias towards positivity.



Most tweets are clustered at the low end of the subjectivity spectrum. This means users are mostly tweeting factual or objective information which could be possibly quoting news or sharing research. However, there is also some dispersion along higher levels of subjectivity, meaning that there are quite a number of tweets that include subjective opinions.

**Visualizing Sentiment Relationships and Distribution**

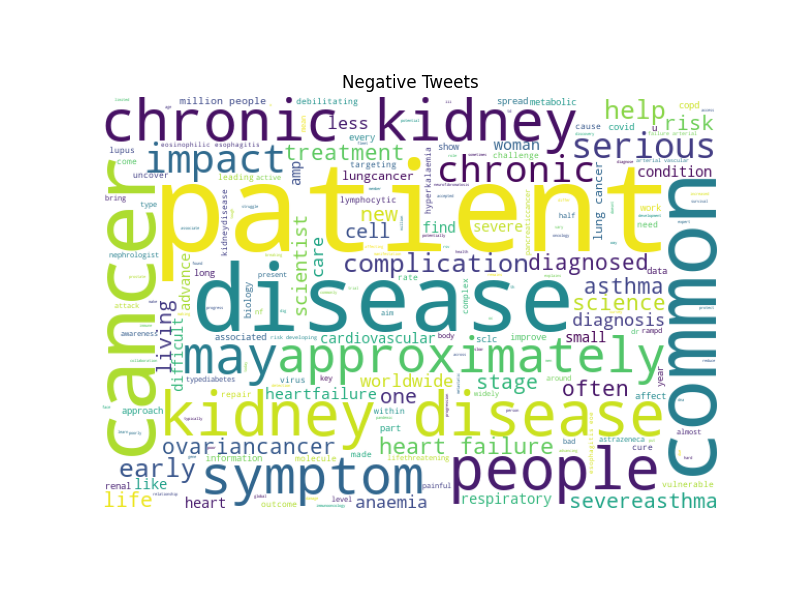
To get an idea of the relationship between polarity, subjectivity, and sentiment labels, we created a Seaborn pair plot. As we can observe from the plot, Neutral tweets bunch around zero polarity and low subjectivity, Positive tweets have a broad spread over both, but especially subjectivity. Negative tweets are fewer in number and bunch more on the low polarity side.

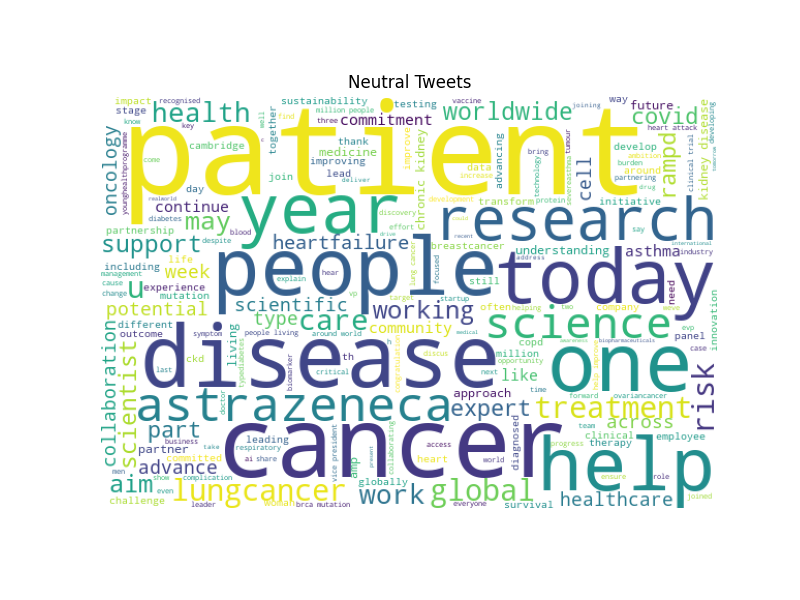


**Word Clouds of Frequent Words by Sentiment**

Another useful exploratory visualization is a Word Cloud for each sentiment class. A word cloud displays the most common words in the tweets, with more common words in larger font. This can give a qualitative sense of what topics or what words are prominent in positive vs negative tweets. We will create separate word clouds for Positive, Negative, and Neutral tweets. By filtering the tweets by sentiment label and grouping the text, we can see which words are strongly associated with each sentiment.







Negative tweets often include serious health terms like cancer, chronic, disease, and symptom. Positive and neutral tweets often include patient, research, help, and science, with positive tweets including terms like progress and new year.

**Model Training and Evaluation**

We used a Multinomial Naive Bayes (MultinomialNB) classifier with CountVectorizer to convert cleaned text to numerical features.

Train-Test Split and Feature Extraction:

● Training set size: 1241

● Test set size: 311

● Vocabulary size: 3246 unique words

Model Results:

● MultinomialNB Accuracy: 0.688

Confusion Matrix: [[ 7 10 10], [ 2 93 41], [ 3 31 114]]

Classification Report:

● Positive: Precision 0.691, Recall 0.770, F1-score 0.728

● Negative: Precision 0.583, Recall 0.259, F1-score 0.359

● Neutral: Precision 0.694, Recall 0.684, F1-score 0.689

MultinomialNB model achieved an accuracy of 68.8%. It worked very well on Positive and Neutral sentiments but failed to work very well on Negative tweets, and this was largely because of class imbalance.

**Conclusion and Recommendation**

Overall, this project presented successful sentiment analysis using text classification techniques. MultinomialNB model worked effectively, especially on positive and neutral classes. However, recall on the negative class was poor.

For performance improvement, we recommend the application of oversampling methods or class weight modification for better representation of negative tweets. Other models such as SVM or neural models such as BERT, and vectorization hyperparameter tuning, can also help in improving the accuracy of the classification. These steps may help ensure more balanced public opinion on significant healthcare concerns.