Summer 2022 Programming Challenge: Sentimental Analysis

-Mohammed Zubair Khan (mkhan217@usc.edu)

Web Scrapping and Data Clean-up

- Module 'beautifulsoup' is used to extract the content present in the html. The output contains all sorts of noises or unwanted data which must be cleaned before using them. Along with this, only required information must be extracted.
    - First I extracted the links from the news home page using the class_="gc__content".
    - Then I scrapped each of these links to get articles about those headlines.
    - In article web pages I selected all the tag 'p' except the one with class_='site-footer__social-section-title css-0'.
    - I have created a custom class "articles" to store the information about each articles(link, title, content)

Data Cleaning

- For data cleaning I used regular expressions, nltk and string operations (split,replace and strip)
- Regular expression was used for removing http links from data.
- Nltk was used to separate each sentence
- And string operations to remove spaces, '/n' , double and single quotes.

Sentimental Analysis

- **Method 1: Using Textblob**

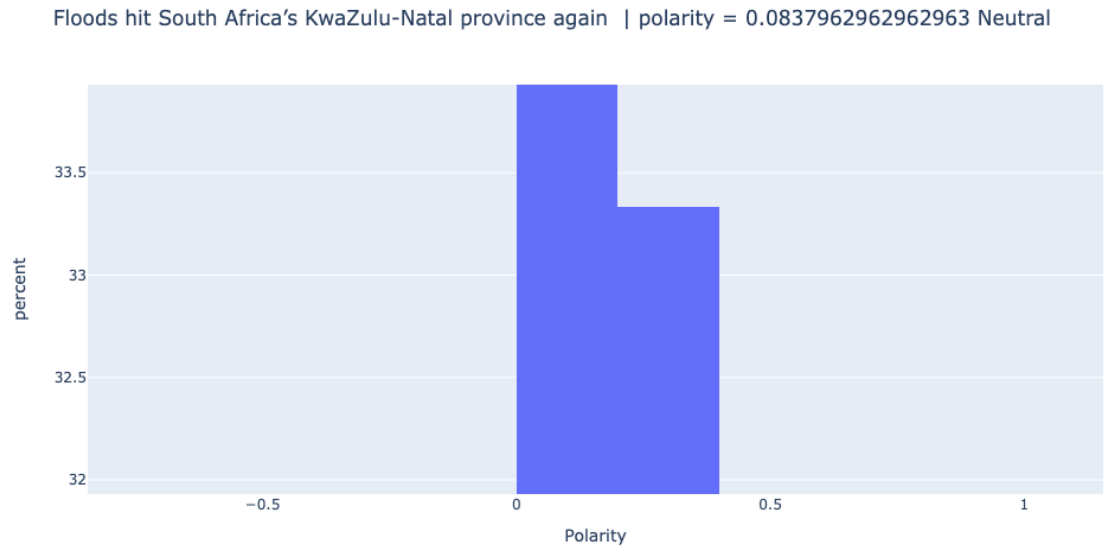
Textblob is a Python text processing and natural language processing module. The sentiment analysis functions of Textblob are built-in. For polarity and subjectivity, the function returns a score. Subjectivity ranges from 0 to 1, and Polarity can be positive or negative.

We created a array of sentences, these sentences were provided to text blob sentimental analysis model. Each sentence got a polarity and subjectivity score. We stored the sentences and scores in a data frame. Then we used plotly to, display histogram of each articles, where values were the polarity and subjectivity scores. For polarity we took a mean of all the polarity scores of all the sentences of an article to show the polarity of an article. We did the same with subjectivity scores. Then we passed polarity score to classify_sentiment() function, which will decide the sentiment. If the mean score is greater than 0.1 then the function returns "Positive". And if the mean score is less than 0, then it returns "Negative". If the mean score is between 0 and 0.1 then it returns "Neutral".

The polarity histogram plots we created had article title, polarity mean score and Sentiment.

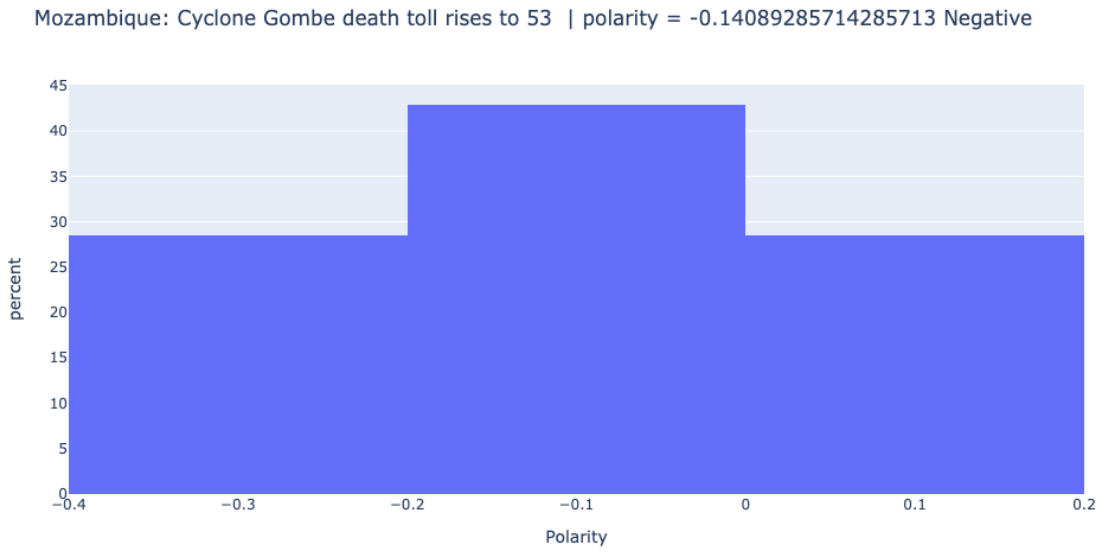
Article 1 :Floods hit South Africa’s KwaZulu-Natal province again.

The model predicted this article to be Neutral



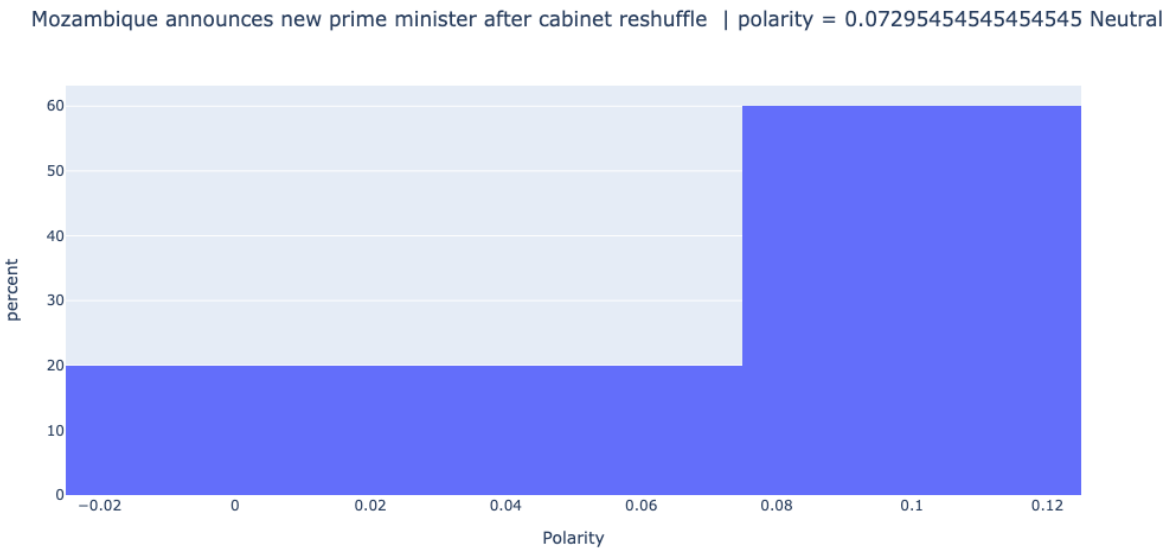
Article 2 :Mozambique: Cyclone Gombe death toll rises to 53

The model predicted this article to be Negative



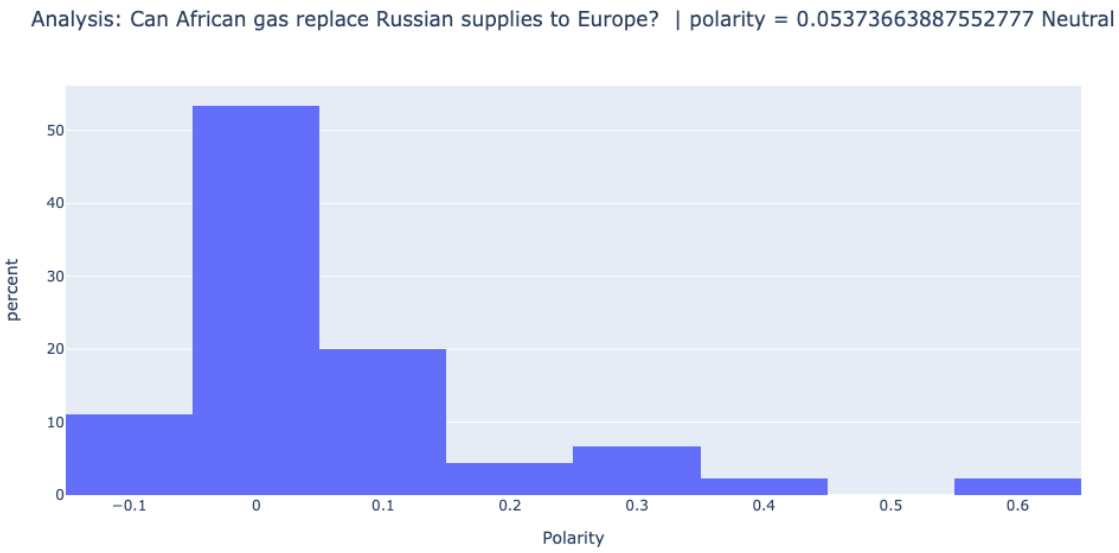
Article 3 :Mozambique announces new prime minister after cabinet reshuffle

The model predicted this article to be Neutral



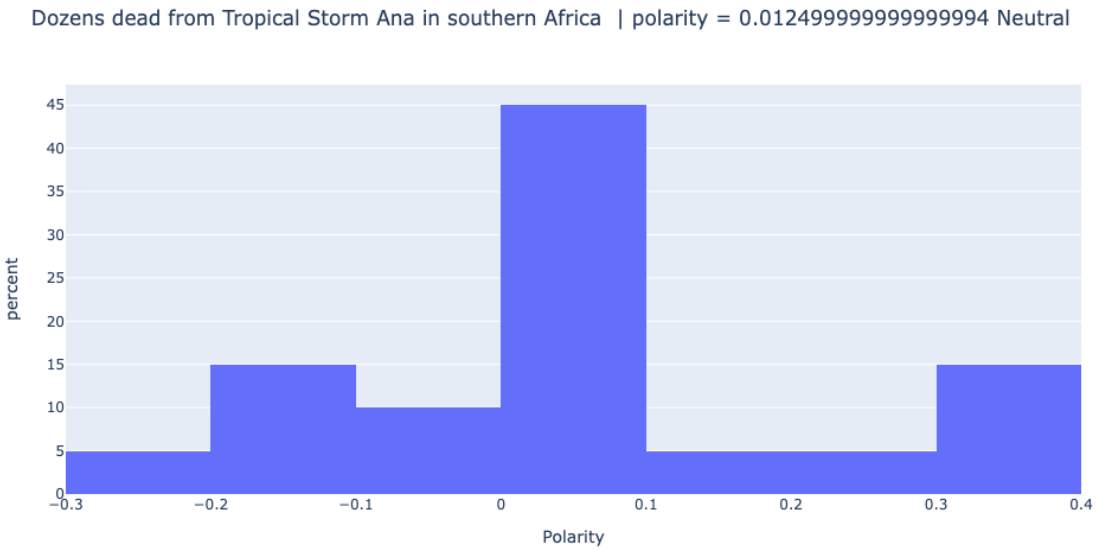
Article 4 :Analysis: Can African gas replace Russian supplies to Europe?

The model predicted this article to be Neutral



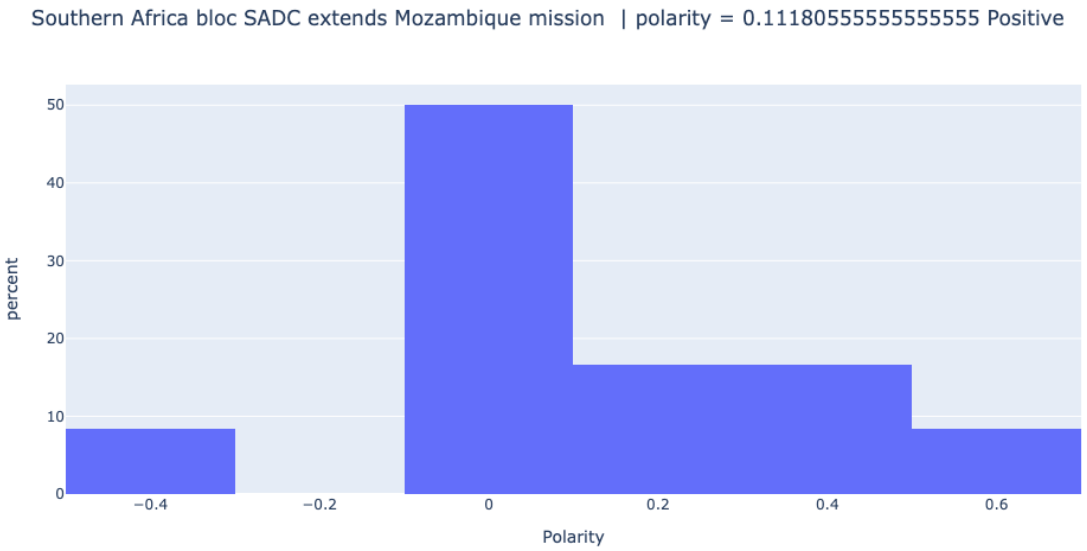
Article 5 :Dozens dead from Tropical Storm Ana in southern Africa

The model predicted this article to be Neutral



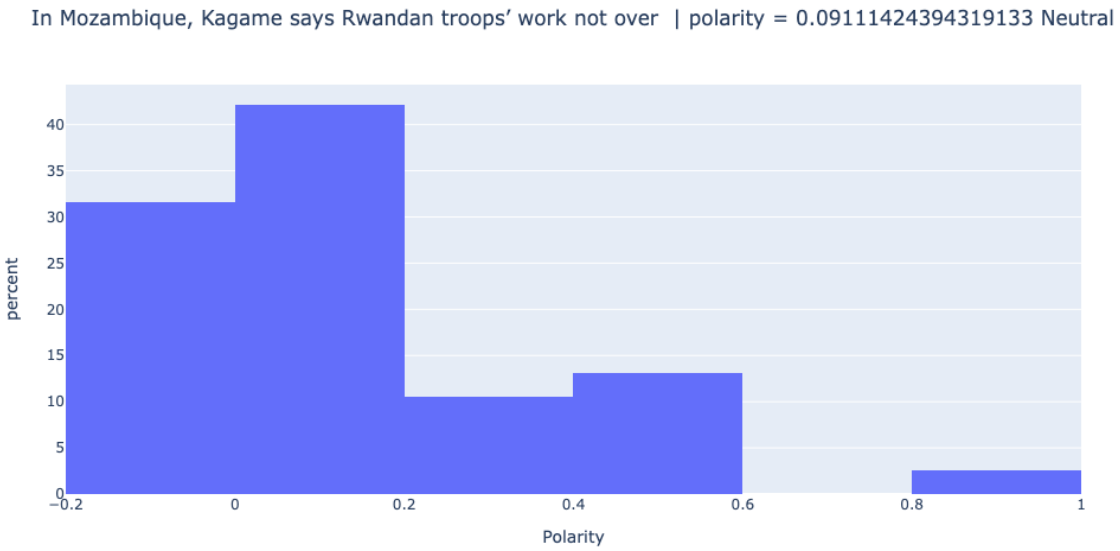
Article 6 :Southern Africa bloc SADC extends Mozambique mission

The model predicted this article to be Positive



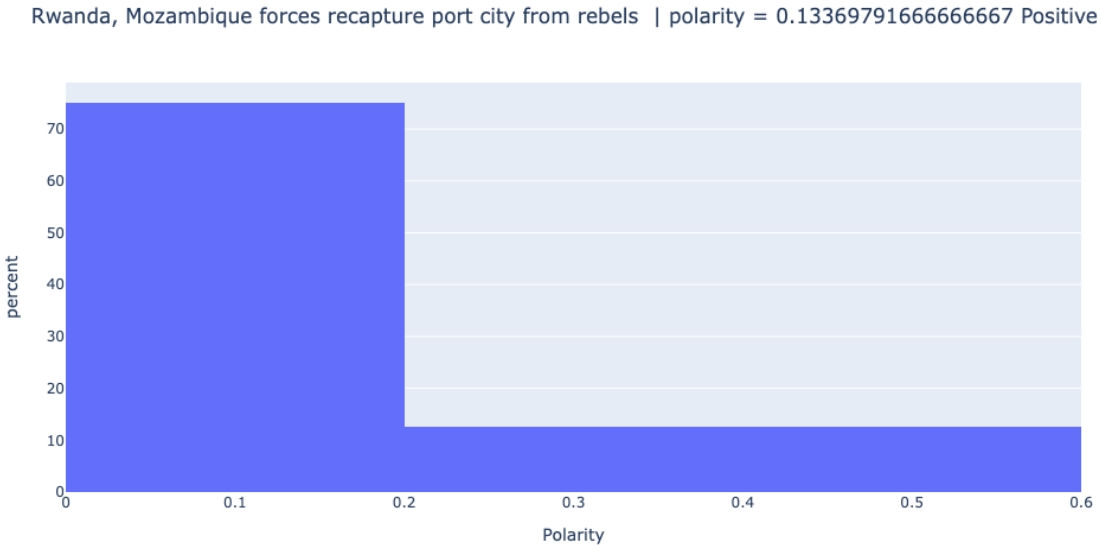
Article 7 :In Mozambique, Kagame says Rwandan troops' work not over

The model predicted this article to be Neutral



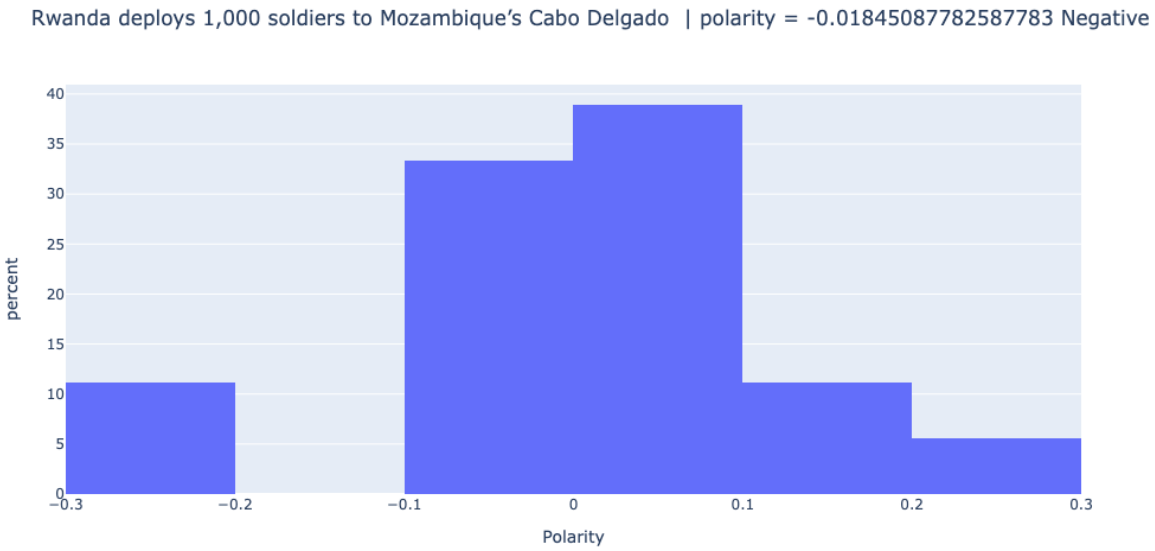
Article 8 :Rwanda, Mozambique forces recapture port city from rebels

The model predicted this article to be Positive



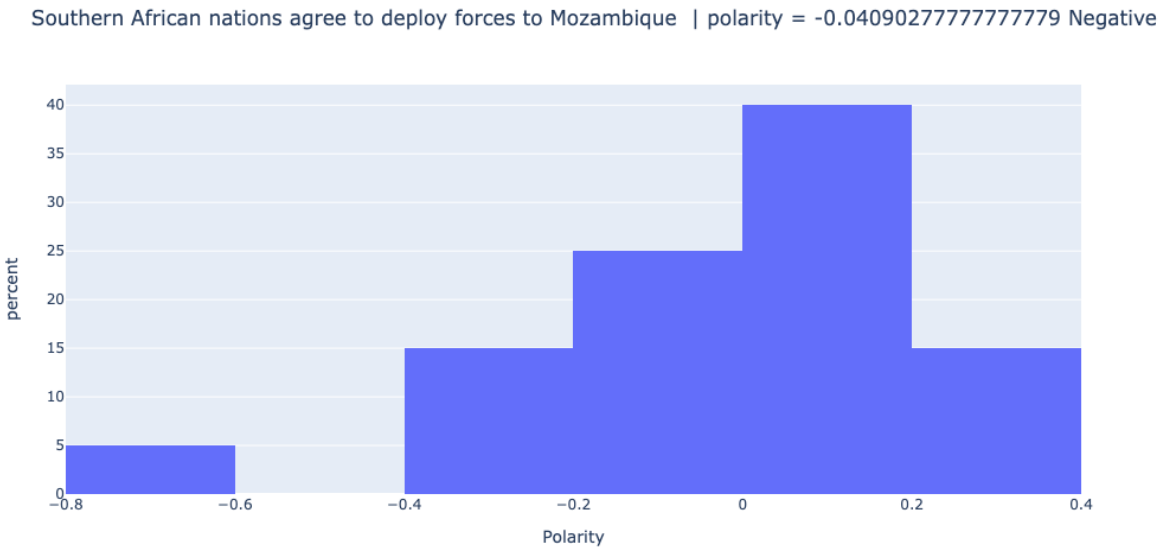
Article 9 :Rwanda deploys 1,000 soldiers to Mozambique’s Cabo Delgado

The model predicted this article to be Negative



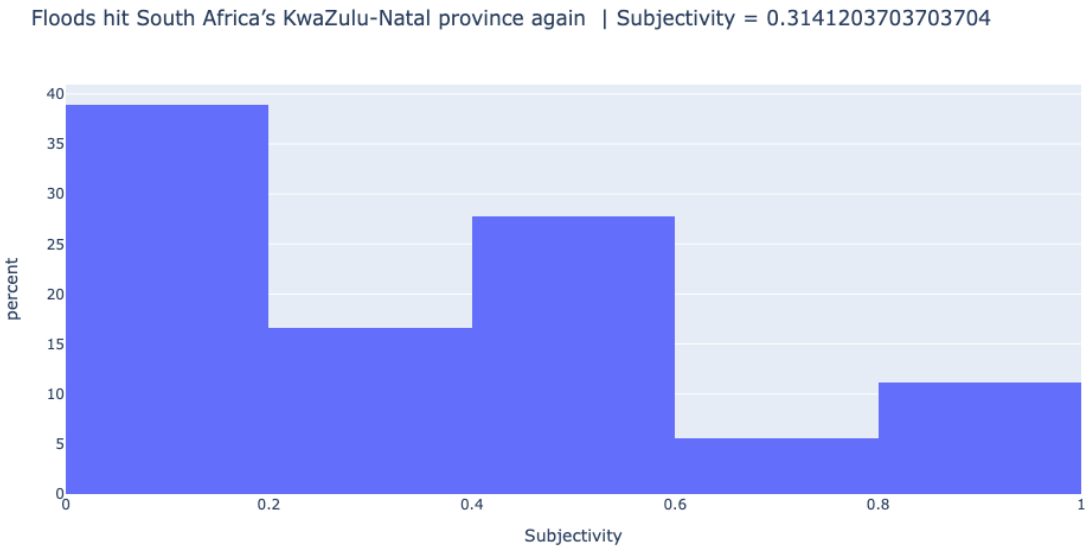
Article 10 :Southern African nations agree to deploy forces to Mozambique

The model predicted this article to be Negative

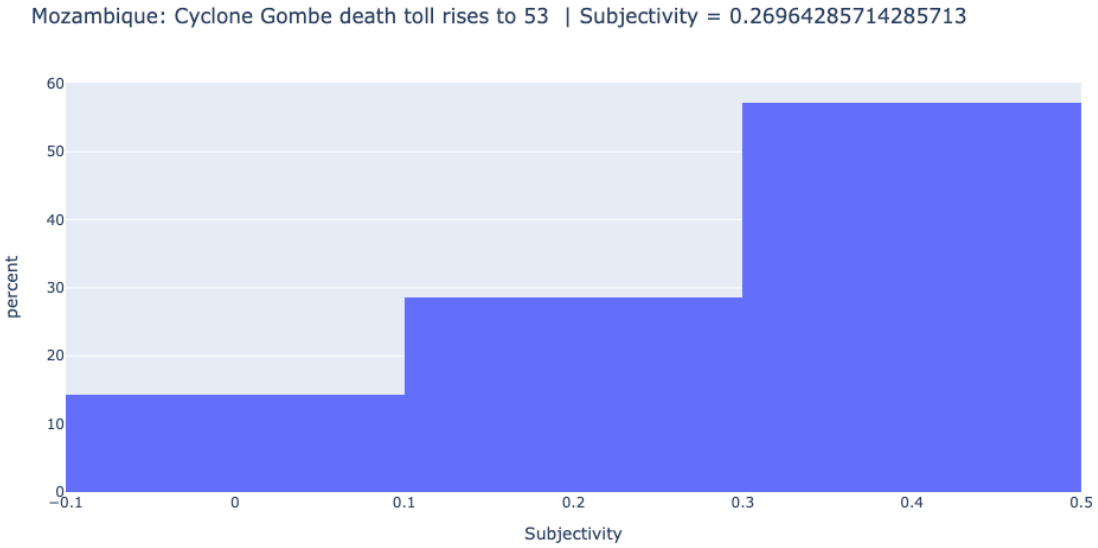


The subjectivity histogram plots we created had article title, subjectivity mean score.

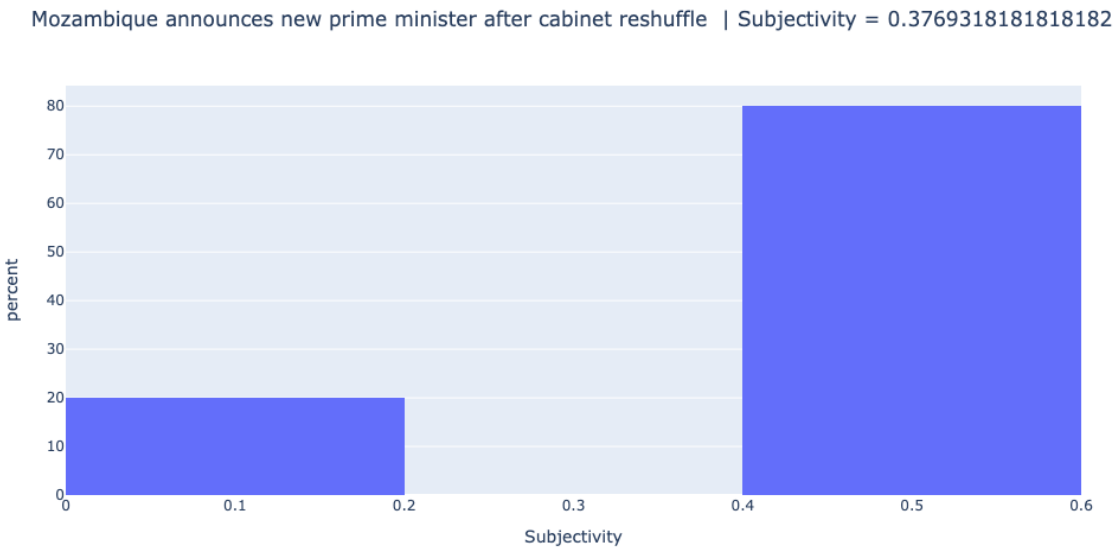
Article 1 :Floods hit South Africa’s KwaZulu-Natal province again



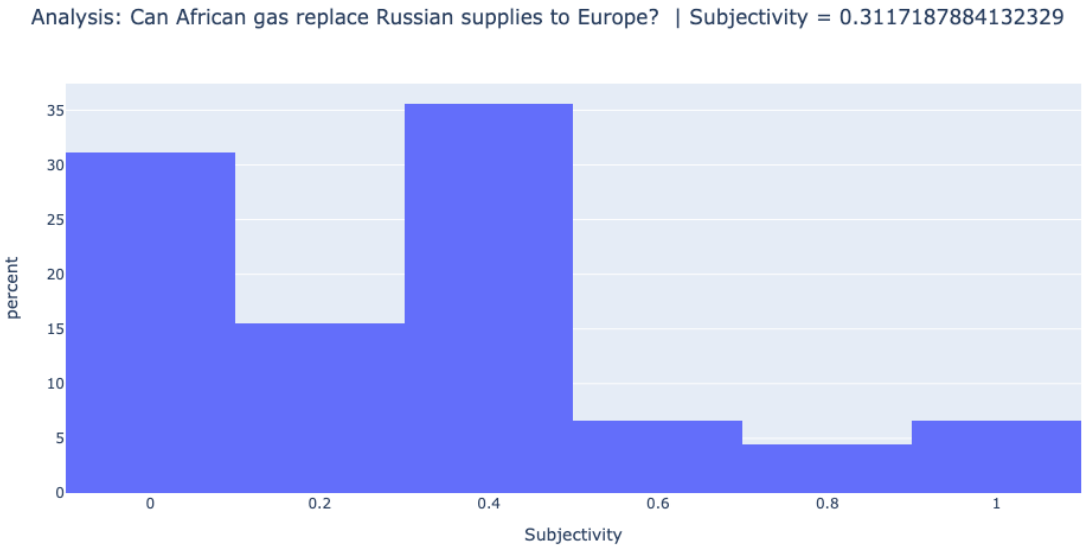
Article 2 :Mozambique: Cyclone Gombe death toll rises to 53



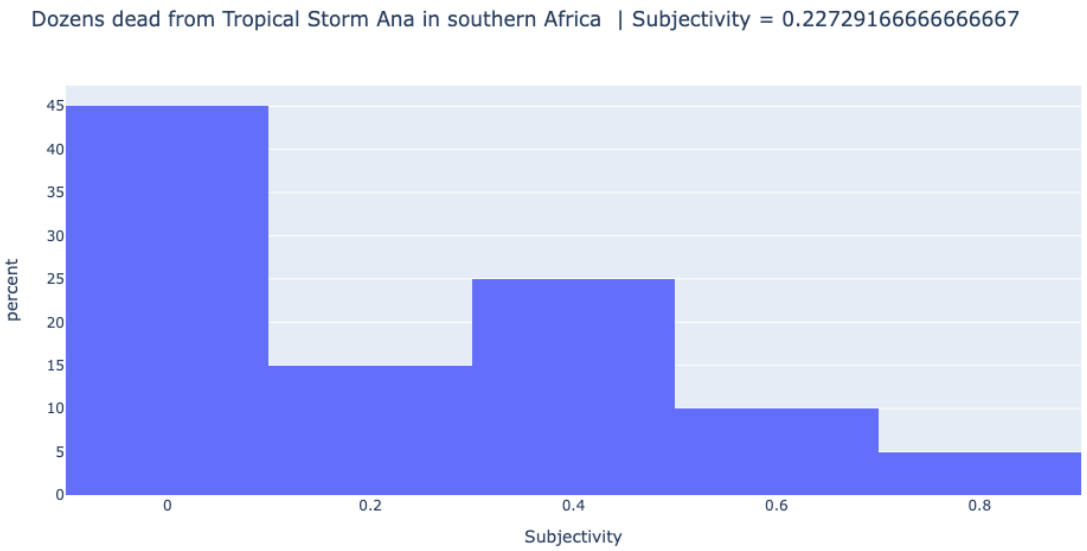
Article 3 :Mozambique announces new prime minister after cabinet reshuffle



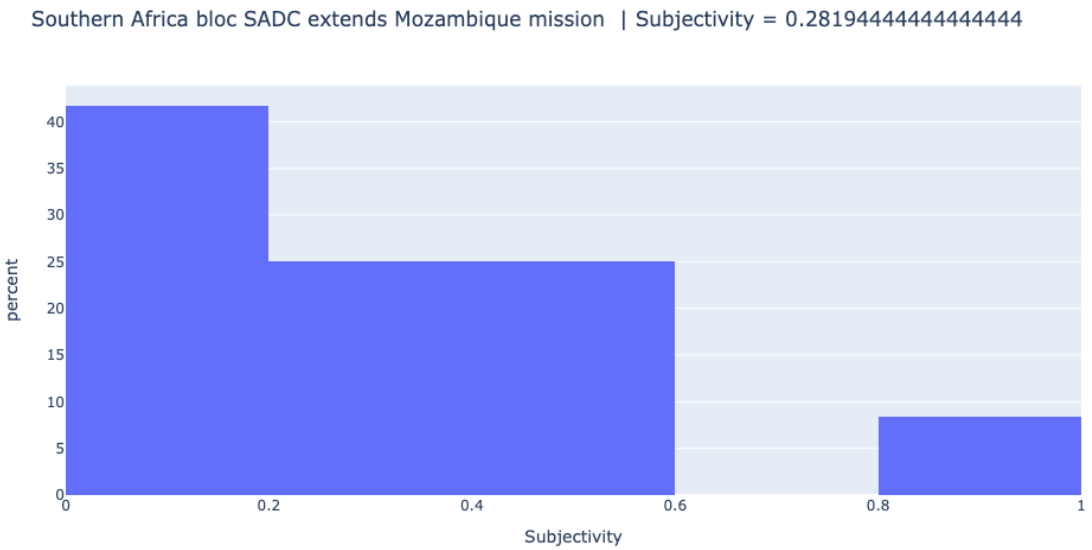
Article 4 :Analysis: Can African gas replace Russian supplies to Europe?



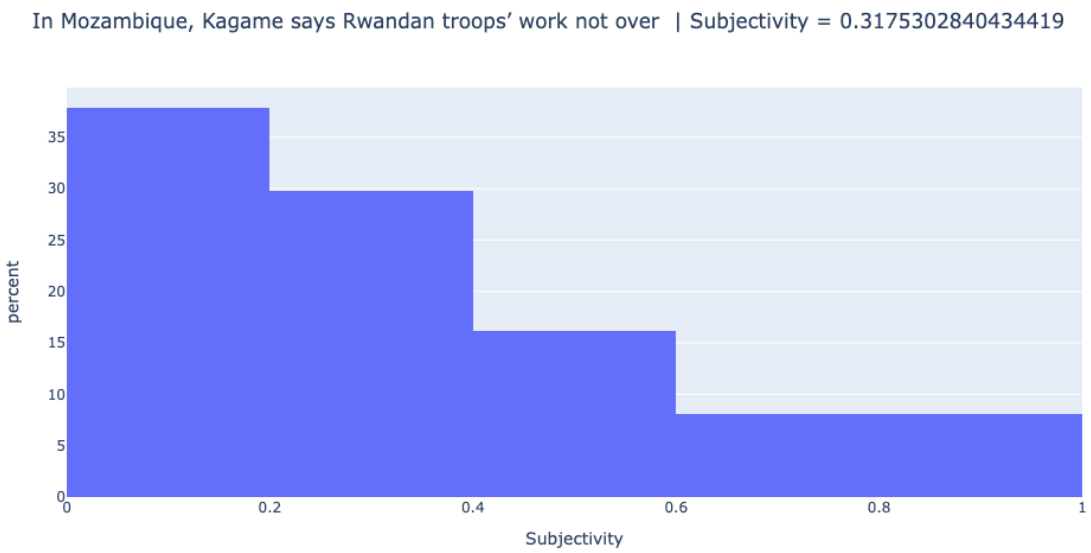
Article 5 :Dozens dead from Tropical Storm Ana in southern Africa



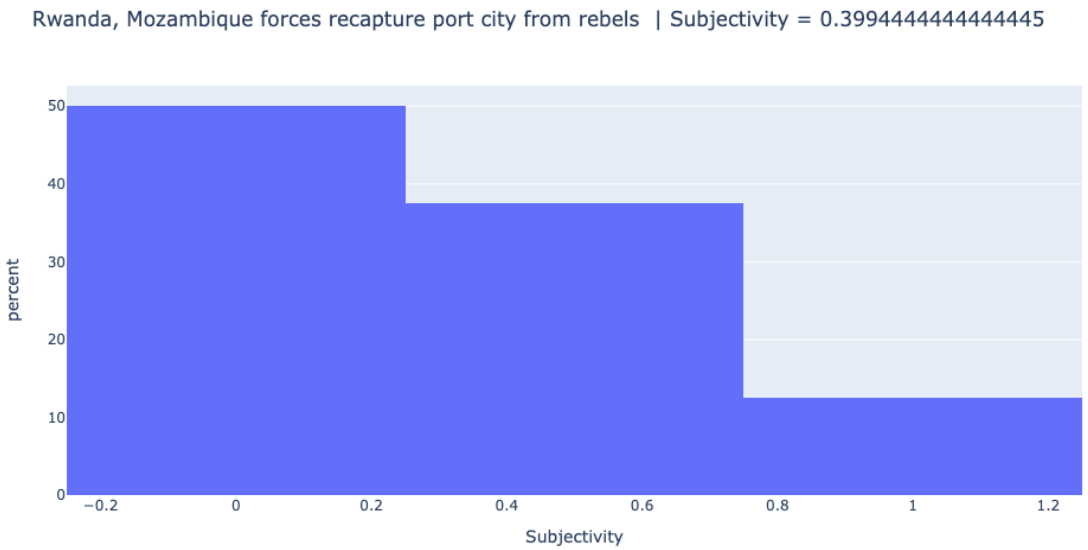
Article 6 :Southern Africa bloc SADC extends Mozambique mission



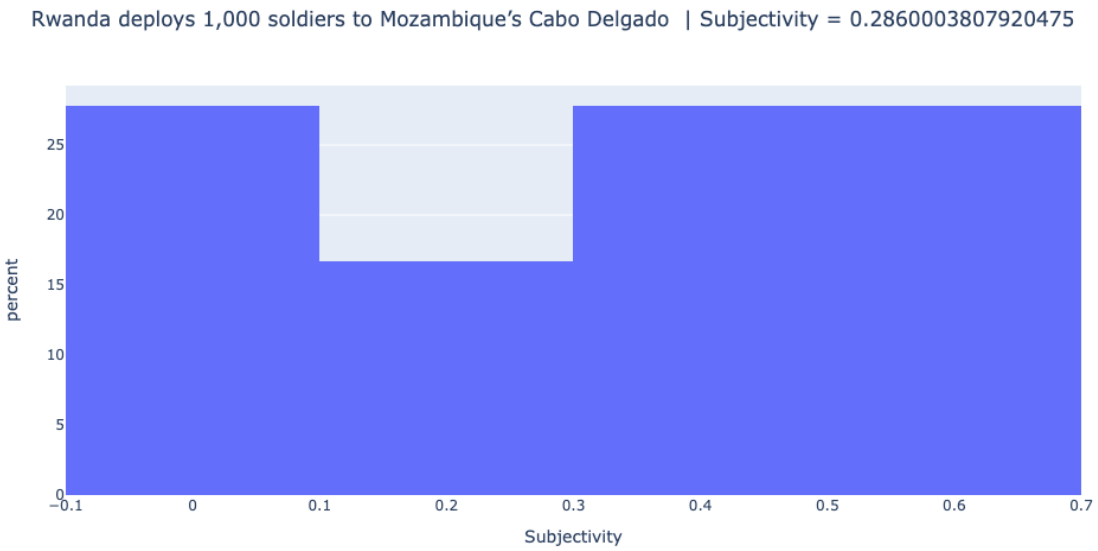
Article 7 :In Mozambique, Kagame says Rwandan troops’ work not over



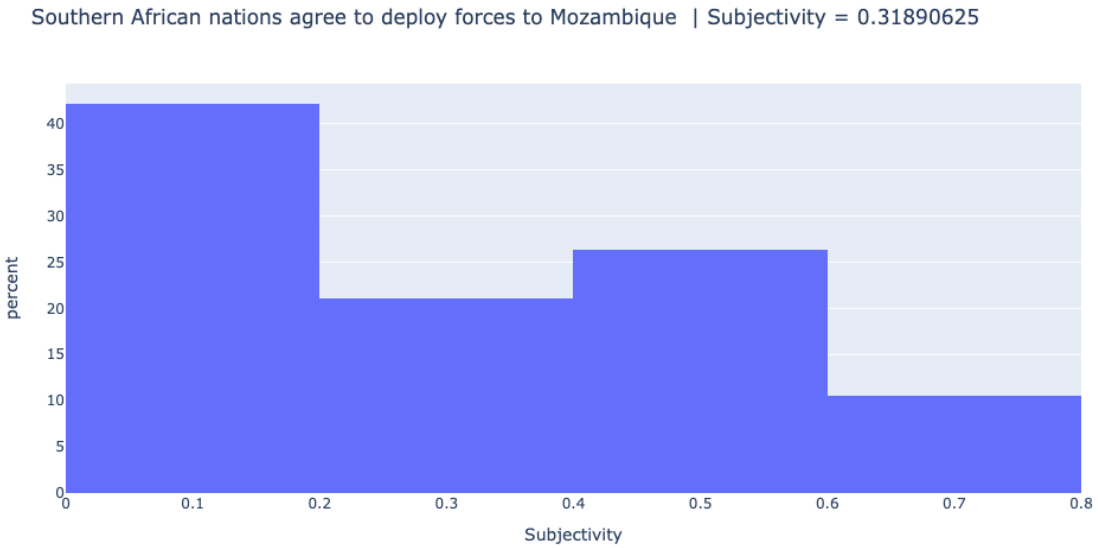
Article 8 :Rwanda, Mozambique forces recapture port city from rebels



Article 9 :Rwanda deploys 1,000 soldiers to Mozambique’s Cabo Delgado



Article 10 :Southern African nations agree to deploy forces to Mozambique

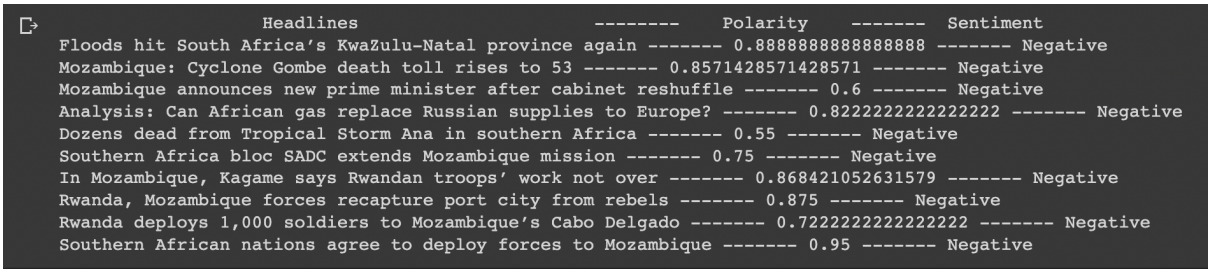


- **Method 2: Using LSTM model**

  This model was built from scratch. As I did not get any news article sentiments dataset. I used the airline review sentiment analysis dataset. The model is an LSTM model classifying the sentences into two classes(1- negative and 0-positive). We tokenized the sentences into words using TensorFlow tokenizer. Then we converted each word into a sequence using tokenizer.texts_to_sequences(). We trained our LSTM model for 5 epochs. Then, I used this model on our article sentences and got its classification score(0 to 1). Then I took a mean of

all the sentences scores and got the polarity score for each article. I printed the article title, polarity score and sentiment on console.

Below image of console output.



```
                    Headlines                        --------    Polarity    -------  Sentiment
Floods hit South Africa’s KwaZulu-Natal province again ------- 0.8888888888888888 ------- Negative
Mozambique: Cyclone Gombe death toll rises to 53 ------- 0.8571428571428571 ------- Negative
Mozambique announces new prime minister after cabinet reshuffle ------- 0.6 ------- Negative
Analysis: Can African gas replace Russian supplies to Europe? ------- 0.8222222222222222 ------- Negative
Dozens dead from Tropical Storm Ana in southern Africa ------- 0.55 ------- Negative
Southern Africa bloc SADC extends Mozambique mission ------- 0.75 ------- Negative
In Mozambique, Kagame says Rwandan troops’ work not over ------- 0.868421052631579 ------- Negative
Rwanda, Mozambique forces recapture port city from rebels ------- 0.875 ------- Negative
Rwanda deploys 1,000 soldiers to Mozambique’s Cabo Delgado ------- 0.7222222222222222 ------- Negative
Southern African nations agree to deploy forces to Mozambique ------- 0.95 ------- Negative
```

In summary our first solution using Textblob came very close to exact solution, when I read and classified the articles and compared my results with textblob results. But the LSTM performed poorly, it classified all the articles as negative. The model was overfitting and this happened because the dataset it was trained on had only reviews, which were just few word sentences. So, when the model saw any negative word in the sentence it classified it as Negative, without understanding the whole context. I did the LSTM solution, just to show how we can do sentimental analysis from scratch using custom dataset. This solution was done just for showing that we can build a specific sentiment analysis model.

Improvements: Since LSTM is built on feedback connections and can process both single data points and sequence of data, if it is trained on a better annotated & balanced dataset, the performance can be improved. It can out-qualify text blob and can be used for specific purposes.