User Manual

Project Title: PandemicAnalysis: A data analysis platform for COVID-19 using Twitter

Student 1 Name: Jaime de Vivero Woods

Student ID: 19447494

Student 2 Name: Alen Tom Joy

Student ID: 18313576

Stream: CASE4

Project Supervisor Name: Renaat Verbruggen

Document Type: User Manual

Date of completion: 06/05/2023

Abstract

This project performs sentiment analysis on publicly available tweets about the COVID-19 pandemic to determine how negative or positive they were (sentiment). The tweets were stored in a database and a number of classifier models were evaluated to find the most accurate. We developed a Django web-app to provide graphs and statistics of the models. The website allows users to select time frames and graphs to display results, the sentiment is then calculated for that time frame and results are displayed.

Installation

The web application runs from a django project. The command “python manage.py runserver” runs the web application. Due to the way the project is developed to increase performance, the user does not need to apply any migrations or install the database for the UI to display the graphs.

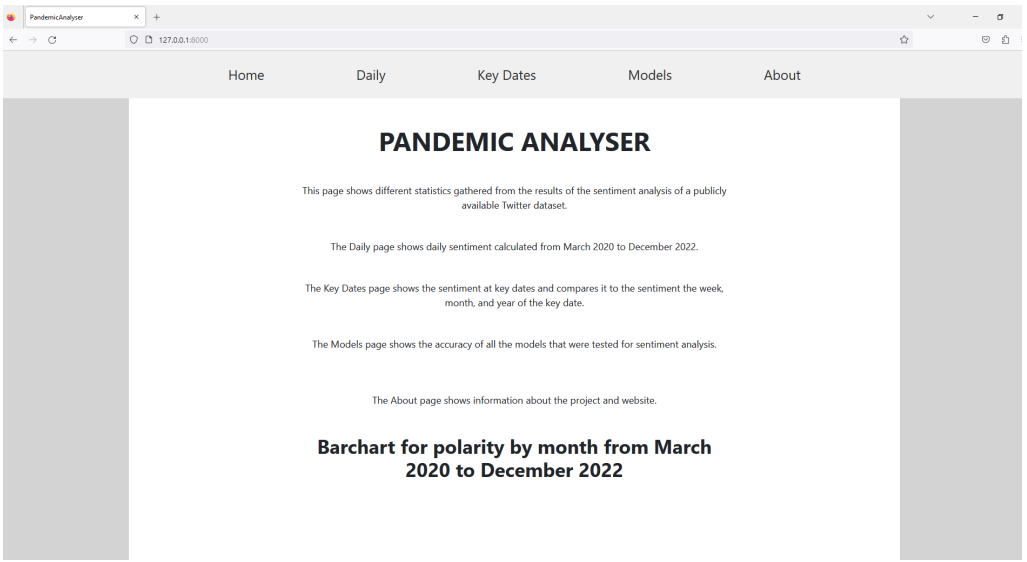
User Guide

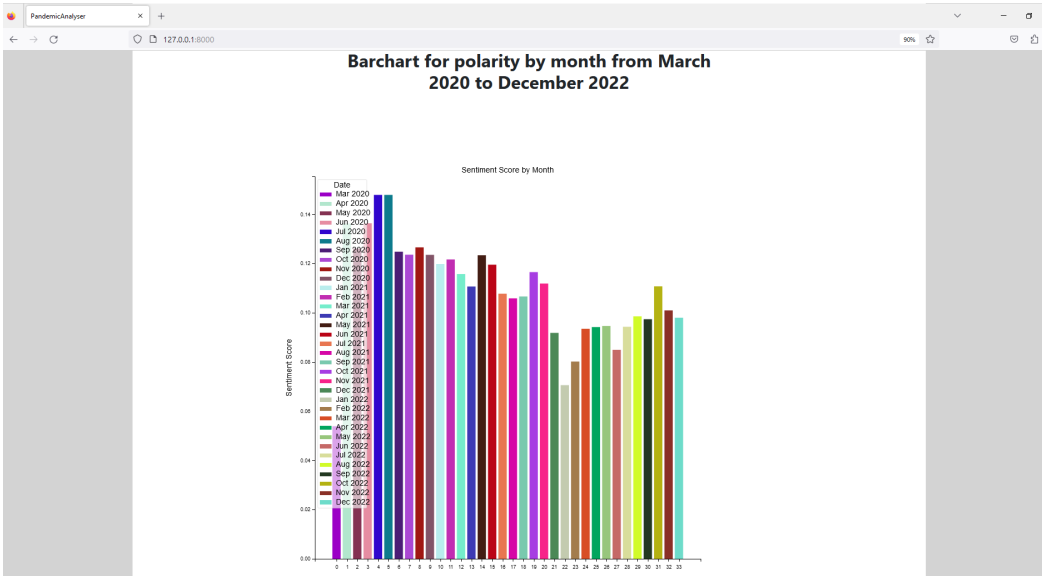
UI Functions

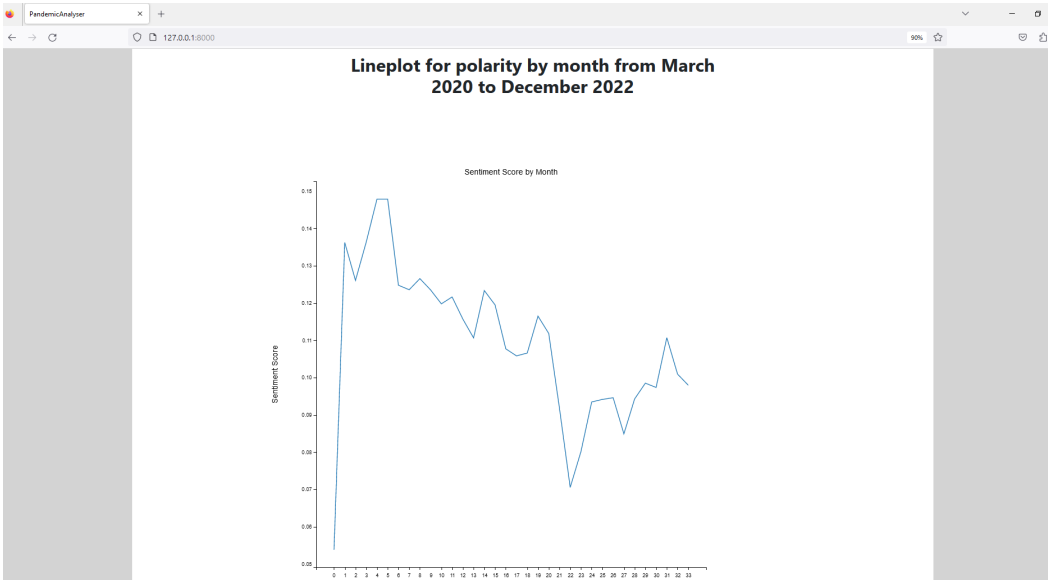
The web application contains functions to view line plots and barchart for all months in the database, view heatmap on a monthly basis, view daily sentiment in a line plot, view the stats of key dates from our results, view different statistics from the developed models and lastly view information about the project. This is shown in five different pages, “Home”, “Daily”, “Key Dates”, “Models” and “About”.

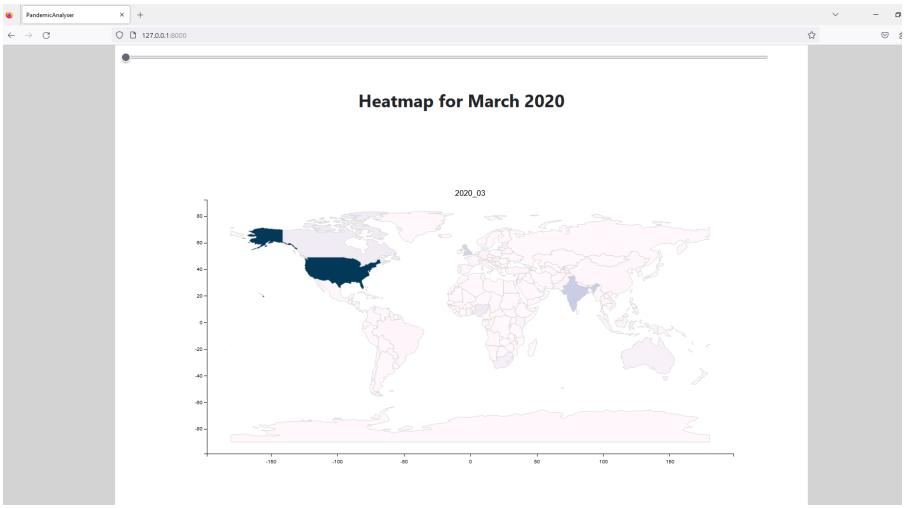
Home Page

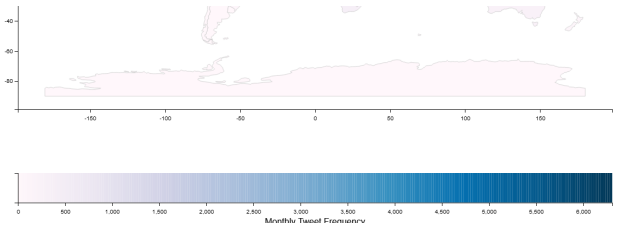
The “Home” page contains three different functionalities. These are a barchart of monthly sentiments, a line plot of monthly sentiments, and lastly, a series of heatmaps showing the total number of tweets per country. The heatmap to be shown can be selected through the use of a slider.




PandemicAnalyser
127.0.0.1:8000
90%
Barchart for polarity by month from March 2020 to December 2022
Sentiment Score by Month
Date
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Sep 2020
Oct 2020
Nov 2020
Dec 2020
Jan 2021
Feb 2021
Mar 2021
Apr 2021
May 2021
Jun 2021
Jul 2021
Aug 2021
Sep 2021
Oct 2021
Nov 2021
Dec 2021
Jan 2022
Feb 2022
Mar 2022
Apr 2022
May 2022
Jun 2022
Jul 2022
Aug 2022
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Sentiment Score


PandemicAnalyser
127.0.0.1:8000
90%
Lineplot for polarity by month from March 2020 to December 2022
Sentiment Score by Month
Sentiment Score

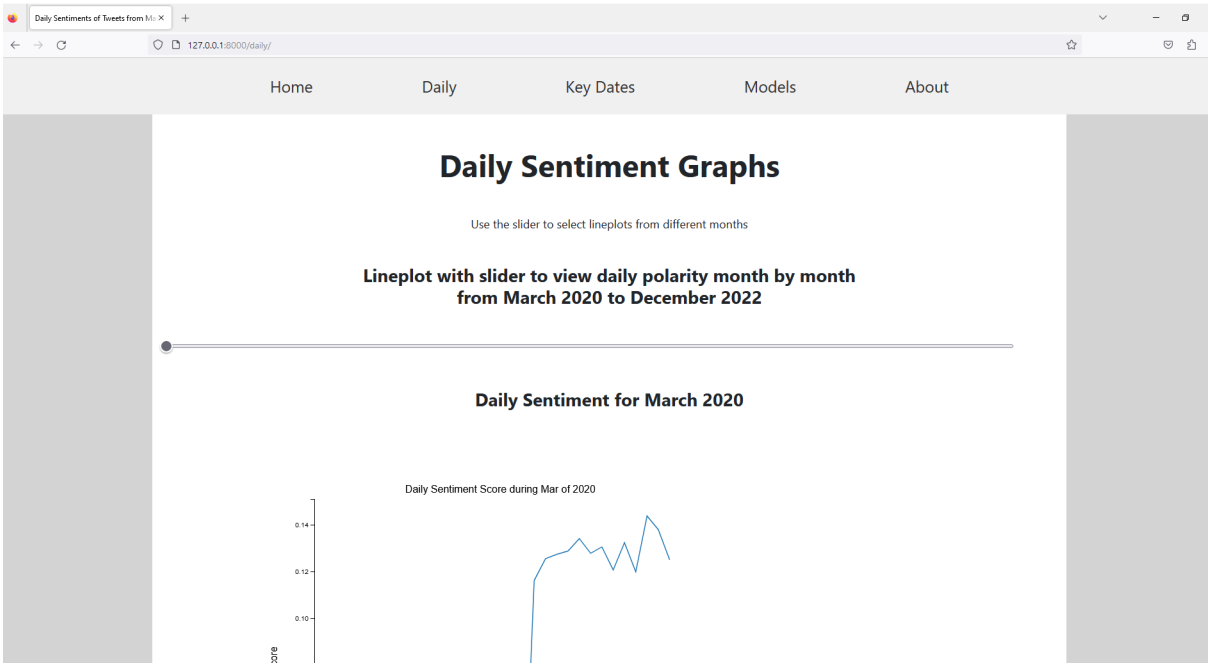

PandemicAnalyser
127.0.0.1:8000
Heatmap for March 2020
2020_03

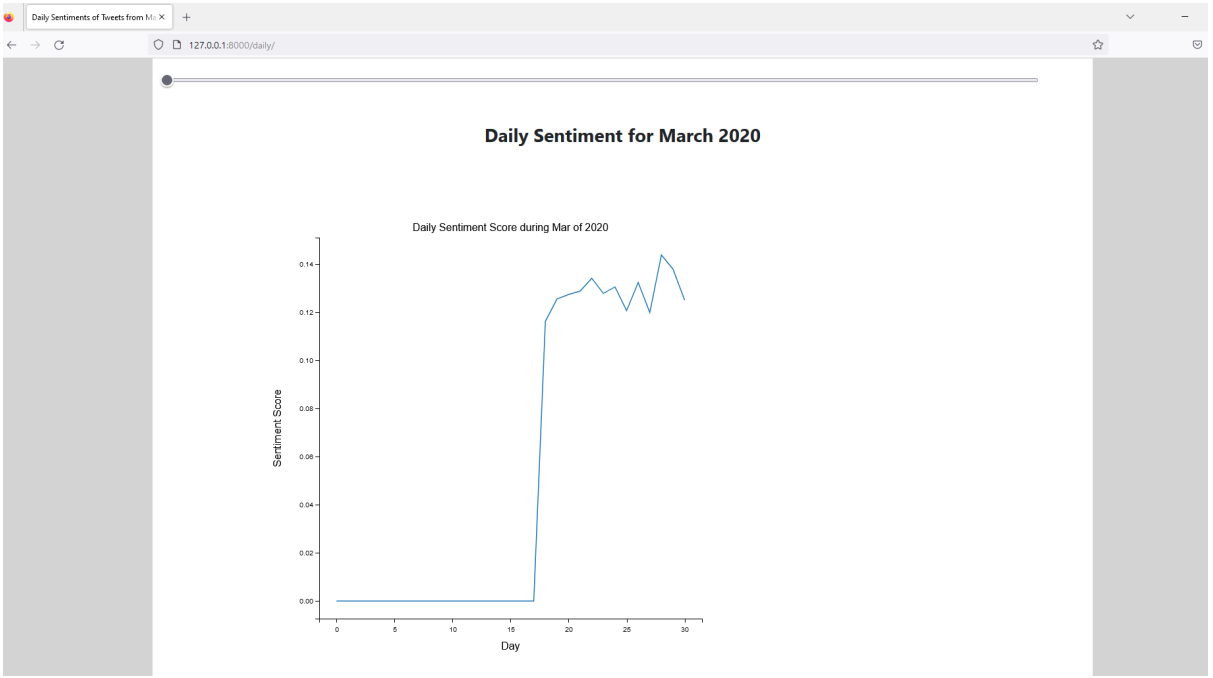

-40
-60
-80
-150
-100
-50
0
50
100
150

0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
Monthly Tweet Frequency

Daily Page

The “Daily” page contains two different functionalities. The first is a series of line plots that show the polarity of every day of a month. Specific months can be selected through the use of a slider. The second functionality is a table showing the major changes in sentiment, the table contains four tables: “Date”, “Change”, “Possible Reasoning” and “News Article”. “Date” shows the date of the major change, “Change” is the percentage of change over the time range. “Possible Reasoning” shows a major news headline that could explain the change. “News Article” shows the url link to the news headline shown in “Possible reasoning”.





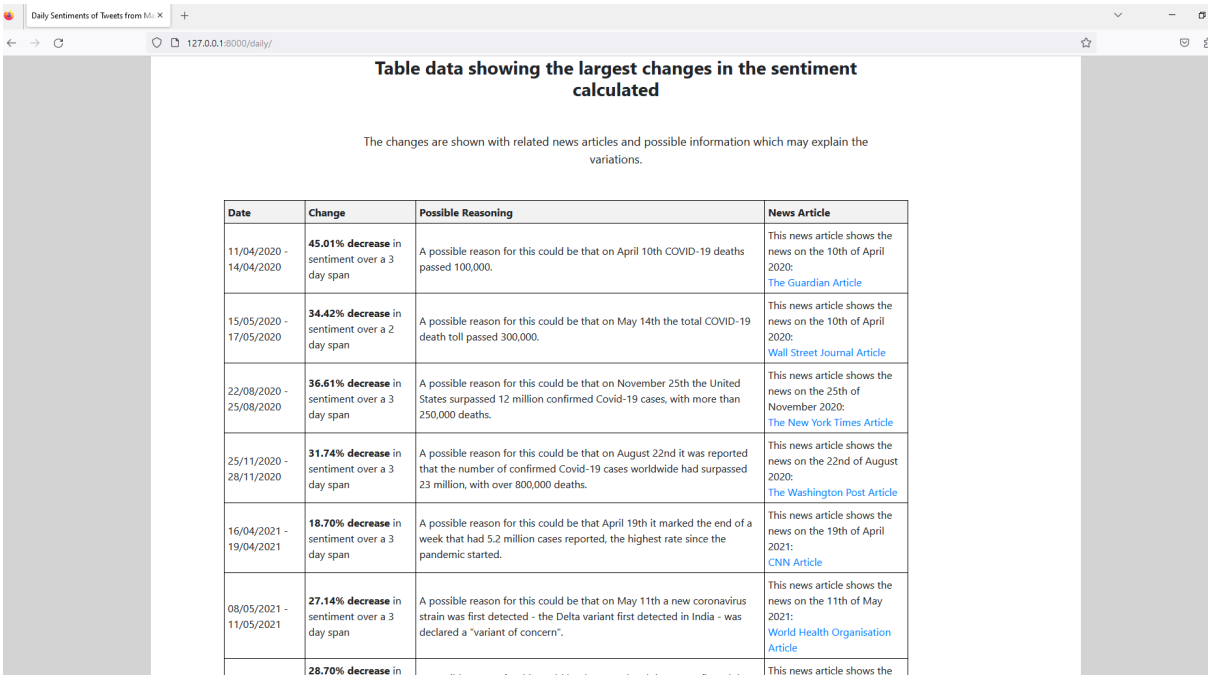


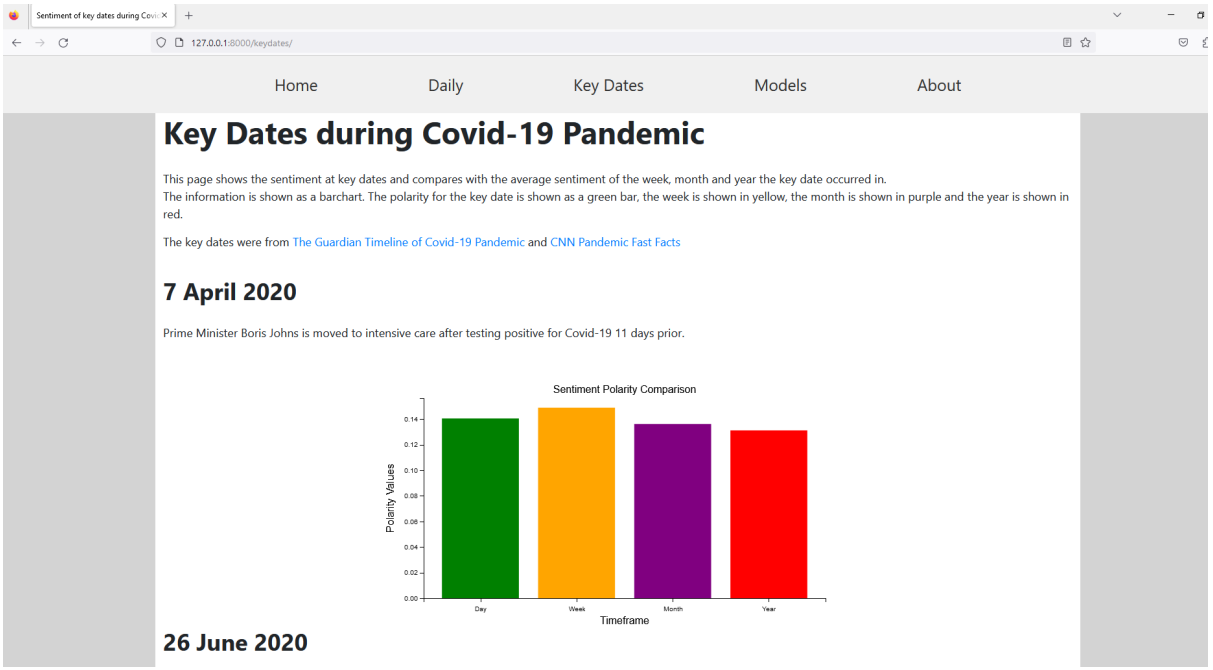
Table data showing the largest changes in the sentiment calculated

The changes are shown with related news articles and possible information which may explain the variations.

Date	Change	Possible Reasoning	News Article
11/04/2020 - 14/04/2020	**45.01% decrease** in sentiment over a 3 day span	A possible reason for this could be that on April 10th COVID-19 deaths passed 100,000.	This news article shows the news on the 10th of April 2020: The Guardian Article
15/05/2020 - 17/05/2020	**34.42% decrease** in sentiment over a 2 day span	A possible reason for this could be that on May 14th the total COVID-19 death toll passed 300,000.	This news article shows the news on the 10th of April 2020: Wall Street Journal Article
22/08/2020 - 25/08/2020	**36.61% decrease** in sentiment over a 3 day span	A possible reason for this could be that on November 25th the United States surpassed 12 million confirmed Covid-19 cases, with more than 250,000 deaths.	This news article shows the news on the 25th of November 2020: The New York Times Article
25/11/2020 - 28/11/2020	**31.74% decrease** in sentiment over a 3 day span	A possible reason for this could be that on August 22nd it was reported that the number of confirmed Covid-19 cases worldwide had surpassed 23 million, with over 800,000 deaths.	This news article shows the news on the 22nd of August 2020: The Washington Post Article
16/04/2021 - 19/04/2021	**18.70% decrease** in sentiment over a 3 day span	A possible reason for this could be that April 19th it marked the end of a week that had 5.2 million cases reported, the highest rate since the pandemic started.	This news article shows the news on the 19th of April 2021: CNN Article
08/05/2021 - 11/05/2021	**27.14% decrease** in sentiment over a 3 day span	A possible reason for this could be that on May 11th a new coronavirus strain was first detected - the Delta variant first detected in India - was declared a "variant of concern".	This news article shows the news on the 11th of May 2021: World Health Organisation Article
	28.70% decrease in		This news article shows the

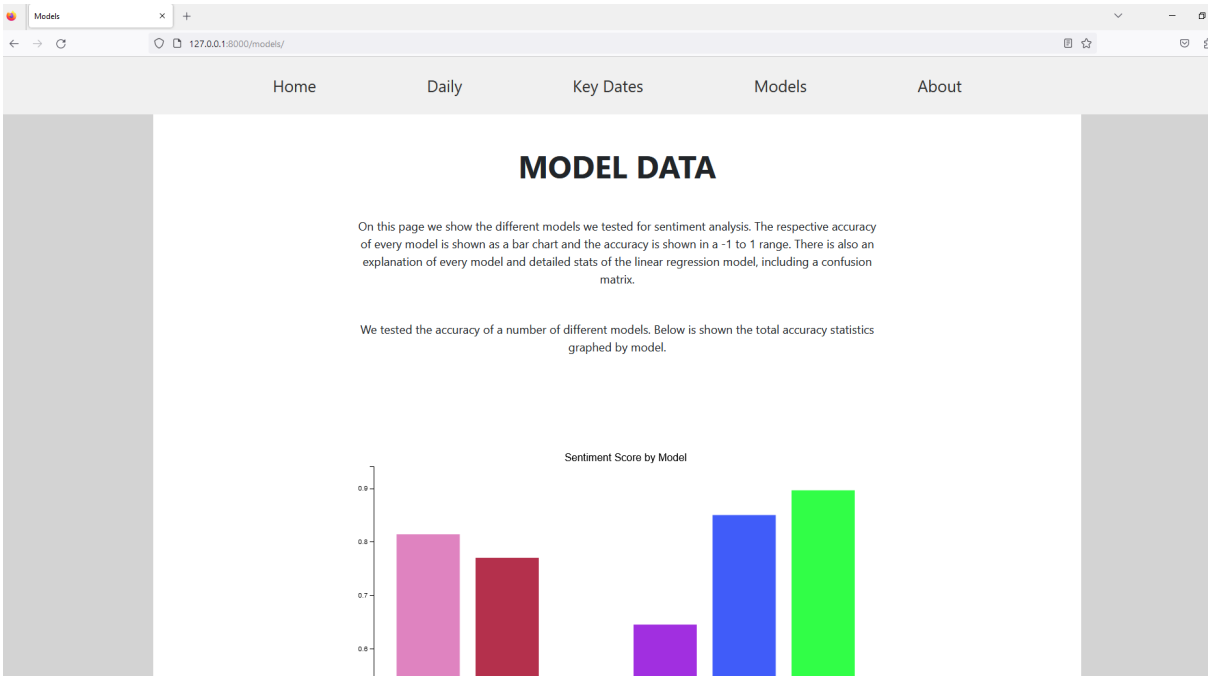
Key Dates Page

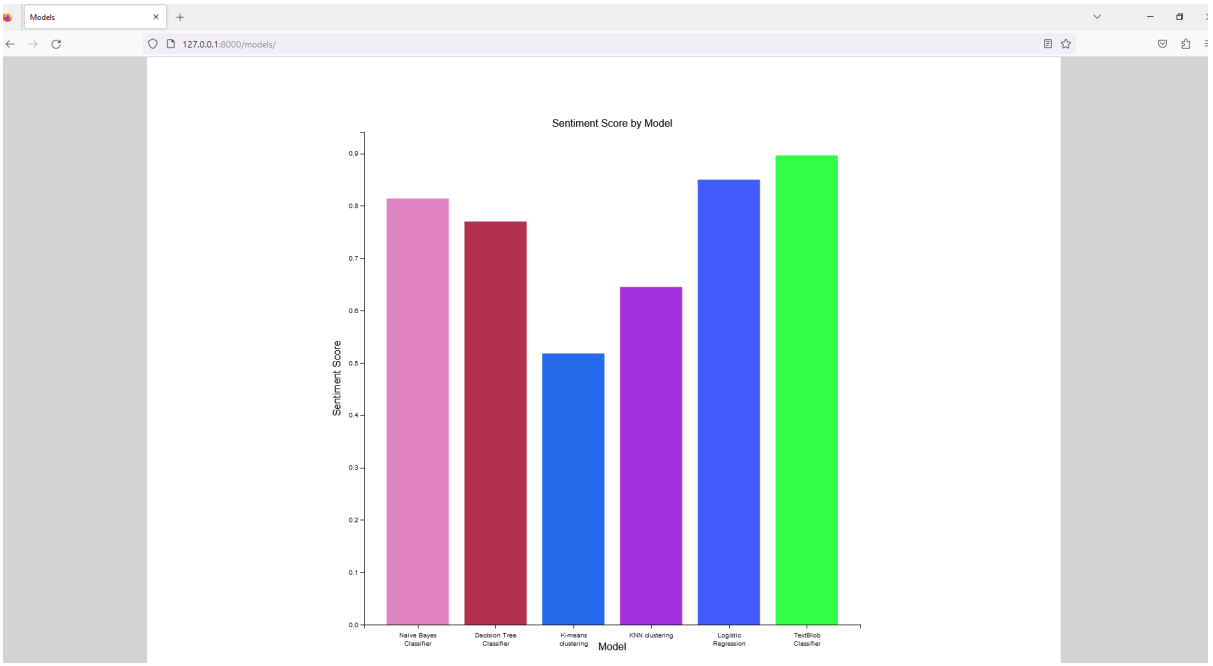
The “Key Dates” page shows the sentiment at key dates and compares with the average sentiment of the week, month and year the key date occurred in. The information is shown as a barchart, with a short text above the graph explaining what happened on the given date. The polarity for the key date is shown as a green bar, the week is shown in yellow, the month is shown in purple and the year is shown in red.



Models Page

The “Models” page contains many different graphs, statistics and information of the developed models. The accuracy of each model is shown as a bar chart. There is an explanation of the algorithm of each model, the accuracy shown in our results and then a confusion matrix and ROC curve is displayed for each model section.





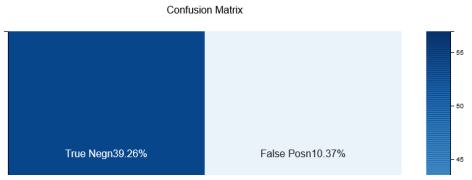
Naïve Bayes Classifier

Classification is where a model or classifier is constructed to predict categorical labels. It is the process of learning a model that can differentiate between different classes of data. Naïve Bayes is a classifier method that applies Bayes' Theorem with the assumption that the predictors used are independent. Bayes' Theorem uses conditional probabilities among predictors to determine the probability of a class variable occurring, given that another variable has already taken place.

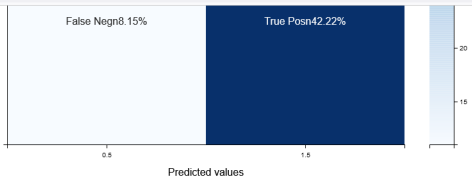
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We utilized the Text Blob library to develop a Naïve Bayes Classifier. We generated a set of training and test data and refined it to achieve the highest levels of accuracy. Our final results demonstrated an overall accuracy rate of **81.4%**.

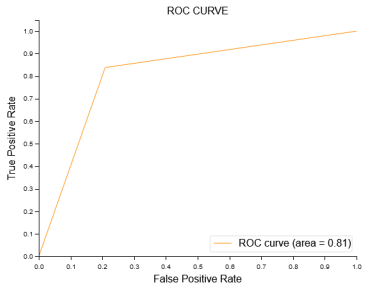
The confusion matrix for the Naïve Bayes classifier developed is shown below:



Models
127.0.0.1:8000/models/



The ROC curve for the Naïve Bayes classifier developed is shown below:



Decision Tree Classifier

About Page

The “About” page contains some information about how the project idea came about, who the authors of the page are, an explanation of the idea and motivation of the project and advantages of using Twitter data.