Climate Change Belief Analysis

Presented By:

SWAT Analytics

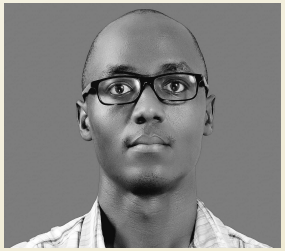
SWAT Analytics Consulting

Our team members

Jacqueline wangombe

Winfred Mutinda

Emmanuel Okoro



Jack Kamire

Titus Wanjohi

Project Outline

1. Introduction
2. Problem Statement
3. Data
4. Exploratory Data Analysis
5. Data Engineering
6. Modeling
7. Conclusion

1 Introduction

Have you noticed any of the following occurrences:-

1. Hotter Temperatures
2. Increased storm severity
3. Increased drought frequency
4. Shifts in rainfall patterns

What could be causing this:-

Climate Change:- a change in the pattern of weather, and related changes in oceans, land surfaces and ice sheets, occurring over time scales of decades or longer

Key drivers of climate change:-

Greenhouse gases

Aerosol emissions

Land use change

Problem Statement

The problem:-

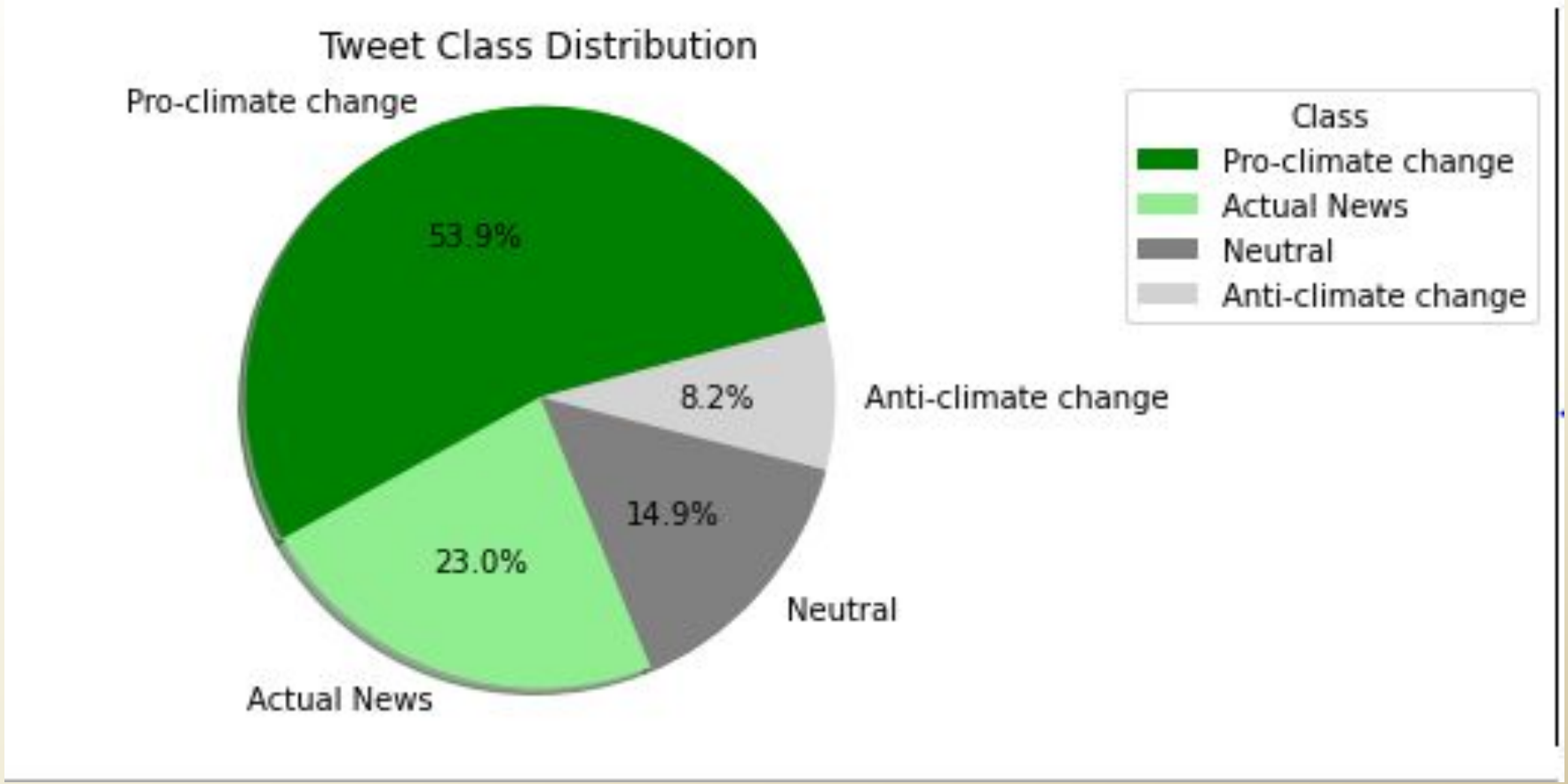
- In an effort to reduce their carbon footprint, our clients are offering products/services that are environmentally friendly and sustainable.

- Is climate change real? What is the prevailing opinion in the public domain?

The solution:-

- Our team conducted a tweet sentiment analysis by creating a classifier model that predicts whether a tweet is pro-climate change or not.

Data


Tweet Class Distribution
Pro-climate change
53.9%
8.2%
Anti-climate change
14.9%
23.0%
Neutral
Actual News
Class
Pro-climate change
Actual News
Neutral
Anti-climate change

Exploratory Data Analysis (EDA)

What is Exploratory data analysis (EDA)?

This is the process of analysing and investigating data sets and summarizing their main characteristics, often employing both non-graphical and graphical methods.

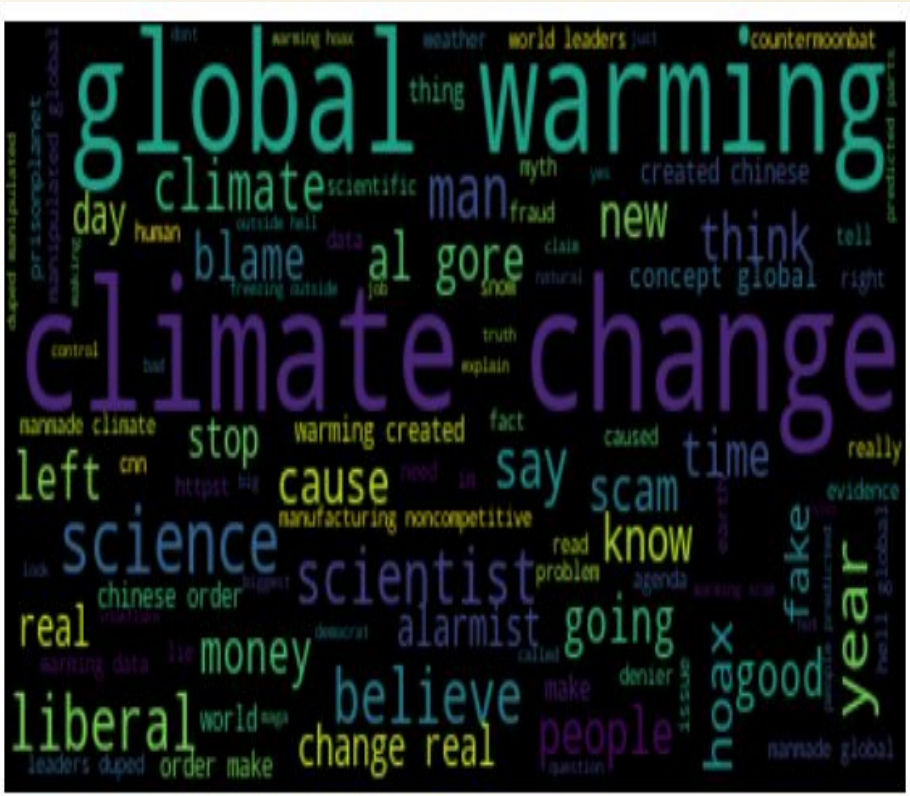
Why is conducting EDA important?

It aids in determining how best to manipulate data to get the required answers, expose trends, patterns, and relationships that are not readily apparent i.e. get insights into the dataset.

Exploratory Data Analysis (EDA)

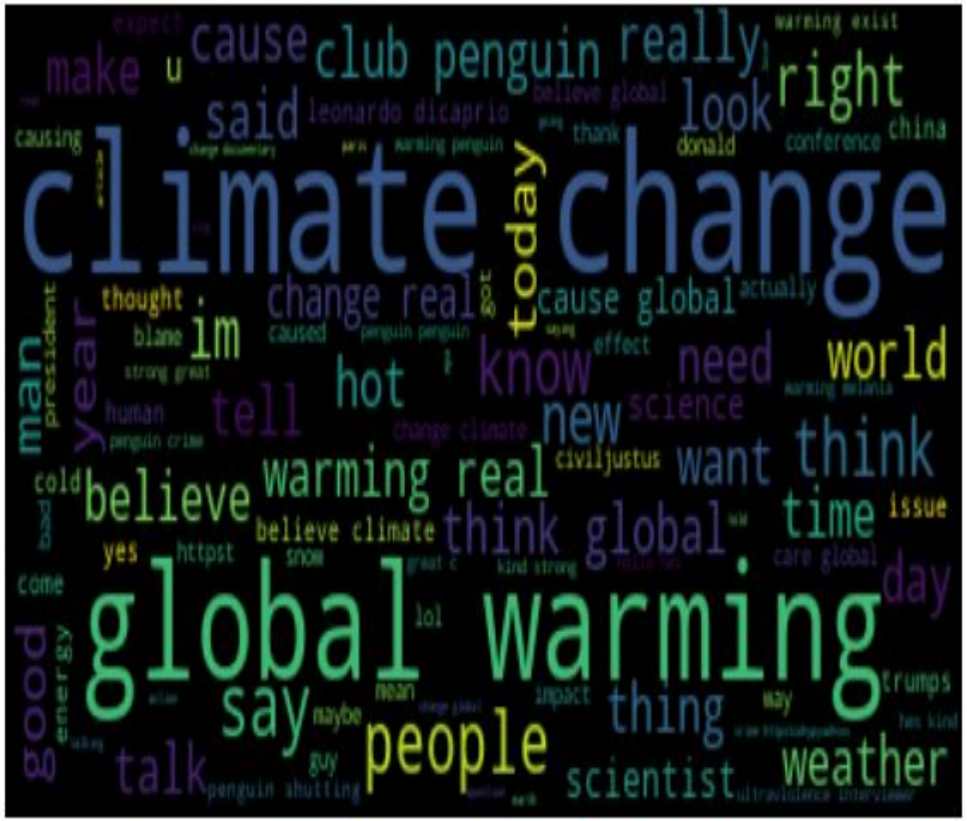
Pro Climate Change

Anti Climate Change



Exploratory Data Analysis (EDA)

Neutral



News Information

Exploratory Data Analysis (EDA)

Key Insights:

- The tweets contain punctuation marks and web urls.
- The tweet text is in mixed case version i.e. both lower and uppercase.
- The tweet words contain stopwords.
- There is uneven distribution of observations among the classes in the sentiment column i.e. class imbalance.
- The key words in the word cloud include climate change, global warming, believe climate, trump.

Data Engineering

Noise Removal

Web-urls, Punctuation, text lowercase, punctuation, # and @

Tokenization, Stemming and Lemmatization

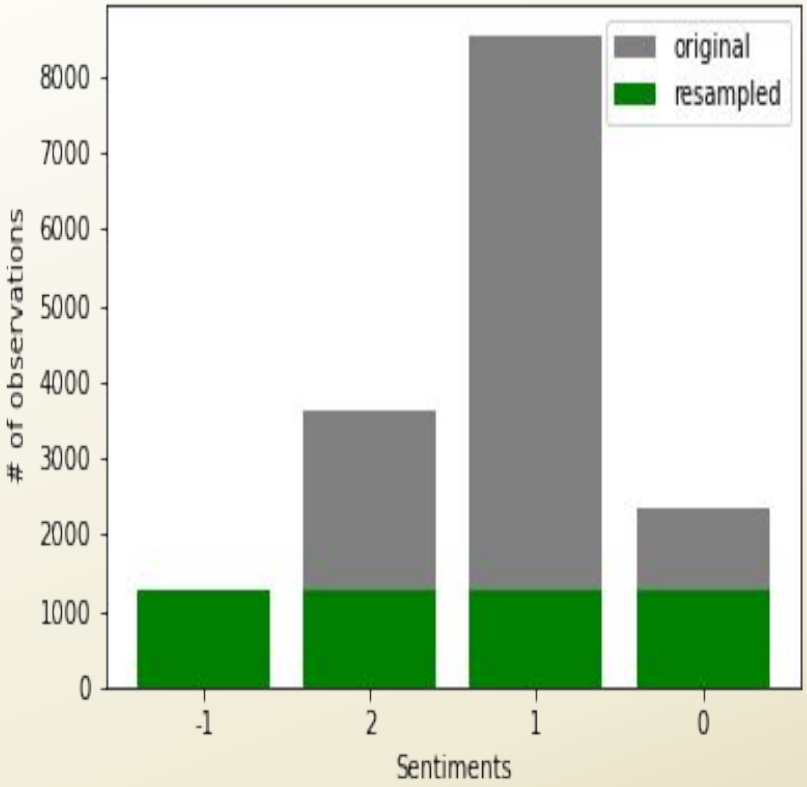
Splitting text and reducing words into their root forms to facilitate vectorization.

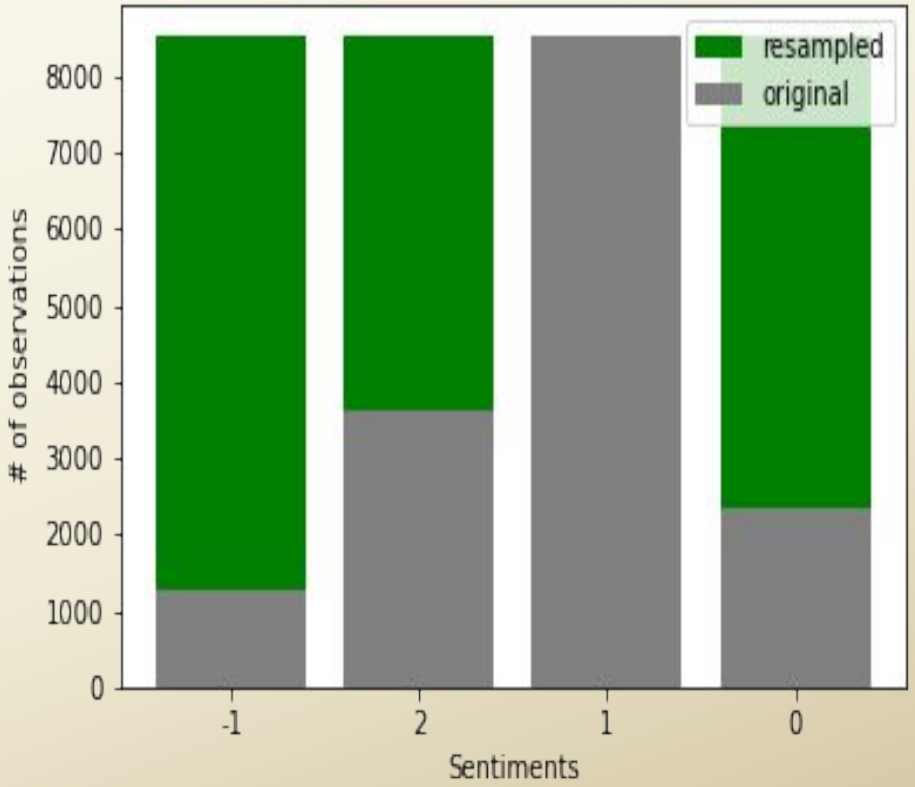
Remove stopwords..

Vectorization, Bag of Words and Data Balancing

- Bag of words aids in extracting features from the text
- Vectorization converts text into numerical form for the model.
- Resampling deals with data imbalance data to increase model accuracy.

Data Engineering



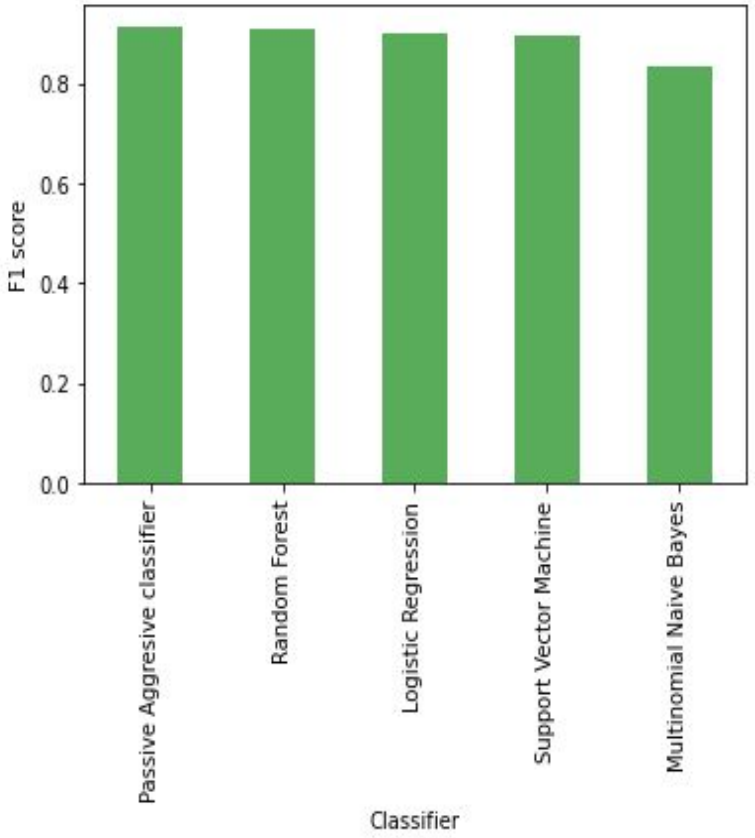


Modeling

Classification models used:

- Logistic regression
- Passive Aggressive classifier
- Support Vector Machines
- Random Forest
- Naive Bayes

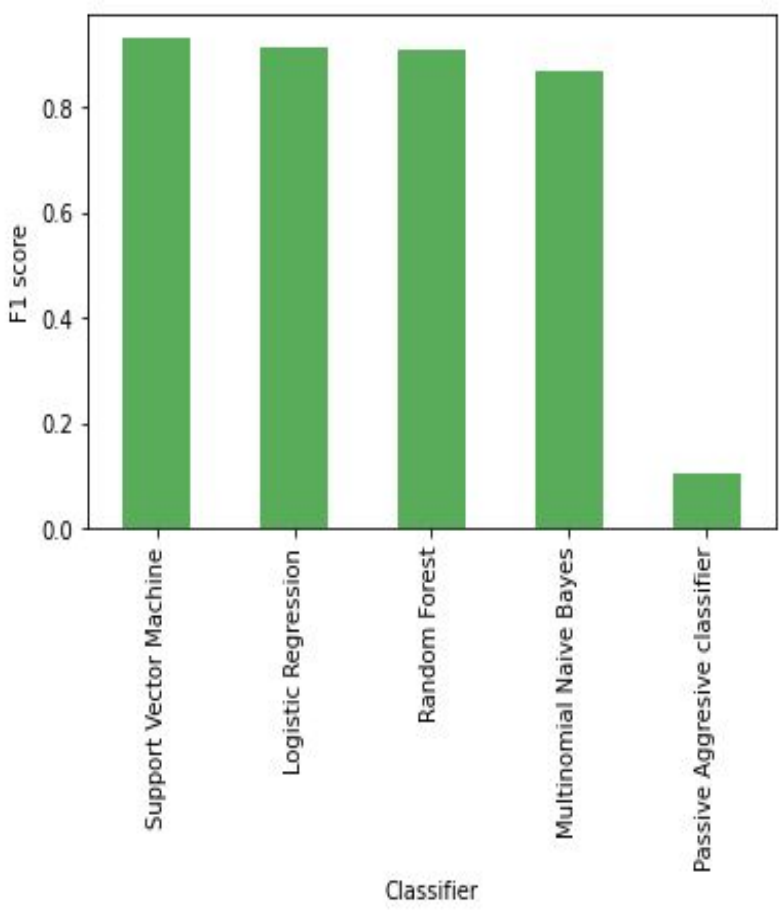
Model Performance (Models with the default hyperparameters)



F1 Scores Comparison

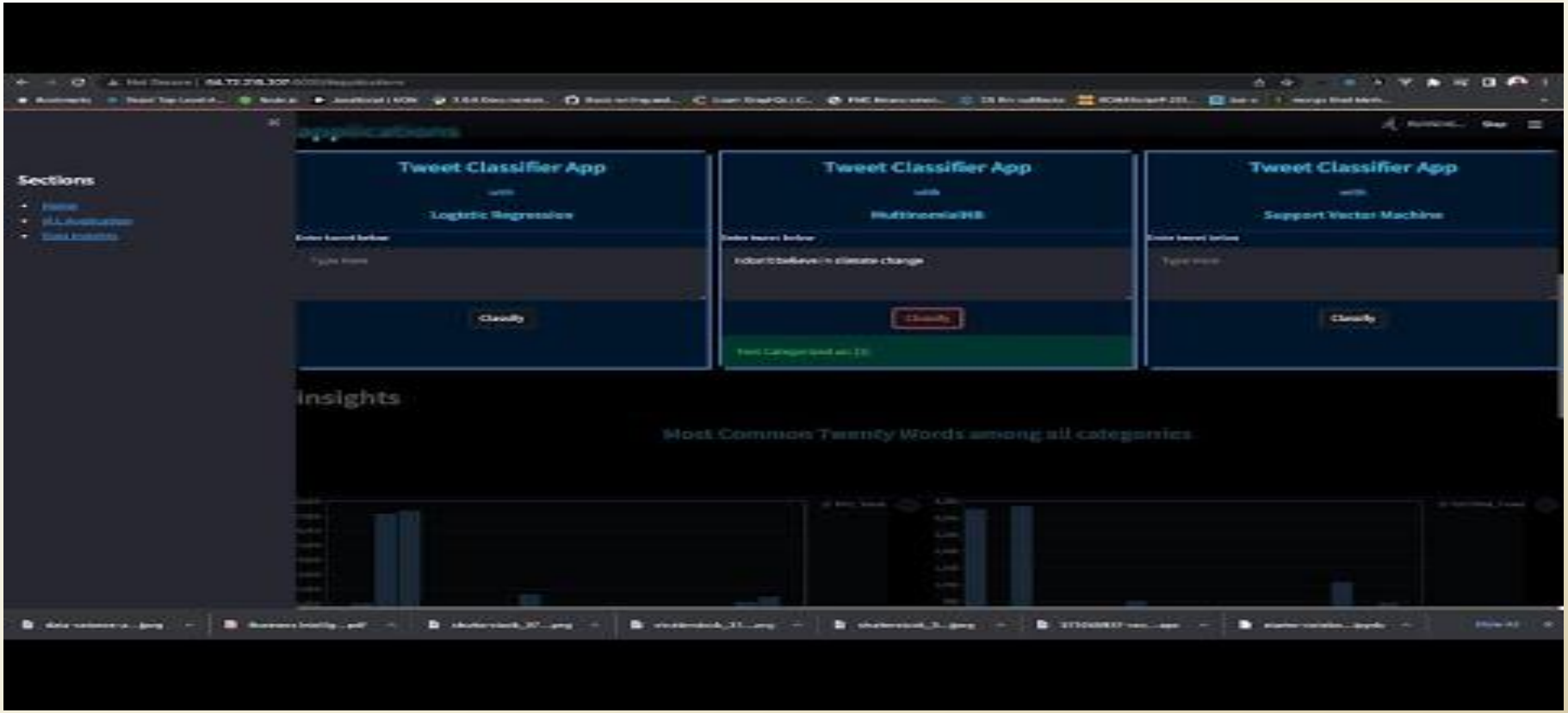
- F1 score is the harmonic mean of precision and recall.
- Maximum F1 Score means precision and recall are equal.
- Models average over 0.8 for the training data.

Model Performance (Models after hyperparameter tuning)



- F1 score is the harmonic mean of precision and recall.
- Maximum F1 Score means precision and recall are equal.
- Models F1 score for all models except Passive Aggressive Classifier was over 0.8.
- Even though the **Logistic Regression** model was not the best with the training data, it was the best with test data and with a score of 0.73.

Tweet Classification Application (Demo Powered by Streamlit)



Conclusion

Key Insights:

- Sentimental analysis has a significant impact on understanding human interactions relying on text based data.
- Model Performance is directly influenced by data preprocessing and feature engineering.
- Based on the Tweet data, majority of sampled Tweets believe in climate change.
- Context impacts sentiment analysis interpretation and application.
- Politics influences climate change discussions and news.

Recommendations and Limitations:

- Timestamps provide context in text data analysis.
- Computing power is a limiting factor in sentiment analysis.
- Natural Language Processing improves with exposure to additional data.
- Metadata about location can inform context given the connection between climate change and politics.

Thank you