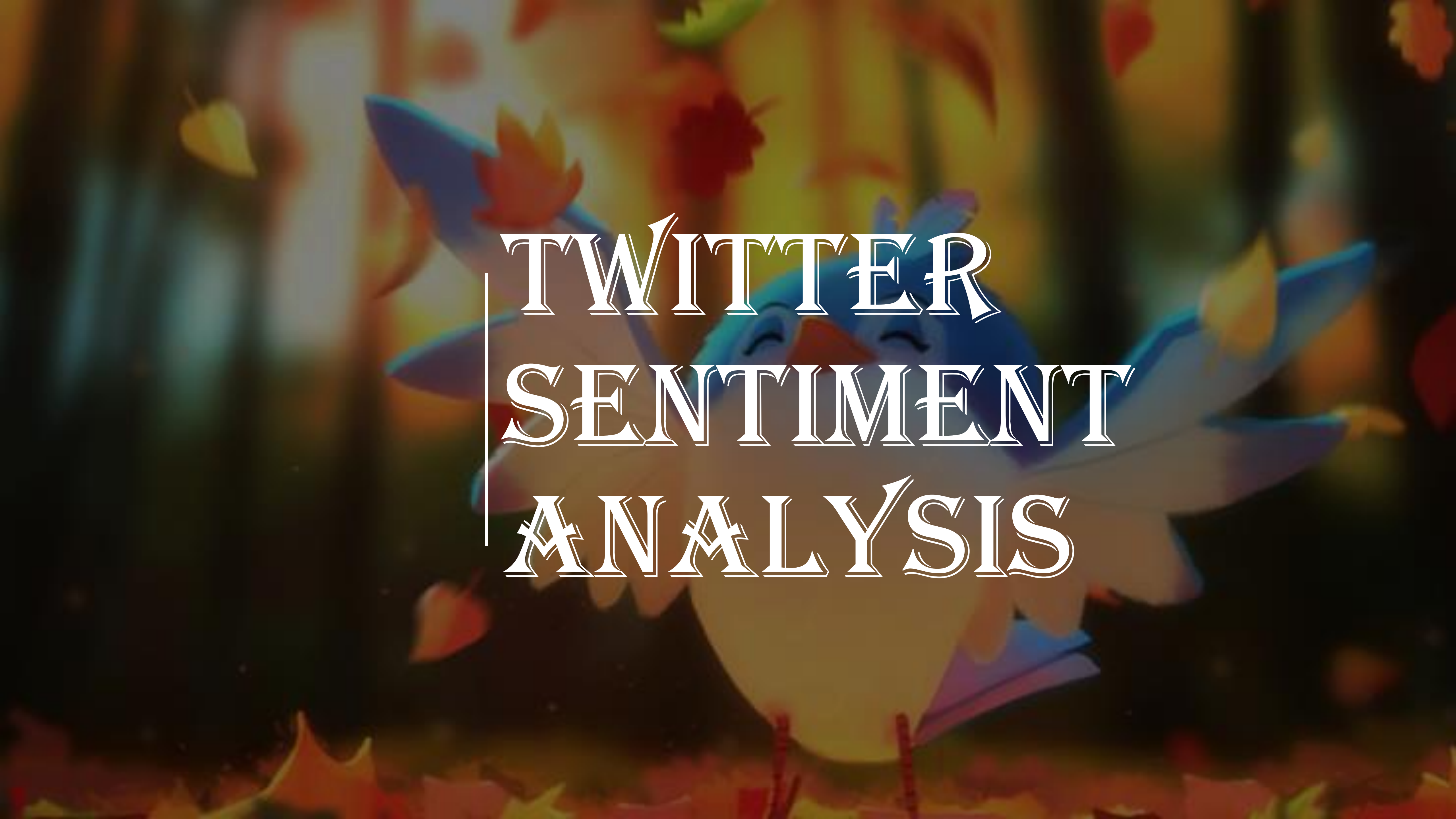


TWITTER SENTIMENT ANALYSIS

OUTLINE

- Sentiment Analysis
- How to get Tweets
- Problem Statement
- Feature Extraction
- Data Cleaning
- Data Visualization

- Machine learning
- What have we Achieve
- Results
- Challenges
- Future Plan
- References

Twitter – What it is?

- Social media platform to exchange views, ideas and thoughts.
- Shows trends and is refreshed in every minute with ever increasing data.
- Keep everyone updated with all the latest news, trends and information.

- Since its ever-growing impact with all the information being posted, this Twitter could result in a dangerous toolkit for the controlling media and governing organizations.
- Another way it may also become a massive opposing body to our society which can imbalance the harmony of this society.

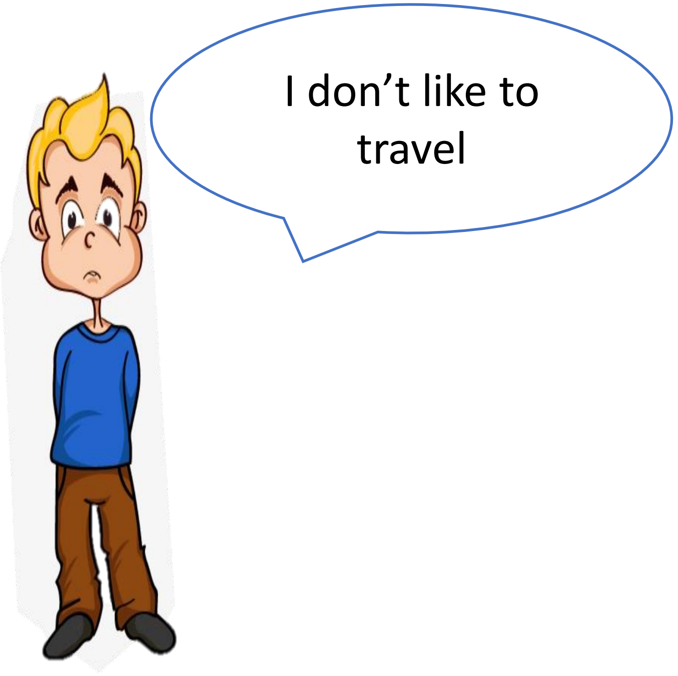
PROBLEM STATEMENT

- We are given a message, decide whether the message is of positive, negative, or neutral sentiment.
- For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

SENTIMENT ANALYSIS

I Love Pizza

Positive Sentiment

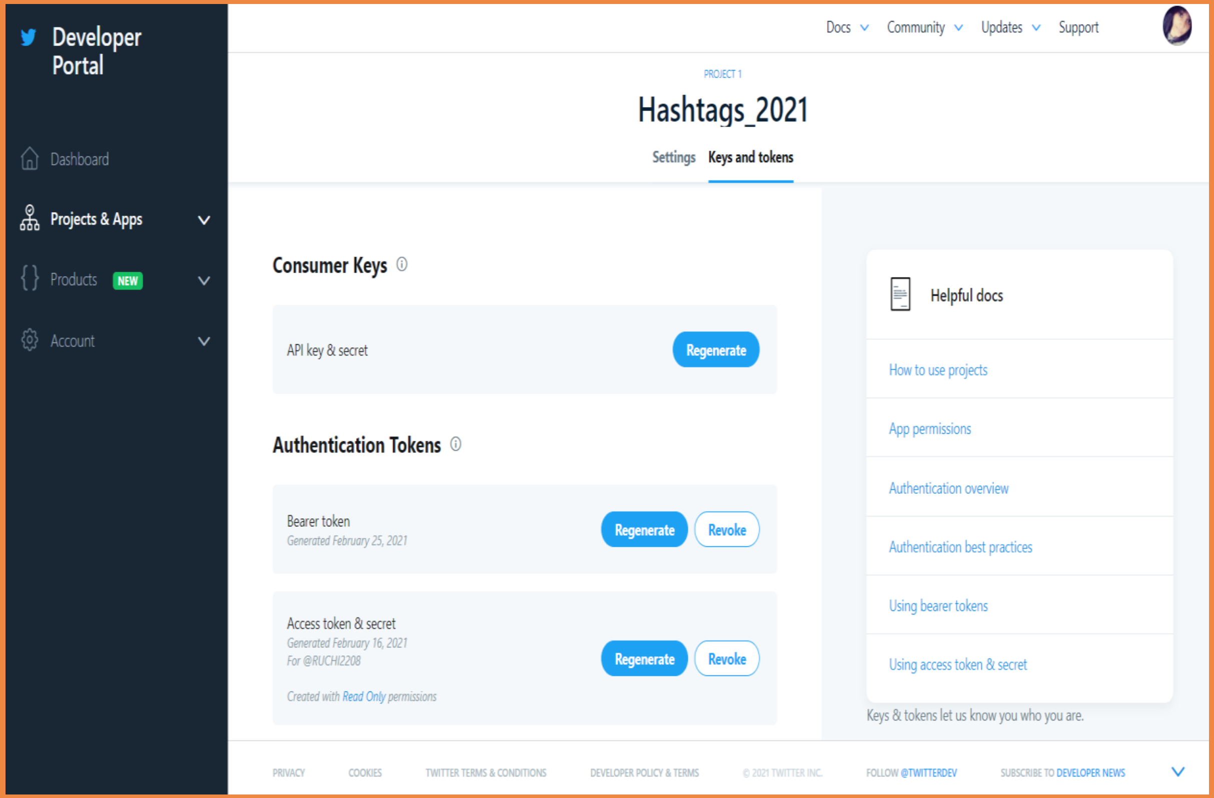


Negative Sentiment

HOW TO GET DATA FROM TWITTER

Create a Developers Account in Twitter API.

Collect the Access Token keys, Consumer Keys to get the live tweets from the API



PROJECT 1

Hashtags_2021

Consumer Keys

API key & secret

Authentication Tokens

Bearer token
Generated February 25, 2021

Access token & secret
Generated February 16, 2021
For @RUCHI2208

Created with Read Only permissions

Helpful docs

How to use projects

App permissions

Authentication overview

Authentication best practices

Using bearer tokens

Using access token & secret

Keys & tokens let us know you who you are.

Obtaining data from Twitter API

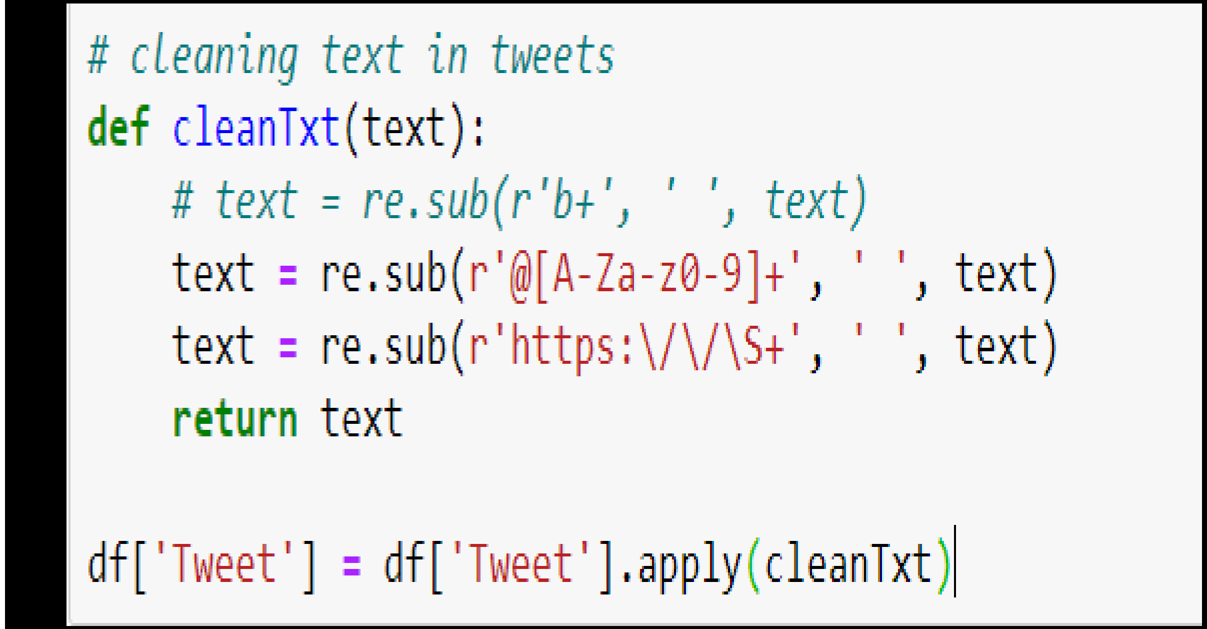
- As discussed earlier data in API follows REST framework. This data can be accessed in different file formats say CSV or JSON.
- We chose to save data in comma separate value as we intent to get only the tweet content.
- ‘Tweet’ is the content which is created by a twitter user to share the news, information or their thought process.
- Tweet is represented in 3 forms – original tweet, a retweet and quote tweet.
- For our project we chose to specifically exclude Retweets because we don’t have to quantify our data model and we need only the original data and then also collect those data in a certain limit.

FEATURE EXTRACTION

- Words and their frequencies
- Parts of speech tags
- Opinion words and phrases
- Position of terms
- Negation
- POS tag polarity score(Noun, Preposition, Adjectives)
- Special characters
- Count of repetition words
- Count of Non-English words
- Count of Acronyms

HOW DID WE CLEAN DATA?

- Removing unwanted patterns - Removing unwanted text patterns from the tweets.
- Removing Twitter Handles (@user).
- Removing Punctuations, Numbers, and Special Characters.
- Removing Short Words.
- Text Normalization.

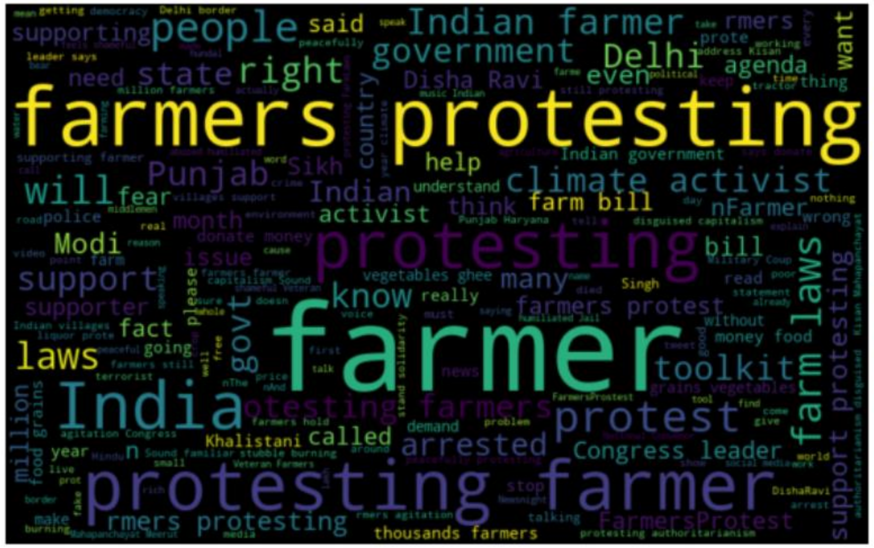


```
# cleaning text in tweets
def cleanTxt(text):
    # text = re.sub(r'b+', ' ', text)
    text = re.sub(r'@[A-Za-z0-9]+', ' ', text)
    text = re.sub(r'https:\/\/\S+', ' ', text)
    return text

df['Tweet'] = df['Tweet'].apply(cleanTxt)
```

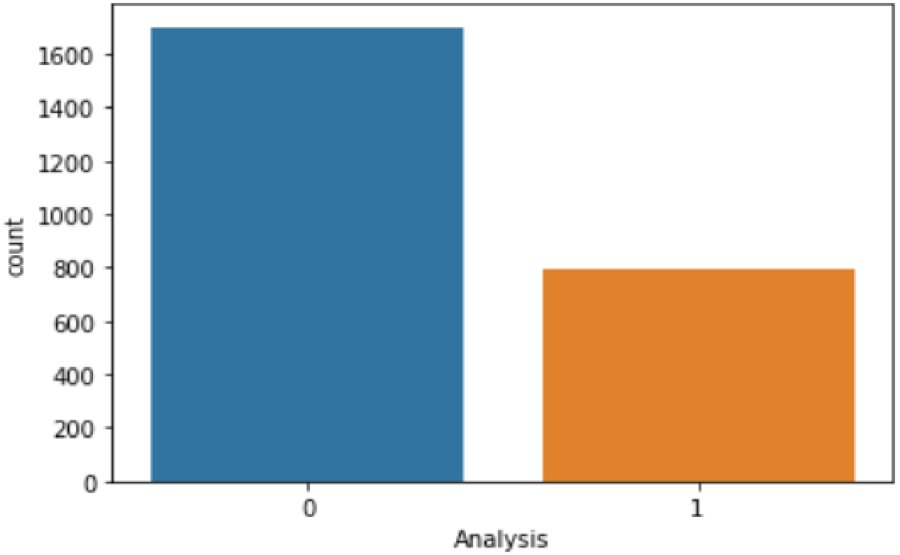
NORMALIZING THE DATA

- **Text normalization** is the process of transforming a **text** into a canonical (standard) form.
- For example, the word “gooood” and “gud” can be transformed to “good”, its canonical form.
- Another example is mapping of near identical words such as “stopwords”, “stop-words” and “stop words” to just “stopwords”.

Visualizing



#FarmersProtest tweet data, **Farmer** is the most repeated word in **wordcloud.**

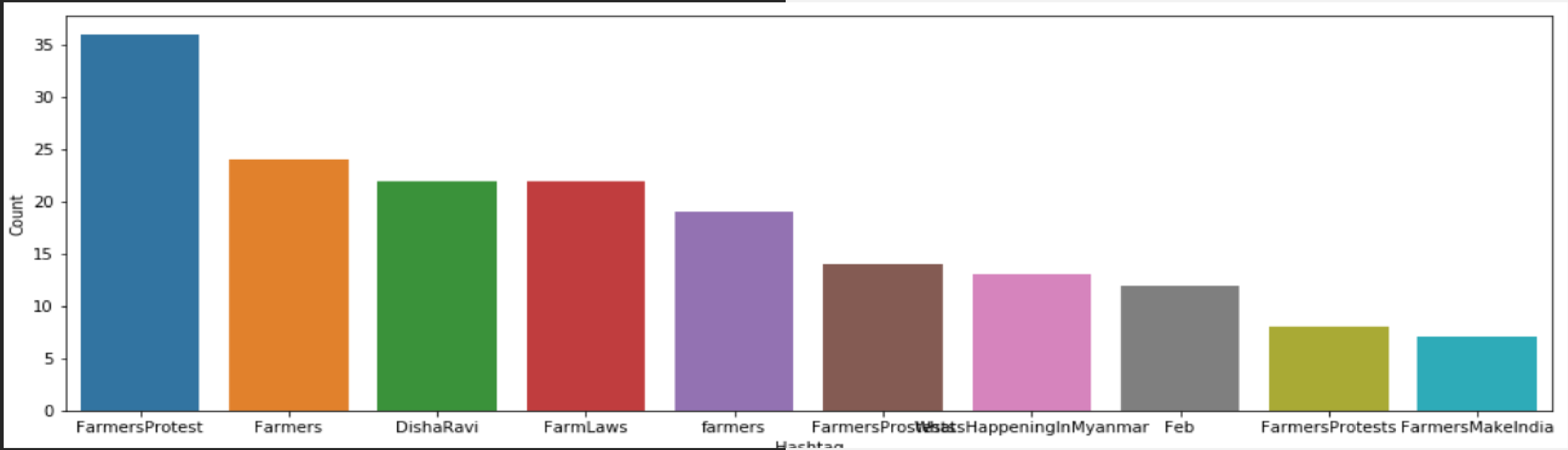


0- Negative Tweets

1- Positive Tweets

CONTINUE..

- From below representation we can say which hashtags has mostly tweeted.
- #FarmersProtest is highest as seen followed by #Farmers

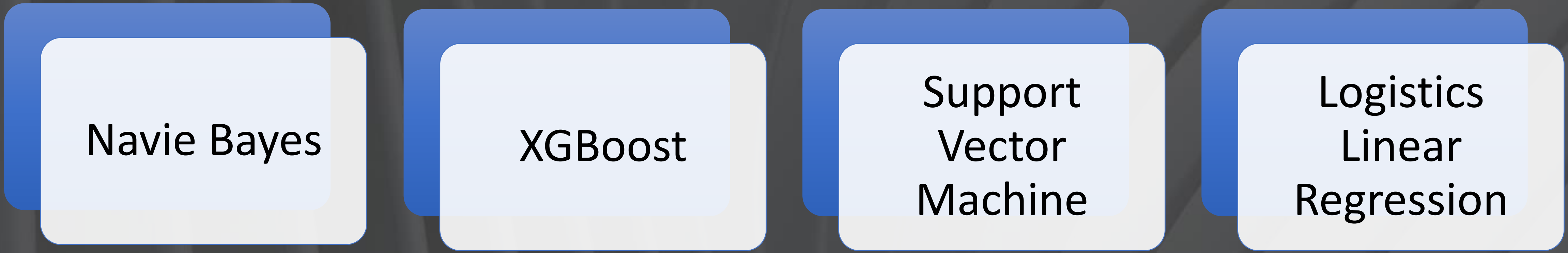


MACHINE LEARNING

Machine Learning provides smart alternatives to analyzing vast volumes of data.

By developing fast and efficient algorithms and data-driven models for real-time processing of data, **Machine Learning** can produce accurate results and analysis.

MACHINE LEARNING MODELS



Navie Bayes

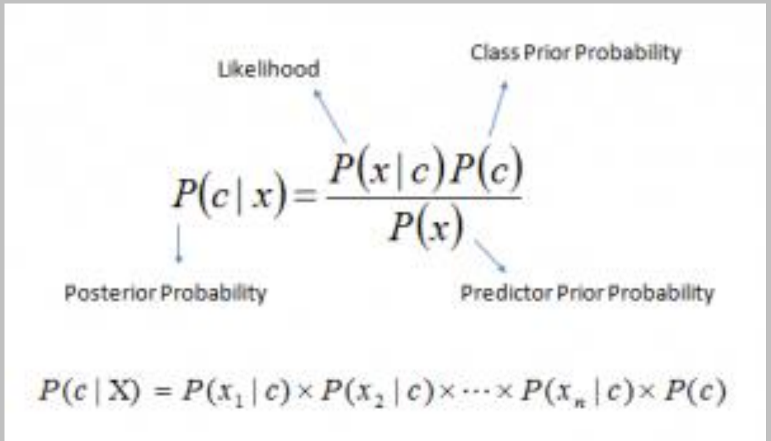
XGBoost

Support Vector Machine

Logistics Linear Regression

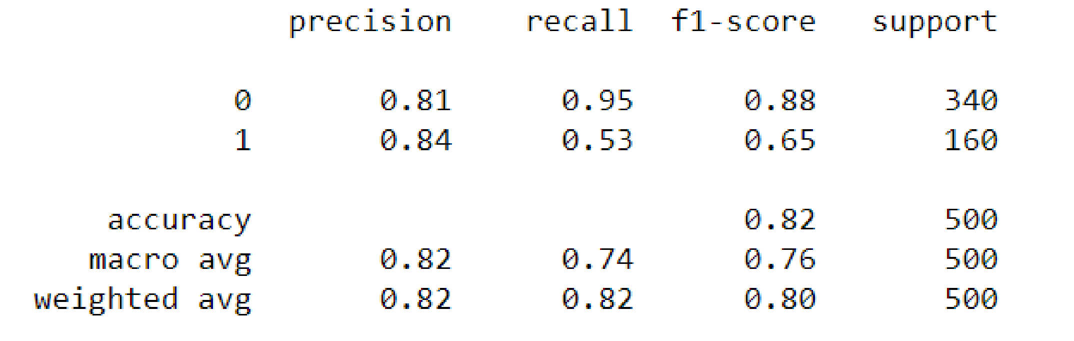
NAIVE BAYES

- **Naïve Bayes Classifier** is one of the simple and most effective **Classification** algorithms which helps in building the fast **machine learning** models that can make quick predictions.
- It is a probabilistic **classifier**, which means it predicts based on the probability of an object.



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

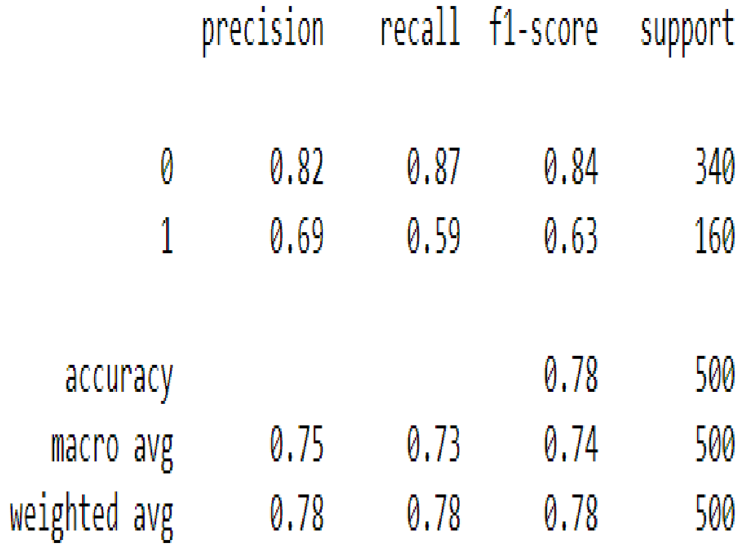


```
              precision    recall  f1-score   support

           0       0.81      0.95      0.88       340
           1       0.84      0.53      0.65       160

    accuracy                           0.82       500
   macro avg       0.82      0.74      0.76       500
weighted avg       0.82      0.82      0.80       500
```

XGBOOST

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

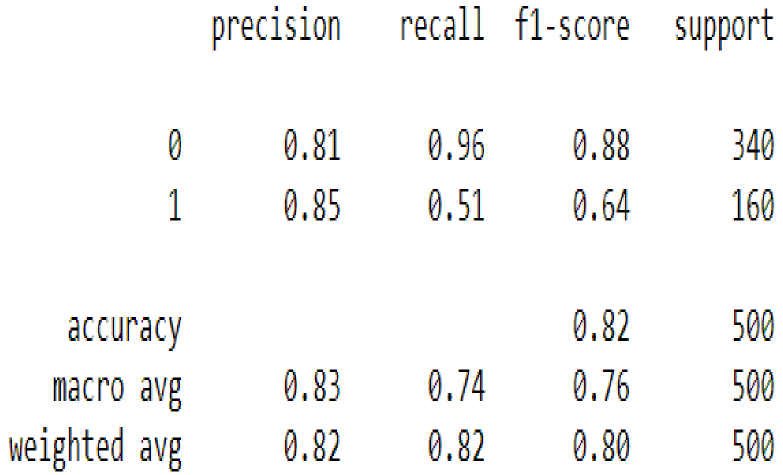
In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.



```
              precision    recall  f1-score   support

           0       0.82      0.87      0.84       340
           1       0.69      0.59      0.63       160

    accuracy                           0.78       500
   macro avg       0.75      0.73      0.74       500
weighted avg       0.78      0.78      0.78       500
```

Support Vector Machine

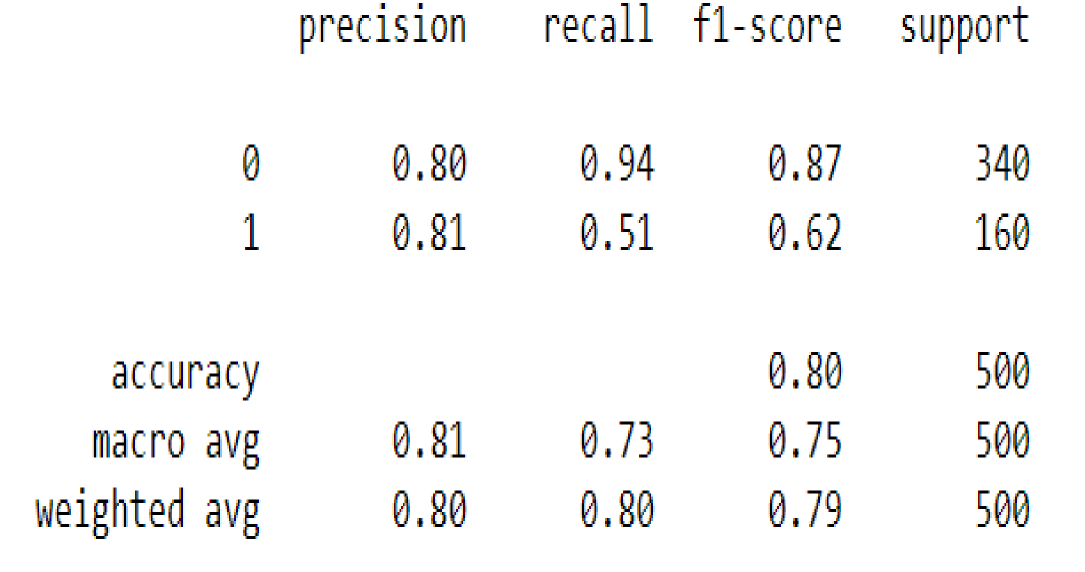
- “**Support Vector Machine**” (**SVM**) is a supervised **machine learning** algorithm which can be used for both classification or regression challenges.
- However, it is mostly used in classification problems. Support Vectors are simply the co-ordinates of individual observation.



```
              precision    recall  f1-score   support

           0       0.81      0.96      0.88       340
           1       0.85      0.51      0.64       160

    accuracy                           0.82       500
   macro avg       0.83      0.74      0.76       500
weighted avg       0.82      0.82      0.80       500
```

LOGISTICS LINEAR REGRESSION

- **Logistic regression** is considered a generalized **linear model** because **the** outcome always depends on **the** sum of **the** inputs and parameters.
- Or in other words, **the** output cannot depend on **the** product (or quotient, etc.) of its parameters!



```
              precision    recall  f1-score   support

           0       0.80      0.94      0.87       340
           1       0.81      0.51      0.62       160

    accuracy                           0.80       500
   macro avg       0.81      0.73      0.75       500
weighted avg       0.80      0.80      0.79       500
```

RESULTS

Models	XGBoost	Logistics Linear Regression	SVM	Navie Bayes
Accuracy	0.78	0.80	0.82	0.82

SVM **NAVIE BAYES**

Best accuracy

Challenges

- Tweet content shared by different people is vastly different.
- In opinion texts, lexical content alone can be misleading
- Intra-textual and sub-sentential reversals, negation, topic change are common in original tweets by people.
- Rhetorical devices/modes such as sarcasm, irony, implication, etc.
- Geographic location and language filters make huge impact on variation in tweet data as people from different areas tweet in different languages and twitter is multilingual in all parts of the world.
- Unstructured and non-grammatical as people don’t follow rules of grammar, lexical resource on a public platform like twitter.
- Out of Vocabulary words can result in many of impromptu tweets.
- Extensive usage of acronyms like asap, lol, afaik.

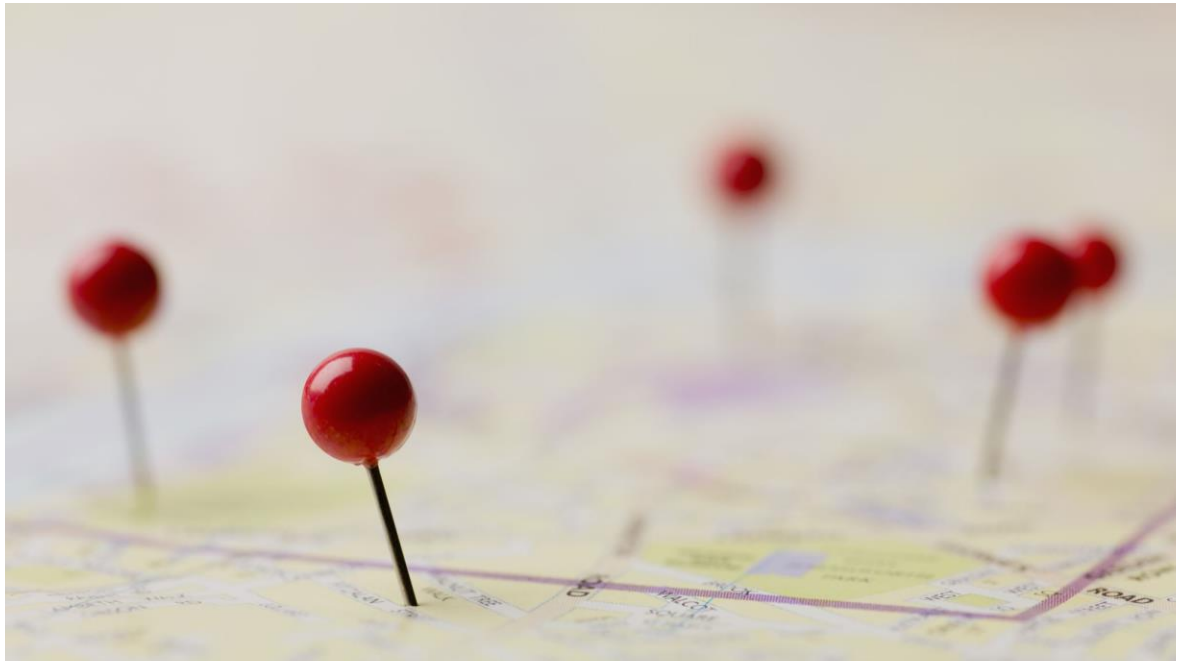
FURTHER PLANS

We will be applying the same model on other social media data gathered from Facebook and Instagram in order to compare how these social media channels fare in spreading hatred.

We will be making an attempt to create a web dashboard for this project to make it user friendly and interesting.

REFRENCES

- https://www.justintodata.com/twitter-sentiment-analysis-python/
- https://www.brandwatch.com/blog/understanding-sentiment-analysis/#:~:text=Sentiment%20analysis%20is%20extremely%20useful,public%20opinion%20behind%20certain%20topics.&text=Being%20able%20to%20quickly%20see,and%20plan%20for%20the%20future
- https://developer.twitter.com/en/portal/projects/1358920922967736321/apps/20069377/keys.



THANK YOU

Presented By:

- Aayush Seth
- Khushboo Patel
- Ruchita Tamboli