Project Requirement and Specification

On

Twitter Sentiment analysis

(CSE V Semester Mini project)

2022

\

Name: YASH BISHT

University Roll No: 2018872

Class Roll No: 62

Section: G

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GRAPHIC ERA HILL UNIVERSITY, DEHRADUN

CERTIFICATE

Certified that Mr. Yash Bisht (Roll No.-2018872) has developed mini project on “Twitter sentiment analysis” for the CSE V Semester Mini Project Lab (PCS-504) in Graphic Era Hill University, Dehradun. The project carried out by Students is their own work as best of my knowledge.

(Dr. Indrajeet)
Project Guide
Resource Person
(CSE Department)
GEHU Dehradun

ACKNOWLEDGMENT

I would like to express our gratitude to The Almighty God, the most Beneficent and the most Merciful, for completion of project.

I wish to thank our parents for their continuing support and encouragement. I also wish to thank them for providing us with the opportunity to reach this far in our studies.

I would like to thank particularly my Project Guide Dr. Indrajeet for his patience, support and encouragement throughout the completion of this project and having faith in us.

At last but not the least I greatly indebted to all other persons who directly or indirectly helped us during this work.

YASH BISHT

Roll No-2018872

CSE-G-V-Sem

Session: 2020-2024

GEHU, Dehradun

TABLE OF CONTENTS

ABSTRACT

This project tackles the issue of tweet sentiment analysis, which involves categorising tweets into those that indicate good, negative, or neutral mood. Twitter is a social networking Website/Application that enables users to post 140-character limit microblogs and status updates. It is a continuously growing service with Over 300 million people have signed up for the service, of which 185 million users are active and most of them log in everyday, resulting in over 550 million tweets. We wish to depict the mood of the public by studying the ideas expressed in the tweets due to the high usage. People's attitudes must be analysed for several purposes, such as determining the market worth of their area, predicting election results, and socioeconomic phenomena such as stock exchange. The purpose of this project is to develop a practical classifier capable of reliably and automatically classifying the sentiment of an unlabeled twitter stream.

PROJECT INTRODUCTION AND MOTIVATION

About Project

Sentimental analysis, often known as information extraction, is a submachine literacy job that seeks to ascertain the common opinion of a given content. We can value the personal info of a communication and try to categorise it according to its nature, such as positive, neutral, or negative, using computer literacy methods and natural language processing. It's a pretty valuable analysis since we may perhaps identify the total opinion about a selling product, or forecast stock requests for a certain firm, for example, if most people believe positively about it, stock requests may grow, and so on. The complexity of the language (objectivity/subjectivity, negation, lexicon, syntax) makes sentiment analysis far from being a solved problem, but it is one of the basis why it is so fascinating to work on. In this project, I decided to create a model based on probabilities to attempt to categorise messages from Twitter into "practical" or "impractical" emotion. By building a model based on probabilities. Twitter is a micro blogging website/application where people can share their thoughts quickly and spontaneously by sending a tweet limited by 140 characters. By including the @ target symbol and the hashtag, you may contact someone directly in a tweet or join a discussion. Twitter is a wonderful source of information to ascertain the current general opinion about anything due to its popularity. Since we believe Twitter to be a more accurate reflection of public opinion than traditional online articles and web blogs, I have opted to work with it. The reasoning is that, when compared to traditional blogging platforms, Twitter has a far bigger volume of relevant material. Furthermore, the response on Twitter is increasingly prevalent (since the users who tweet is considerably more than those people who write blogs on a daily basis). When forecasting macro-scale socioeconomic events such as a company's stock market rate, public sentiment analysis is useful & extremely important. This might be achieved by looking at the entire public perception of that company through time and applying economics techniques to identify the relationship between consumer opinion and a company's stock market valuation. Additionally, businesses can predict how well their product is performing in the market and which regions of the market are receiving positive and unfavourable response (Because Twitter allows us to acquire streams of geotagged tweets for specified locations. If companies can access this information, they will be able to examine the causes of geographically different reactions and sell their goods more effectively by finding out applicable solutions, such as the development of acceptable market segments. Another growing use for sentiment analysis is the ability to predict the outcomes of popular political elections and surveys. According to one of these studies, which was carried digitally in Germany to forecast the results of federal elections, Twitter is an excellent indicator of offline mood.

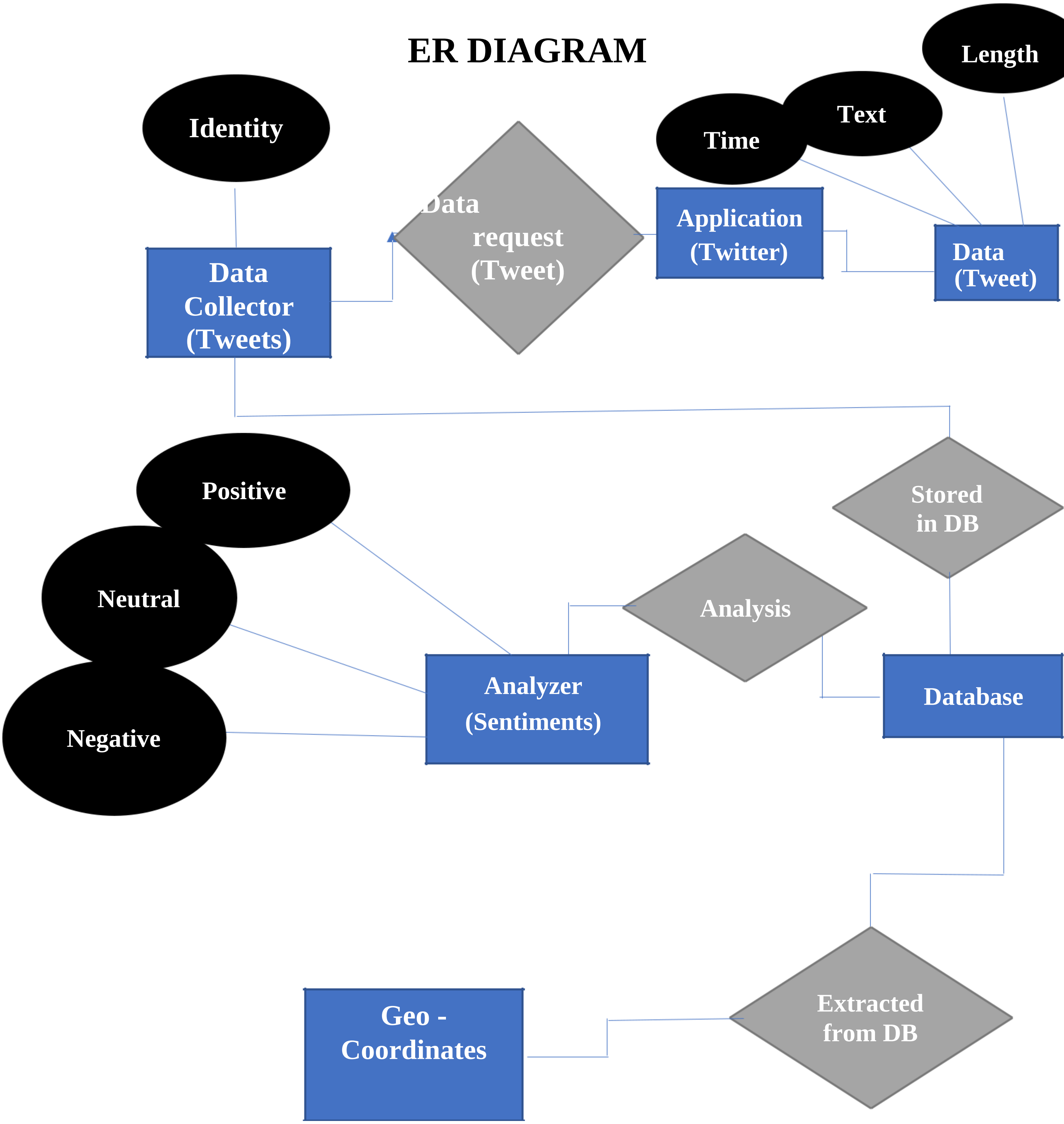
METHODOLOGY

To collect data in the form of raw tweets, the Python library "tweepy," which provides a package for a simple Twitter streaming API, is utilised. This API provides 2 access methods for tweets: Sample Stream and Filter Stream. Sample Stream just takes a brief, random sample of all the tweets that are streaming in real time. Filter Stream sends tweets that meet a set of criteria. It can sort the input tweets using three filters: Twitter user identification by name; a keyword to track or search for in tweets; tweets from a certain individual certain location(s) (only for geo-tagged tweets). Any one of these filtering criteria can be specified individually by a programmer, or they can be combined in various ways. However, we do not face this limitation for our purposes, and So, we'll continue to use the Sample Stream mode. Instead of gathering all of the data at once, we divided it up into smaller chunks at various times because we wanted to improve the breadth of our data. If we had chosen the second option, the flexibility of the tweets could have been harmed because a substantial number of people would have been highlighting a certain hot topic and hence sharing a roughly similar attitude or mood. When we were looking over our sample of collected tweets, we noticed this tendency. For instance, the sample collected around Deepawali, and Holi had a higher percentage of tweets referencing these pleasant occasions and were, as a result, of a typically positive nature. Thus, sampling our data in chunks at varying moments in time would attempt to mitigate this issue. It is implemented as a Python "dictionary" data type., which contains a variety of key-value pairs. The following are some key-value pairs:

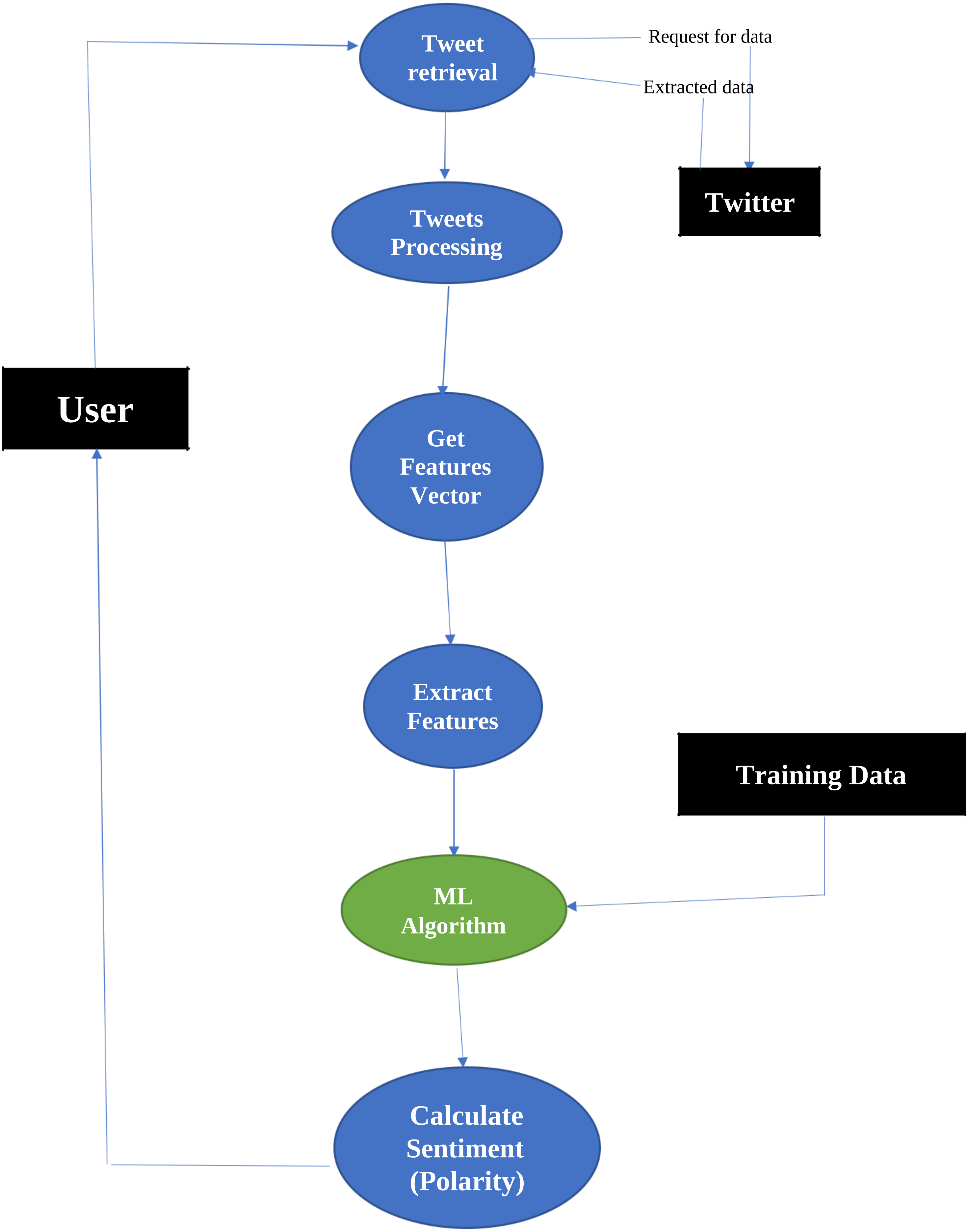
- Whether a tweet has been favourited
- User ID
- Screen name of the user
- Original Text of the tweet
- Presence of hashtags
- Whether it is a retweet
- Language of account registration in twitter
- Geo-tag location of the tweet
- Date and time when the tweet was created

Because this is a lot of information, we only filter out what we need and discard the rest. We loop over all of the tweets in our sample for our unique application. and store the true text content of the retrieved tweets in a separate file if the user's account's language is set to English. The dictionary key "txt" refers to the original text content of the tweet, while "lang" refers to the language of the account.

ER DIAGRAM


Identity
Data Collector (Tweets)
Data request (Tweet)
Time
Text
Length
Application (Twitter)
Data (Tweet)
Positive
Neutral
Negative
Analyzer (Sentiments)
Analysis
Stored in DB
Database
Extracted from DB
Geo - Coordinates

Data Flow Diagram



Data Flow Explanation:

The steps involved in the analysis of tweets includes how the data is retrieved from twitter this can be understood in the following steps:

- **Tweet retrieval :** This process is aided by the tweepy python library , which takes the input of requested data to twitter application where tweets are actually posted and extract data/tweets from it.

- **Tweet Processing :** The tweets retrieved are then processed on the basis of their key value pairs. Example of these are: The originality of a tweet, that is it should not be a retweet as it can cause redundancy, User id etc.

- **Get Features Vector :** The processed tweets are now filtered on the basis of the query string given by the user. The filtered tweets only contain tweets of user's interest. For example user searched for a company's name all tweets displayed will have company's name somewhere mentioned in the tweet.

- **Extract Features :** Since some tweets are very huge and for calculating polarity of them a small part is required to be sent to the machine learning algorithm. The important features are extracted and sent to the ML algorithm.

- **ML Algorithm :** The machine learning algorithm along with trained dataset checks the extracted features of tweets and returns 3 values: -1, 0, 1.

- **Calculate Sentiment (Polarity) :** on the basis of returned values from the ML algorithm the polarity is given as:
  -1 = negative
  0 = neutral
  1 = positive
  These results are displayed to user in the front-end.

Machine Learning Model:

Naïve bayes classifier algorithm is used in this project.

This machine learning algorithm is based on bayes theorem. It is used in text classification that is differentiating in different kind of sentences. It is based on the probability of one event on account of the other event which can be understood by bayes theorem. It is a simple and efficient machine learning algorithm used to make quick predictions.

Other use case of this algorithms are spam filtration & classifying articles.

There are three types of Naïve bayes algorithm:

- **Gaussian**: The predictor in this model takes values continuously instead of single discrete value, then using Gaussian distribution these values are sampled.

- **Multinomial**: The data when available in multinomial distribution this is used. The primary use case is classification of different documents based on the different categories such as Sports, Politics, Education etc. Frequency of words is used for the predictors.

- **Bernoulli**: Its function is similar to multinomial, but the difference is that the predictor variable are Boolean which are independent of a particular word present in document. This model is known for its document classification capabilities.

Advantages of Naïve Bayes:

- It is a effective and efficient ML algorithm to predict a set of data.
- Used in binary(0/1) and also in multi class classification.
- It performs better in Multi-Class predictions when compared with other algorithms of similar interest.
- It is widely used and most popular for text classification analyzer.

Disadvantages of Naïve Bayes:

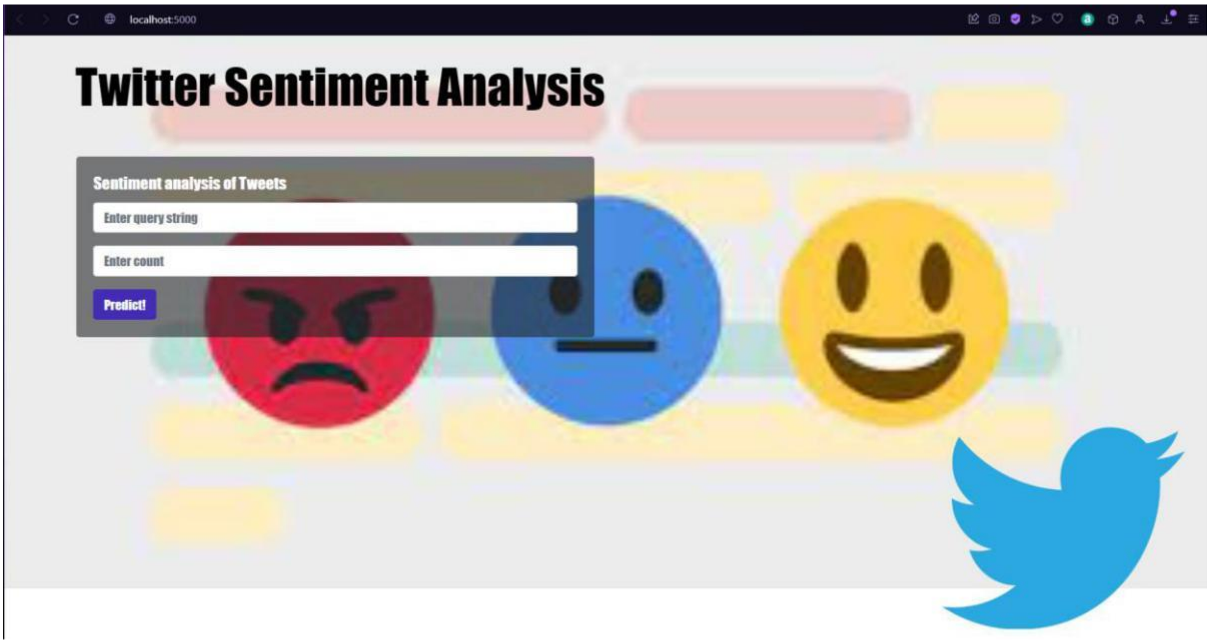
- Since most features are independent this algorithm does not learn about the relationship between features of data.

Uses of Naïve Bayes:

- Calculating credit score.
- Real time predictions of text classifications for example sentiment analysis and spam filtering.

Frontend Screenshots:

1. First Page



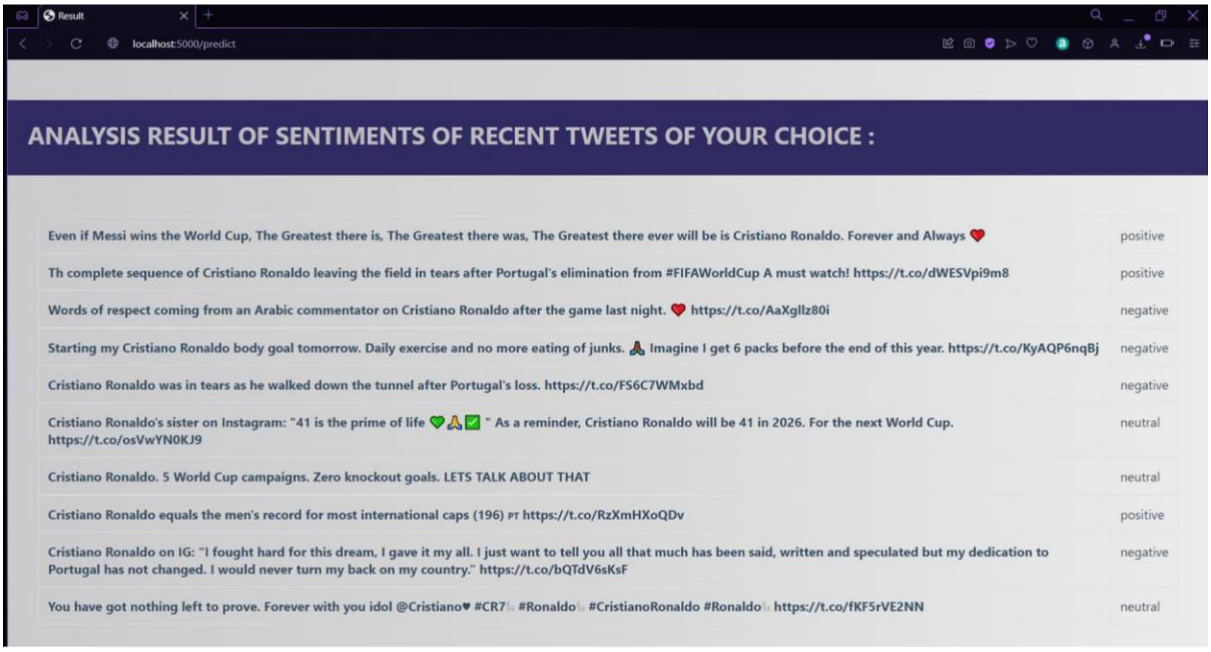
Twitter Sentiment Analysis

Sentiment analysis of Tweets

Enter query string

Enter count

2. Results page



ANALYSIS RESULT OF SENTIMENTS OF RECENT TWEETS OF YOUR CHOICE :

Even if Messi wins the World Cup, The Greatest there is, The Greatest there was, The Greatest there ever will be is Cristiano Ronaldo. Forever and Always ❤️	positive
Th complete sequence of Cristiano Ronaldo leaving the field in tears after Portugal's elimination from #FIFAWorldCup A must watch! https://t.co/dWESVpi9m8	positive
Words of respect coming from an Arabic commentator on Cristiano Ronaldo after the game last night. ❤️ https://t.co/AaXgllz80i	negative
Starting my Cristiano Ronaldo body goal tomorrow. Daily exercise and no more eating of junks. 🙏🏾 Imagine I get 6 packs before the end of this year. https://t.co/KyAQP6nqBj	negative
Cristiano Ronaldo was in tears as he walked down the tunnel after Portugal's loss. https://t.co/FS6C7WMxbd	negative
Cristiano Ronaldo's sister on Instagram: "41 is the prime of life 💚🙏✅ " As a reminder, Cristiano Ronaldo will be 41 in 2026. For the next World Cup. https://t.co/osVwYN0KJ9	neutral
Cristiano Ronaldo. 5 World Cup campaigns. Zero knockout goals. LETS TALK ABOUT THAT	neutral
Cristiano Ronaldo equals the men's record for most international caps (196) 🇵🇹 https://t.co/RzXmHXoQDv	positive
Cristiano Ronaldo on IG: "I fought hard for this dream, I gave it my all. I just want to tell you all that much has been said, written and speculated but my dedication to Portugal has not changed. I would never turn my back on my country." https://t.co/bQTdV6sKsF	negative
You have got nothing left to prove. Forever with you idol @Cristiano♥ #CR7 #Ronaldo #CristianoRonaldo #Ronaldo https://t.co/fKF5rVE2NN	neutral

Conclusion

Twitter sentimental analysis is one of the most popular topics in machine learning these days. Considering the complexity of English language and much more so when other languages such as Chinese are considered, we are still a long way from reliably detecting the sentiments of a corpus of texts.

In this project, we attempted to demonstrate the fundamental method of identifying tweets as positive, negative or neutral using Naive Bayes as a baseline. While working on this project we learned how language models that are related to the Naive Bayes can produce better results. We may enhance our classifier further by extracting more features from the tweets, experimenting with different types of features, altering the parameters of the Naive Bayes classifier, or attempting another classifier entirely.

We observed that the number of neutral sentences is substantially higher than expected, indicating that Twitter sentiment analysis needs to be improved. We have only worked with elemental unigram models till now in this project. This model can be further improved by including additional information like a word's closeness to a negation word and much more.