

# Schmaltz Surveyor

Sentimental Analysis of Twitter

Minor Project – IS6C06

Under the Guidance of Mrs. Nandini BM

By

Nithyashree Arunachalam

Pradyoth P

Tejasvini SJ

Shashank BU



# Introduction

Social media has gained immense popularity and has become a major global platform to stay connected as well as express opinions.

A huge amount of content is created on various topics and comments are posted on these platforms daily.

The feedback received on a certain piece of content can be either negative or positive.

Other than this, receiving negative feedback, on various occasions might affect the mental health of the content creators and, in some cases, might also lead to cyberbullying.

Social media is a necessity in today's time to stay connected, informed, and relevant and therefore such issues must be tackled.

Schmaltz refers to excessive sentimentality. Therefore this project has been named Schmaltz Surveyor, which aims to analyze the sentiments expressed by Twitter users.

- Sentimental Analysis reads people's sentiments or emotions towards particular things or topics.
- Sentiment analysis is a machine learning tool that will help analyze and categorize the texts as positive or negative.

# Literature Survey

- Xing Fang, and Justin Zhan[1], worked on Sentiment Analysis on Amazon Online Products Review by collecting data from amazon.com. They figured out the issues of categorical sentiment polarity using Machine Learning Algorithms. They used many libraries that hold Naive Bayesian, SVM, and Random Forest.
- Faizan[2] built a model for the analysis of feeling using the KNN algorithm with unigram, bigram, and n-gram features. They then performed training and testing of their model on the US Airlines data set, for which they attained an accuracy of 65.33 %.
- Chirag Kariya and Priti Khodke's[3] paper explains various steps involved in the analysis of Twitter sentiments along with the various tools that are used to perform Twitter sentiment analysis. Amongst the various algorithms available, the KNN algorithm is used to increase the efficiency of sentiment analysis whereas Naive Bayes for simple and efficient sentiment analysis by classifying the tweets as either positive, negative, or zero.
- Akshay Amolik et al.[4] proposed sentiment analysis, and they accurately classified tweets by using Feature Vector and classifiers like Naïve-Bayesian and SVM. Exception of lower recall and accuracy, Naïve Bayesian has better precision as a comparison to SVM. SVM gives better results when it comes to accuracy. With the increase of training data, the accuracy of classification will also increase.

# Proposed System

The objective of this whole project is to classify the social media content into positive or negative using machine learning algorithms.

It will help us analyze and see the extent of the positive or negative extent of these comments and content on social media.

A dataset taken consists of comments and texts extracted from various social media platforms.

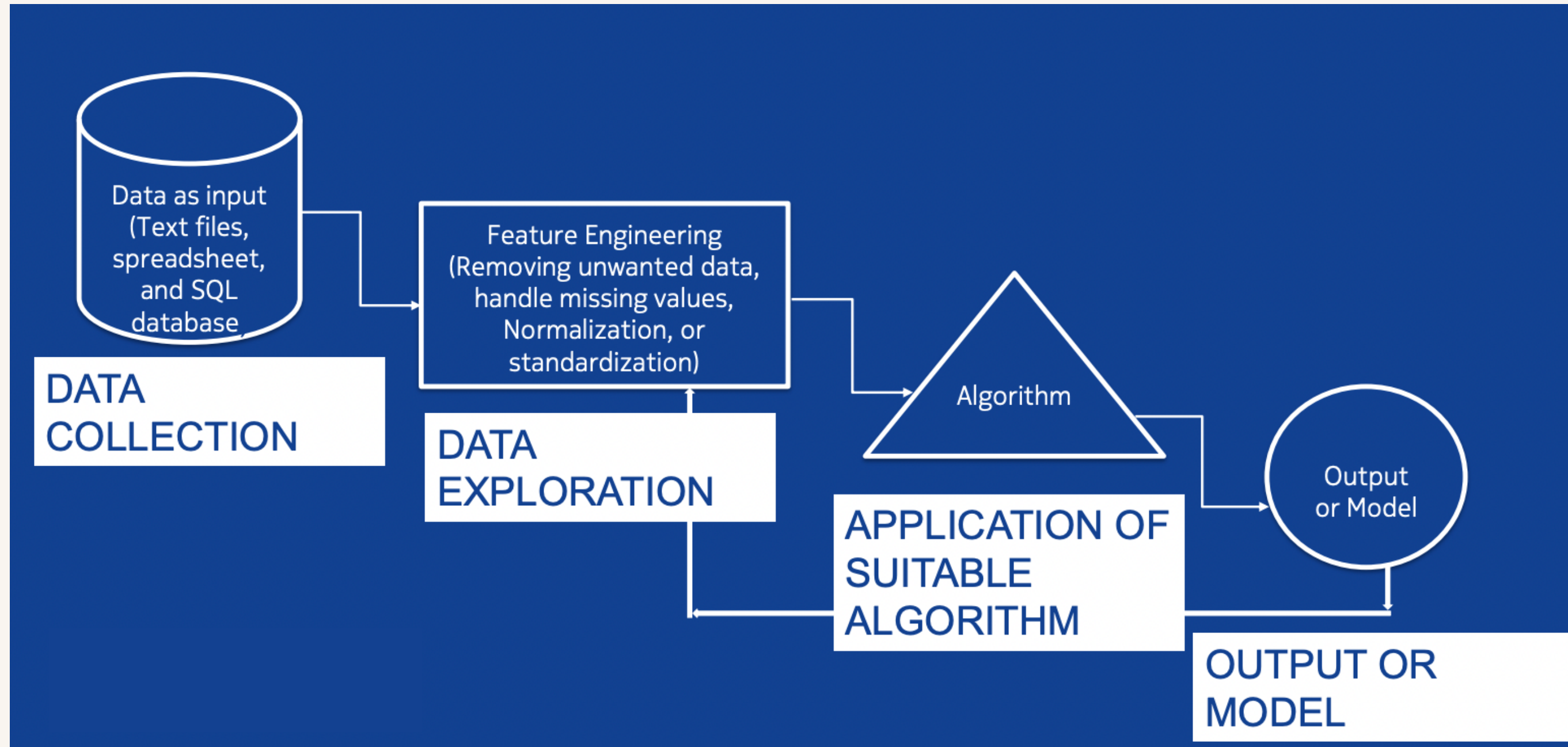
The dataset is then converted into the numerical form using vectorization methods in NLP. These numbers are used to train ML models to make predictions.

The ML models or classifiers as the name suggests classifying the text into positive and negative. These classifiers are nothing but machine learning algorithms that automatically order or categorize data.

**Scikit learn** library which is a python library will be used extensively throughout the whole project. Scikit-learn is one of the most useful libraries for machine learning in Python. It contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction

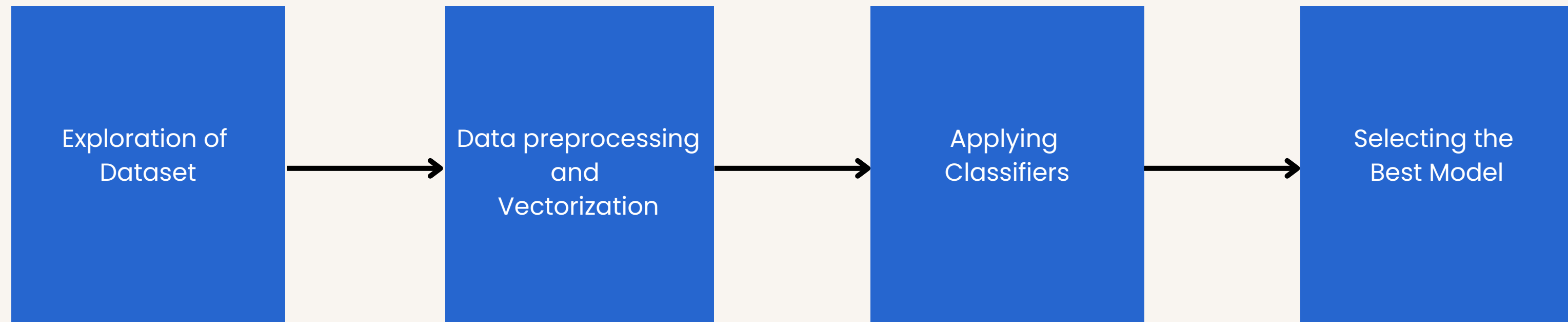
# Flow of the Project

Based on the typical practises, the project will be covered in 4 phases.



# Project Design

The project is partitioned into 4 Phases.





# Exploration of Dataset

- Creating a database that has a list tweets
- This database can be divided into training and testing data.

# Data preprocessing and Vectorization

**Step 1: Checking for missing values.**

**Step 2: Text Normalization.**

Normalize the text data as texts from such online platforms usually contain inconsistent language and the use of special characters in place of letters. To tackle such inconsistencies in data, Regex.

**Step 3: Lemmatization**

Lemmatization is the process of grouping together the different forms of a word so they can be analyzed as a single item. For example, we do not want the Machine Learning algorithm to treat eating, eats, and eat as three separate words because they convey the same message. Lemmatization helps reduce the words to their root form.



# Data preprocessing and Vectorization

## Step 4: Removal of stop words

Stop words are a set of commonly used words in the language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.

## Step 5: Converting to Numerical Form

The dataset needs to be converted into the numeric form so that it can be put through classifiers.

- TF IDF: TF-IDF means Term Frequency – Inverse Document Frequency. This is a statistic that is based on the frequency of a word but it also provides a numerical representation of how important a word is for statistical analysis.

# Term Frequency– Inverse Document Frequency

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document.

The document frequency of word  $i$  represents the number of documents in the corpus with word  $i$  in them. Let us represent document frequency for word  $i$  by  $df_{ij}$ . With  $N$  as the number of documents in the corpus, the tf-idf weight for word  $i$  in document  $j$  is computed by the following formula:

$$w_{ij} = tf_{ij} \times \left(1 + \log \frac{1+N}{1+df_{ij}}\right)$$

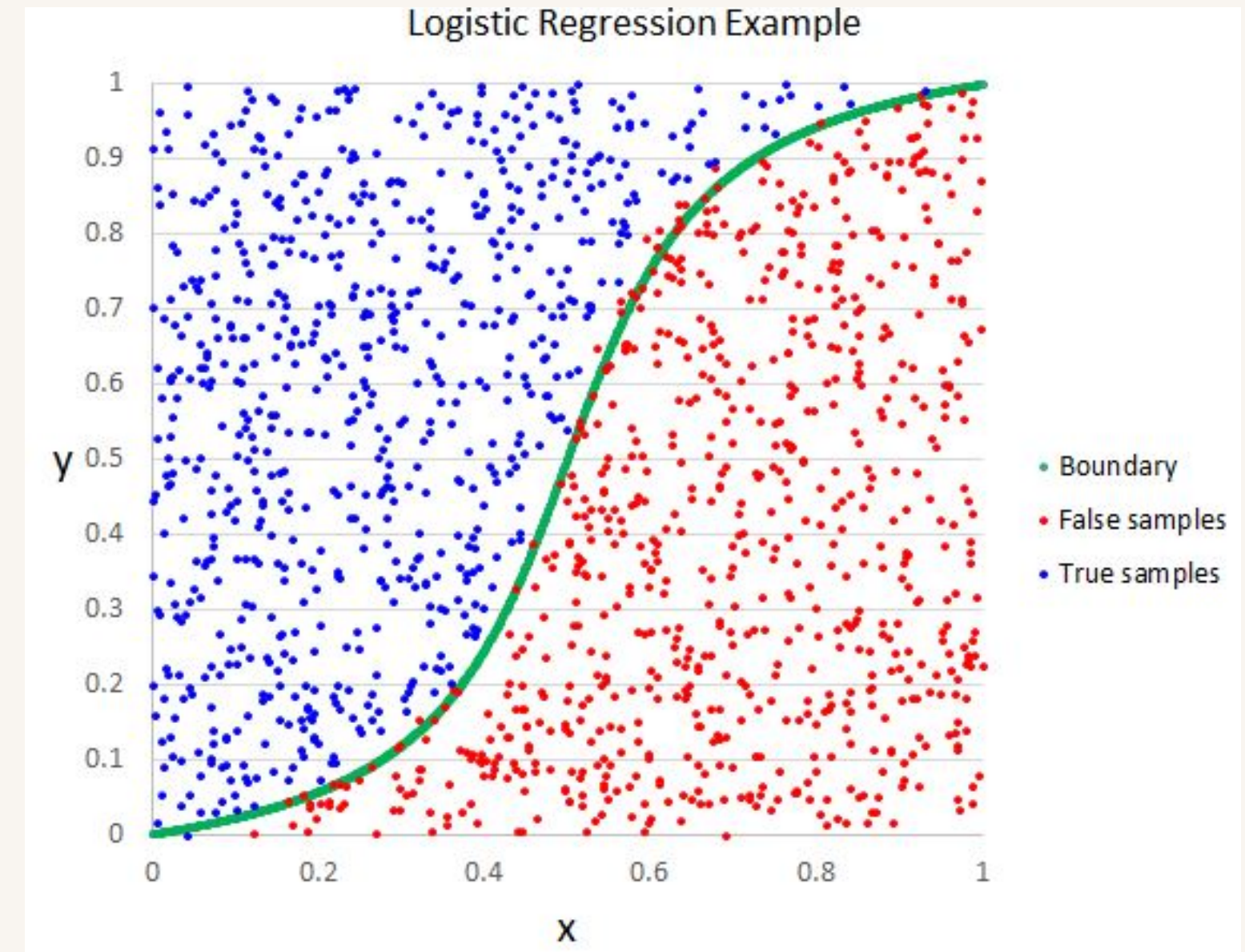
# Applying Classifiers

- **Logistic Regression**
- **Linear Support Vector Classifier or Support Vector Machine**
- **Random Forest**
- **k Nearest Neighbours**

# Logistic Regression

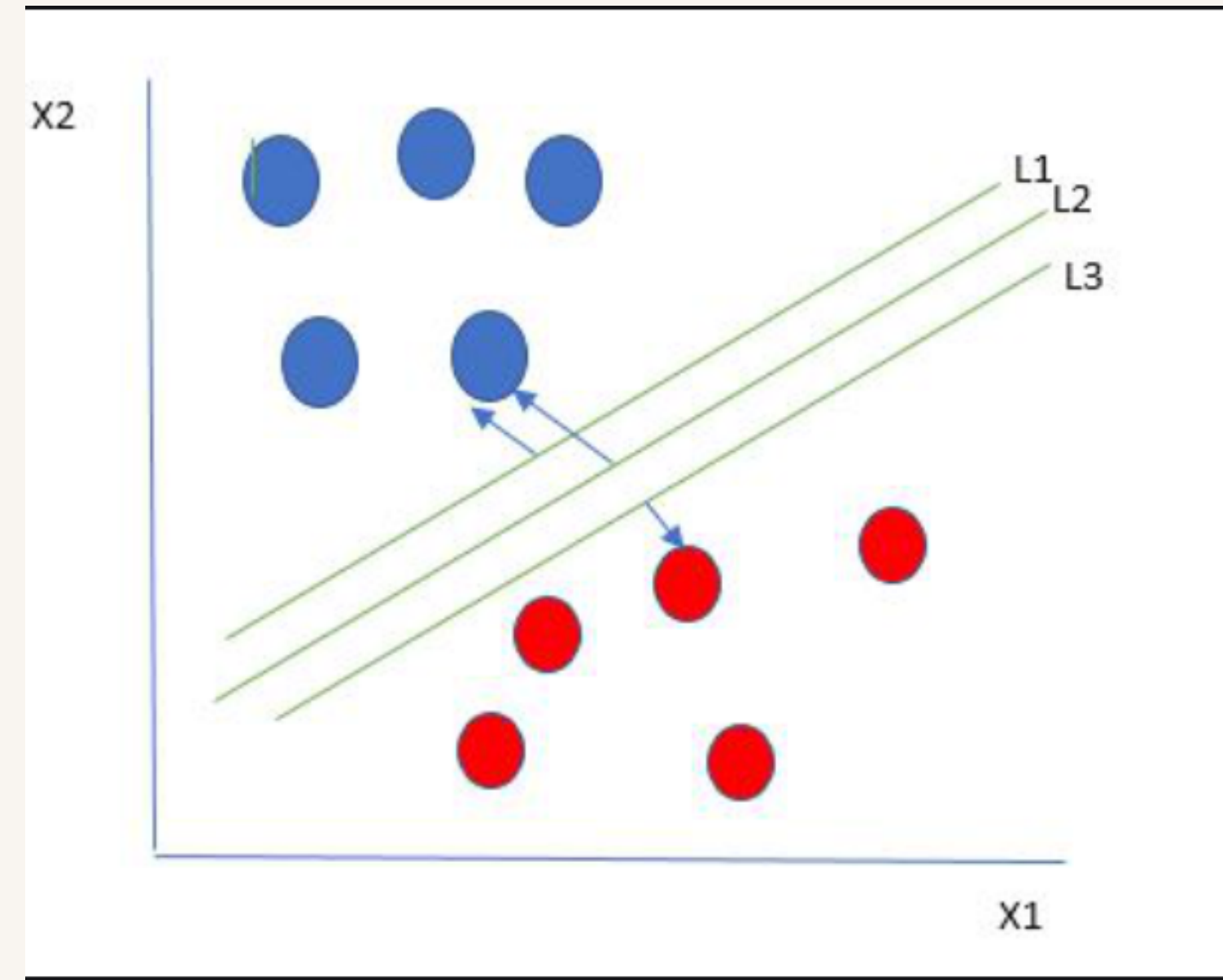
Logistic regression is a statistical analysis method to predict a binary outcome, such as 0 or 1, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.



# Support Vector Machine

- SVM performs classification by finding the hyper-plane that differentiate the classes we plotted in n-dimensional space.
- classifies positive and negative examples, here blue and red data points
- Largest margin is found in order to avoid overfitting. The optimal hyperplane is at the maximum distance from the positive and negative examples



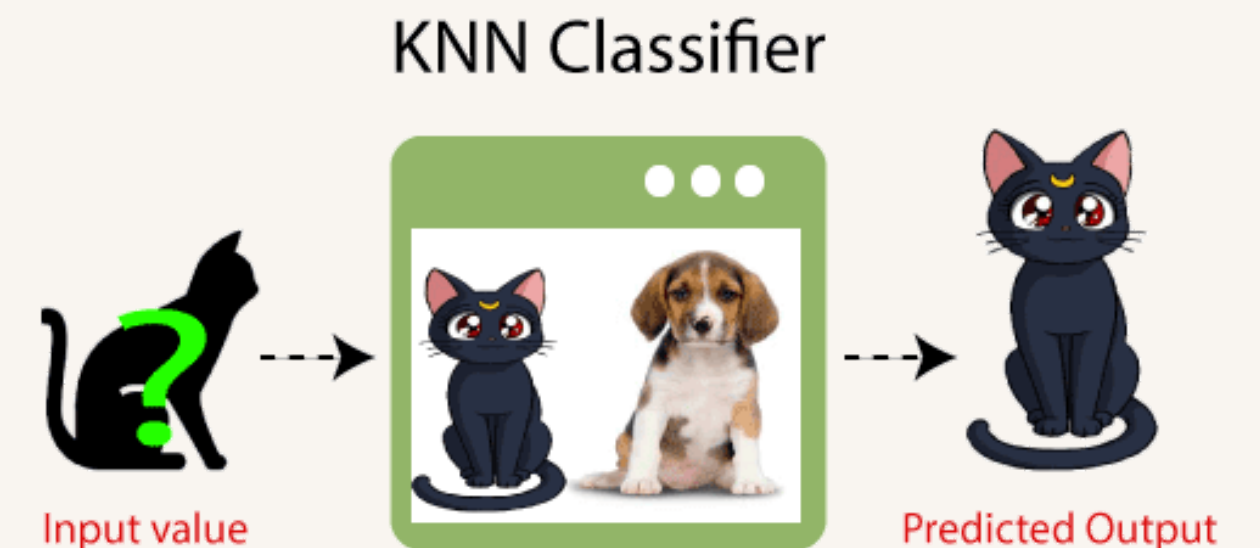


# k Nearest Neighbours

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumptions on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.



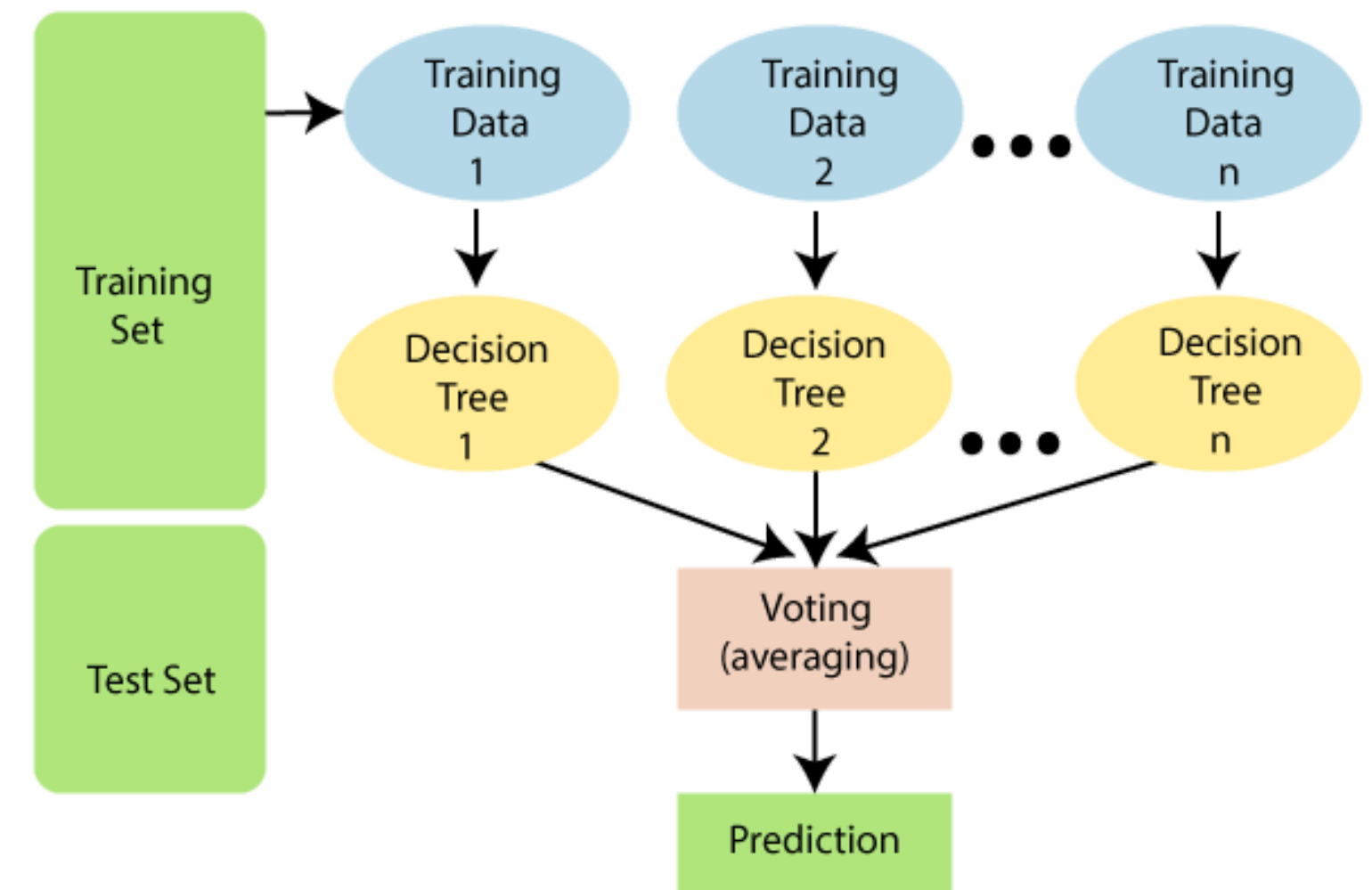


# Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Instead of relying on one decision tree, the random forest takes the prediction from each tree, and based on the majority votes of predictions, it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



# Understanding TP/TN and FP/FN

True => records that the model was able to identify its class

False => records that the model was not able to identify

## Confusion Matrix

A confusion matrix is a table that represents the summary of the prediction results on a classification problem.

# Understanding TP/TN and FP/FN

ACTUAL VALUES	PREDICTED VALUES	
	Positive	Negative
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

PREDICTED VALUES	ACTUAL VALUES	
	Positive	Negative
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)



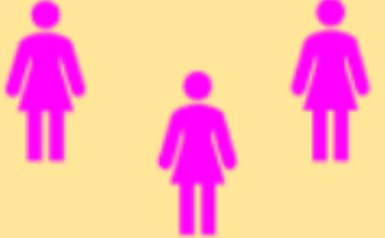

# Understanding TP/TN and FP/FN

## Actual Values:

8 women pregnant and 8 not pregnant women.

## Predicted Values:

1. True Positive: 6 pregnant woman
2. True Negative: 5 not pregnant woman
3. False Positive :3 not pregnant woman
4. False Negative: 2 pregnant woman

		PREDICTED VALUES	
		Pregnant	Not Pregnant
ACTUAL VALUES	Pregnant		
	Not Pregnant		

# Understanding TP/TN and FP/FN

1. **True Positive(TP)**: Values that are actually positive and predicted positive.
2. **False Positive(FP)**: Values that are actually negative but predicted to be positive.
3. **False Negative(FN)**: Values that are actually positive but predicted to be negative.
4. **True Negative (TN)**: Values that are actually negative and predicted to be negative.

Rate is a measuring factor in a confusion matrix.

- **True Positive Rate(TPR):  $\text{True Positive} / (\text{True Positive} + \text{False Negative})$**
- **True Negative Rate(FPR):  $\text{True Negative} / (\text{True Negative} + \text{False Positive})$**

# Understanding TP/TN and FP/FN

*Accuracy:* Accuracy is the ratio of the records that the model correctly classified over the total number of records.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$



# Selecting the Best Model

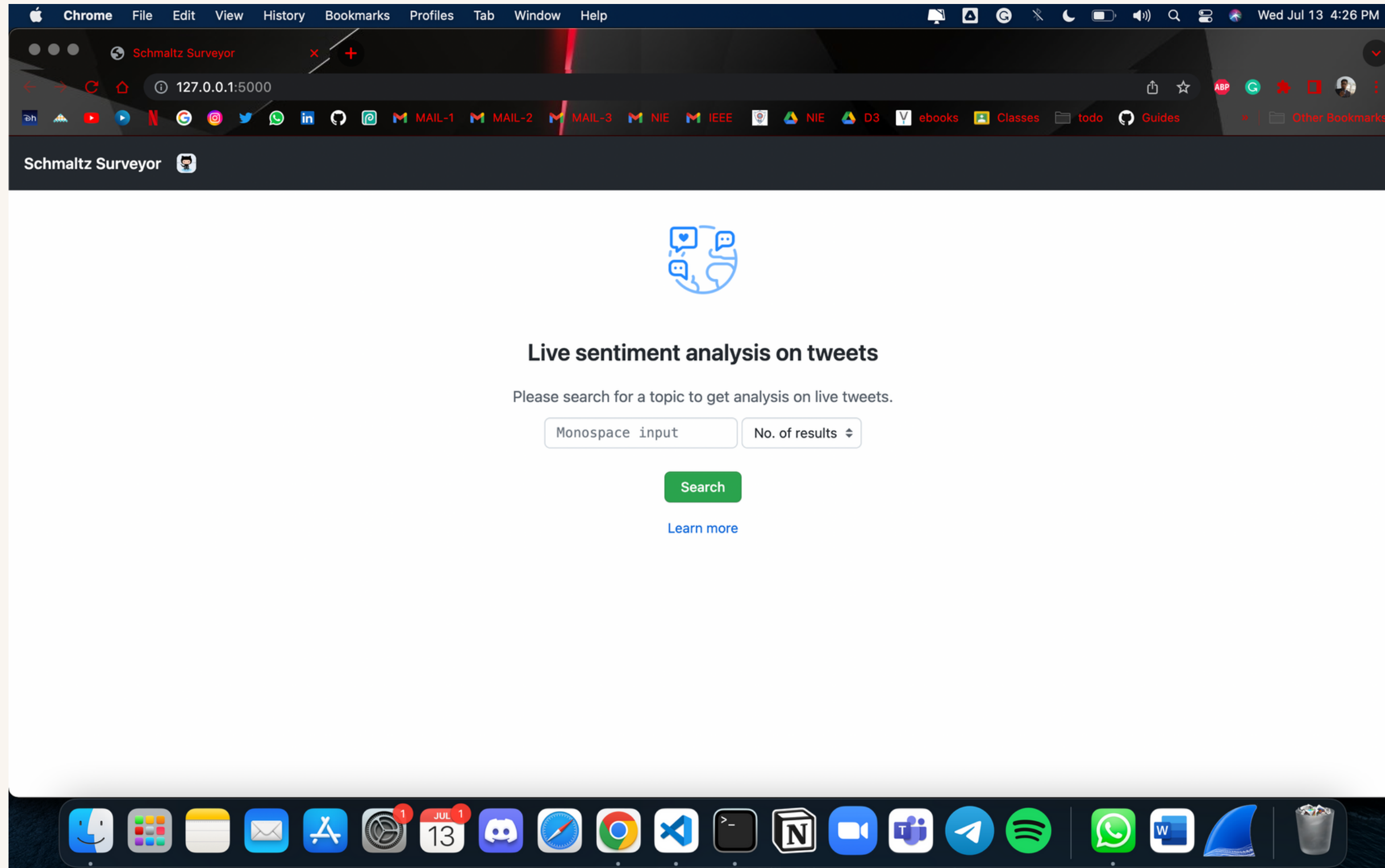
- Drawing comparison among the 4 to choose the model to be used

The Model used Classifier	Accuracy Score (in %)
Support Vector Machines	95.38
Logistic Regression	94.50
k Nearest Neighbors	93.69
<b>Random Forest</b>	<b>95.60</b>

# LIVE Sentimental Analysis

The whole objective of this project was to predict the sentiments of tweets. In the frontend, the user can pull the real time tweets by giving the keyword that the tweet must contain. These tweets are then used as the data to perform sentimental analysis and their sentiment is predicted.

# LIVE Sentimental Analysis




# LIVE Sentimental Analysis

We have the following functionality in the frontend:

- A single dialogue box in which the user can type keywords.
- A small drop down list containing the number(10,100,1000) of tweets that we can pull using the twitter API that have the particular keyword.
- The search button which then gets the number of tweets selected containing the keyword
- The user is redirected to a page which then lists the tweets along with the sentiment that it has predicted.

# LIVE Sentimental Analysis

Schmaltz Surveyor



Live sentiment analysis on india

The following are the tweets on the topic and there sentiments.

Monospace input

Search

✓ No. of results

10

100

1000

[Learn more](#)

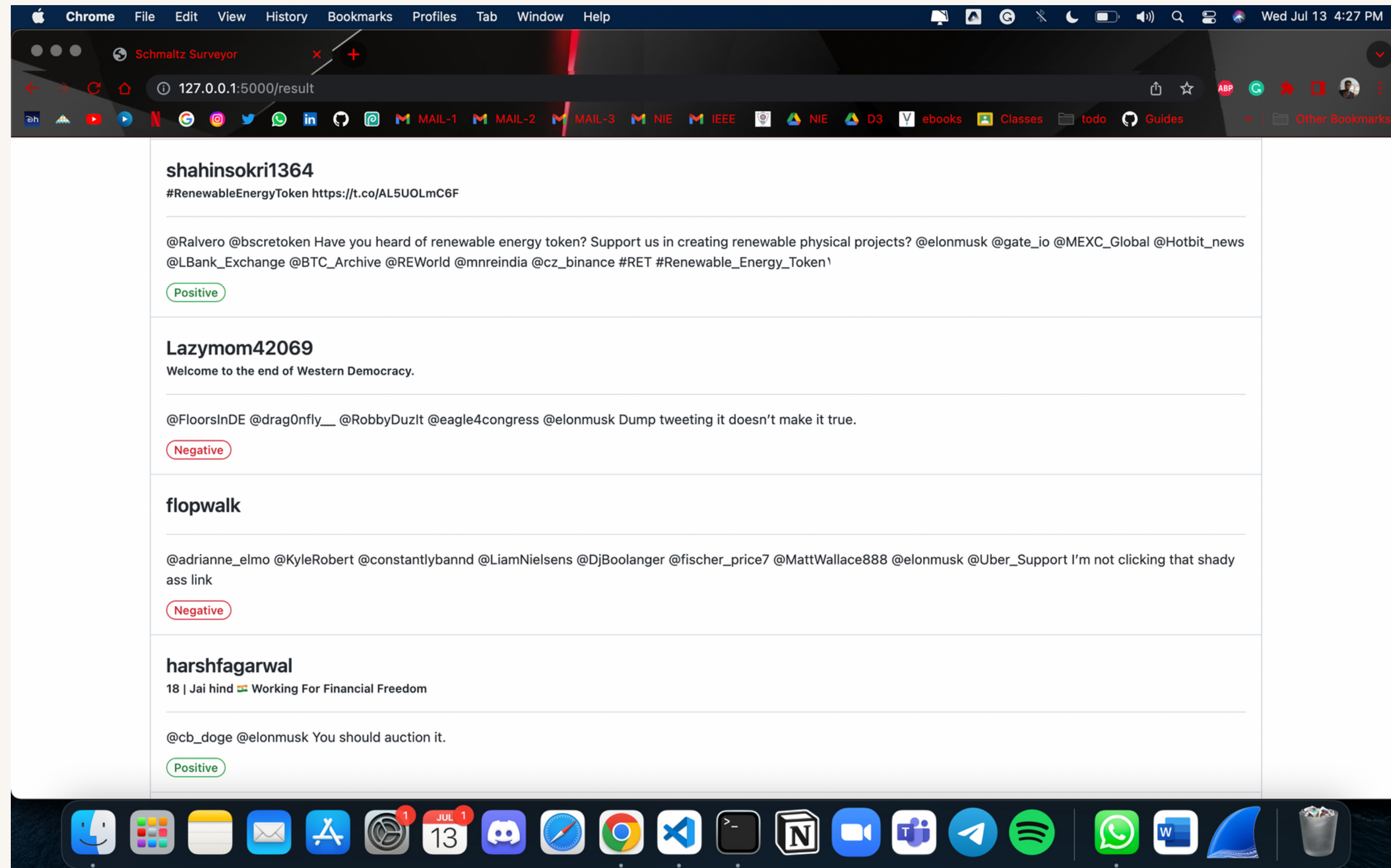
MarvinH2\_G2

I have a defective personality module.

@farid\_\_jalali Or call WHO's bluff & label by place of origin as more meaningful than 1/2/12/2.75 BA.1 Gauteng Strain BA.2 Philippines Strain BA.4 Limpopo Strain BA.5 KwaZuluNatal Strain BA.2.12 == New York Strain. BA.2.75 == India Strain I thought greek letters were to make it easier?

Negative

# LIVE Sentimental Analysis



Chrome File Edit View History Bookmarks Profiles Tab Window Help Wed Jul 13 4:27 PM

Schmaltz Surveyor 127.0.0.1:5000/result

MAIL-1 MAIL-2 MAIL-3 NIE IEEE NIE D3 ebooks Classes todo Guides Other Bookmarks

**shahinsokri1364**  
#RenewableEnergyToken <https://t.co/AL5UOLmC6F>

@Ralvero @bscretoken Have you heard of renewable energy token? Support us in creating renewable physical projects? @elonmusk @gate\_io @MEXC\_Global @Hotbit\_news @LBank\_Exchange @BTC\_Archive @REWorld @mnreindia @cz\_binance #RET #Renewable\_Energy\_Token¹

Positive

**Lazymom42069**  
Welcome to the end of Western Democracy.

@FloorsInDE @drag0nfly\_\_ @RobbyDuzIt @eagle4congress @elonmusk Dump tweeting it doesn't make it true.

Negative

**flopwalk**

@adrienne\_elmo @KyleRobert @constantlybannd @LiamNielsens @DjBoolanger @fischer\_price7 @MattWallace888 @elonmusk @Uber\_Support I'm not clicking that shady ass link

Negative

**harshfagarwal**  
18 | Jai hind 🇮🇳 Working For Financial Freedom

@cb\_doge @elonmusk You should auction it.

Positive

Mac OS dock with icons for Finder, Launchpad, Calendar, Mail, App Store, System Preferences, Calendar (JUL 13), Discord, Safari, Chrome, VS Code, Terminal, Notepad, Zoom, Teams, Telegram, Spotify, WhatsApp, Word, and a trash can.



# LIVE Sentimental Analysis

The screenshot shows a Chrome browser window on a Mac. The address bar displays `127.0.0.1:5000/result`. The page content is a surveyor interface with four entries, each showing a username, a bio, a tweet, and a sentiment label.

Username	Bio	Tweet	Sentiment
MdTanvir224411	I am student	@NikoFiftiii @elonmusk It's very good project. I hope this project is successful.	Positive
MotherCreator22	Mother and Creator	@elonmusk @Pontifex Standing far apart. Body language is awkward. What's going on here? Lol. Is someone a hostage? Which one is it really? Lol. Sorry, not sorry. Let go of JMD's remote Father	Negative
Jonatha33625234	common sense is critical!	@catturd2 @elonmusk @JoeTalkShow @BreitbartNews i have said it for a long time @elonmusk is a schill!	Positive
Biblecollege_	To know & proclaim Jesus Christ from East Asia to the World 야훼를 영화 롭게하고 영원히 예수 그리스도를 누리는 것. 순수한 종교 야고보서 1:27 彌迦 6:8 .只要你行公義好憐憫存謙卑的心與你的神同行	RT @elonmusk: Starship launch site tonight <a href="https://t.co/Len70RGcNf">https://t.co/Len70RGcNf</a>	Positive

# Conclusion

The result of the whole project is to

- Classifying the tweet into positive and negative would help us understand and analyze the extent of positivity and negativity on Twitter.
- Sentiment Analysis is a great way to analyze the response to a particular tweet.
- We have used 4 classifiers and they gave the following accuracy :
- Logistic Regression : 95.38%
- Support Vector Machine : 94.50%
- K Nearest Neighbour : 93.69%
- Random Forest : 95.60%

# Conclusion

- From the above data comparison of the classifiers used, it is clear that Random Forest Model is the way to go with the highest accuracy of approx. 95.60%
- Classifying the tweet into positive and negative would help us understand and analyze the extent of positivity and negativity on Twitter. Sentiment Analysis is a great way to analyze the response to a particular tweet. The model at a backend to a web application helps us analyze the sentiments, as and when needed, and helps us with a dashboard of the same as well.
- model as a backend to a web application that determines the toxicity of a comment which is provided as an input by the user.

# Future Enhancement

- Firstly, English is definitely one of the most largely used languages for communication on social media platforms. But there is a lot of content that is produced in regional languages as well. Detecting sentiment is definitely hard for regional languages as most of the models built are for English. But this is not an impossible task. This software can be used as a starting point for creating models to detect sentiment in regional languages.
- Secondly, such software can be used in a widespread manner and this functionality can be added by users who want to avoid unnecessary negative comments on their page and want to filter out negative content.

# References

- [1]. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1.
- [2] Faizan. "Twitter Sentiment Analysis" International Journal of Innovative Science and Research Technology (2019)
- [3] Chirag Kariya and Priti Khodke. "Twitter Sentiment Analysis" 2020 International Conference for Emerging Technology (INCET) Belgaum.
- [4]Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6.