**SOCIAL MEDIA SENTIMENTAL ANALYSIS**

A Project Report submitted in partial fulfilment of the requirements for the award of the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

By

Janvi Pangoriya(181500292)

Kajol Gupta(181500303)

Nidhi Gupta(181500422)

Pranav Tomar(181500475)

Under the Guidance of

**Mr Abhishek Verma**

**Department of Computer Engineering and Applications**

**Institute of Engineering and Technology**

****

**GLA University**

**Mathura-281406, INDIA**

**Dec, 2020**

**Department of Computer Engineering and Applications**

**GLA University, 17 km. Stone NH#2, Mathura-Delhi Road,**

**Chaumuha, Mathura – 281406 U.P (India)**

**Declaration**

I/we hereby declare that the work which is being presented in the B.Tech. Project **“Social Media Sentiment Analysis”**, in partial fulfillment of the requirements for the award ofthe ***Bachelor of Technology*** in Computer Science and Engineering and submitted to the Department of Computer Engineering and Applications of GLA University, Mathura, is an authentic record of my/our own work carried under the supervision of **Mr Abhishek Verma, Technical Trainer, Dept. of CEA,GLA University.**

The contents of this project report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree.

**Sign**: *Janvi Pangoriya* **Sign**: *Nidhi Gupta*

**Name of Candidate**: Janvi Pangoriya **Name of Candidate**: Nidhi Gupta

**University Roll No**.: 181500292 **University Roll No**.:181500422

**Sign:** *Kajol Gupta* **Sign**: *Pranav Tomar*

**Name of Candidate**: Kajol Gupta **Name of Candidate**: Pranav Tomar

**University Roll No**.:181500303 **University Roll No**.:181500475

**Department of Computer Engineering and Applications**

**GLA University, 17 km. Stone NH#2, Mathura-Delhi Road,**

**Chaumuha, Mathura – 281406 U.P (India)**

**Certificate**

This is to certify that the above statements made by the candidate are correct to the best of my/our knowledge and belief.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Supervisor**

Mr. Abhishek Verma

Technical Trainer

Dept of CEA,GLA University

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Project Coordinator** **Program Co-ordinator**

(Mr. Mayank Srivastava) ( Mr. Shashi Shekar)

Date:

**ACKNOWLEDGEMENT**

**P**resenting the ascribed project paper report in this very simple and official form, we would like to place my deep gratitude to GLA University for providing us the instructor Mr Abhishek Verma, our technical trainer and supervisor.

He has been helping us since Day 1 in this project. He provided us with the roadmap, the basic guidelines explaining on how to work on the project. He has been conducting regular meeting to check the progress of the project and providing us with the resources related to the project. Without his help, we wouldn’t have been able to complete this project.

And at last but not the least we would like to thank our dear parents for helping us to grab this opportunity to get trained and also my colleagues who helped me find resources during the training.

Thanking You.

**Sign:** Janvi Pangoriya **Sign**: Nidhi Gupta

**Name of Candidate**: Janvi Pangoriya **Name of Candidate**: Nidhi Gupta

**University Roll No**.:181500292 **University Roll No**.:181500422

**Sign**: Kajol Gupta **Sign**: Pranav Tomar

**Name of Candidate**:Kajol Gupta **Name of Candidate**: Pranav Tomar

**University Roll No**.:181500303 **University Roll No**.:181500475

**ABSTRACT**

Twitter is an online micro-blogging and social networking platform, which allows users to write short status, updates of maximum length 280 characters. These tweets reflect public sentiment about various topics and events happening. Analysing the public sentiment can help, firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. Sentiment analysis techniques are widely popular for this purpose Twitter is a widely popular micro-blogging platform for users to express their opinions about governmental issues, product items, sports and so forth. A tweet is a text-based post and has

only 280 characters. Twitter is a “what’s happening-right-

now” social network and hence tweets are valuable sources for businesses, government and individuals to determine

public’s opinion or sentiment about an entity (product, people,

topic, event etc.) Tweets reflect those events as seen by the individuals tweeting, and can be aggregated to form the basis of event exploration and visualization. However, the volume of tweets produced by Twitter every day is very vast. Hence, there is a need to automate the process of sentiment

analysis to ease the tasks of determining public’s opinions without having to read millions of tweets manually. This process of analysing and summarizing the opinions expressed in these huge opinionated user generated data is usually called Sentiment Analysis or Opinion Mining.

Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event . The attitude may be a judgment, affective state (the emotional state of the author or speaker), or the intended emotional communication (the emotional effect intended by the author or interlocutor). The sentiment can be classified as either positive, negative or neutral. There are mainly three techniques of sentiment classification. The Lexicon Based Approach uses dictionary, which contains positive and negative words to do the sentiment classification. The Machine Learning Based approach uses various supervised as well as unsupervised algorithms for classification purpose. The Hybrid Approach is the combination of both lexicon based approach and machine learning approach Lexicon based approach for sentiment classification deals with classifying the sentiments using an opinion lexicon. An opinion lexicon is a collection of positive and negative words. To classify a sentence as either positive or negative, the number of positive and negative words in a sentence is calculated. If the sentence contains more number of positive words then the sentence is classified as positive. If the sentence contains more number of negative words, then the sentence is classified as negative. If the sentence contains equal number of positive and negative words, then the sentence is classified as neutral.

Machine learning approach for sentiment classification uses two datasets i.e. training dataset and testing dataset. The supervised and unsupervised machine learning algorithms are first applied on training dataset. The classifier trains itself with respect to differentiating attributes of text. The model obtained after training is applied on test data which is unseen.

Machine learning technique for sentiment classification starts with collection of tweets. These tweets can be labelled or unlabelled. These tweets are noisy. They may contain words, punctuation marks or special characters, which do not express any sentiment. Hence, these tweets are first pre-processed to remove noise. Then the features relevant to sentiment analysis are extracted from the tweets. This forms the training data. A machine learning classifier is trained on this training dataset and it is then tested on the unseen test dataset.

This study proposes a sentiment analysis system using machine-learning approach. The study is based on sentence level sentiment classification. It is useful for finding out sentiments of people regarding any topic or event. It is also useful in predicting socio-economic phenomena like stock market prediction.

As a future work, this study can be expanded to include feature/aspect level classification, which is useful in product review and recommendation system. The number of sentiment classes can be increased to get more refined sentiment prediction. The study is a prototype and is meant to present the potential use of social networking platforms such as Twitter for large scale information gathering and processing for future social media related applications. The study can be extended for applications such as emergency management, social unrest etc. The study can be extended.

**CONTENT**

Title Page……………………………………………………………………………1

Declaration…………………………………………………………………… 2

Certificates…………………………………………………………………….3

Acknowledgement…………………………………………………………… 4

Abstract………………………………………………………………………. 5

Content……………………………………………………………………….. 7

List of Figures ………………………………………………………………...9

CHAPTER 1 Introduction to the Project…………………………………….. 10

* Context………………………………………………………………….10
* Motivation………………………………………………………………10
* Objective………………………………………………………………..11
* Dataset…………………………………………………………………..11

CHAPTER 2 Requirement Analysis…………………………………………. 12

* Impact of Social Media…………………………………………………………………....... 12
* Types of Sentiments…………………………………………..…………………….. 13
* Sentiment Analysis using Machine Learning…………………………………………………………………….15

CHAPTER 3 Software Design………………………………………………...19

* Use-case Diagram…………………………………………………………..19
* Data Flow Diagram(DFD)………………………………………………….20

CHAPTER 4 Methodology ……………………………………………………21

* Flow Chart………………………………………………………………… 23
* Dataset ………………….………………………………………………….24
* Data Pre-processing…………………………………………...……………25
* Data Visualization………………………………………………………..26
* Feature Extraction…………………………………………………………………29
* Algorithm used…………………………………………………………………….… 31

Chapter 5 Experimental Analysis and Results………………………………. 40

* Accuracy…………………………………………………………………..40
* Recall………………………………………………………………………41
* Precision……………………………………………………………………41
* F1-score…………………………………………………………………….42
* Confusion Matrix…………………………………………………………..43
* Output of various algorithms result confusion matrix………………..……44

Chapter 6 User Interface and software testing………...……………………….46

Chapter 7 Conclusion…………………………………………………………..49

*Refrences*

**LIST OF FIGURES**

1. Sentimental analysis using machine learning (flow chart)
2. Bar plot of Positive words in tweets
3. Pie chart of positive words in tweets
4. Bar plot of Negative words in tweets
5. Pie chart of negative words in tweets

**CHAPTER 1: INTRODUCTION .**

**CONTEXT**

This Machine Learning Project “Social Media Sentiment Analysis” has been submitted in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering at GLA University, Mathura supervised by Mr. Abhishek Verma. This project has been completed approximately three months and has been executed in modules, meetings have been organised to check the progress of the work and for instuctions and guidelines.

**MOTIVATION**

Social media can be very influential on society in both positive and negative ways. It gives people a way to stay in touch with people who live far away. It lets people share fun, interesting and informative content. It gives businesses a way to engage with customers.

One of the problems, however, is that anybody can share anything, including material that may not be accurate. In some cases, real harm is done when people spread inflammatory, unverified or outright false information. This can harm private individuals, as when someone is bullied online. It can also have a harmful impact on society as a whole.

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools make that process quicker and easier than ever before, thanks to real-time monitoring capabilities.

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world.

Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

**OBJECTIVE**

Social media sentiment is the attitude and feelings people have about a brand on social media, about a particular event that happened or some incident or a current topic that matters and holds value in the real world. It adds context to all the @-mentions, comments, and shares.

To figure out where the people indulge in, stand on the positive/negative spectrum, we need to analyse the conversations of people in order to protect some other from various comments or thinking that might have a negative impact on the society we are living in. So in this project we are going to analyse the comments and conversation of one of the social media platforms Twitter and classify the tweets among two categories:- Racist and Non-Racist.

Keyword spotting is one of the simplest techniques and leveraged, widely by us in Sentiment Analysis algorithms. After the completion of the project, the fed Input document will thoroughly scanned for the obvious positive and negative words like “sad”, “happy”, “disappoint”, “great”, “satisfied”, and such. The tweets will be marked as racist if it does not contain any negative word or hate speech and the others will be marked as non-racist.

**SOURCES**

The sources of our project (including all the project work, documentations and presentations) will is available at the following link <https://github.com/JanviPangoriya/social_media_sentiments-analaysis> and the dataset we have taken is from kaggle and the link for the dataset is <https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech>

**CHAPTER 2: REQUIREMENT ANALYSIS .**

**IMPACT OF SOCIAL MEDIA**

Social media has a huge impact on individuals and their lives. While some impacts can be positive, social media has been shown to negatively affect things like our moods and stress levels. Addiction is caused by social media too. With access to it anytime of day on our phones, we read a number of things create multiple mindset and post things that might feel offensive to others. Multiple studies have shown that unlimited use of social media causes stress, bad moods and negative mental health. Apart from all this whole of the world is connected at our fingertips. Any news can spread like fire. And among all the news comes the fake news which can create a different mindset among people bullying the ones who are innocent and misbehaving with the others. This needs to be controlled. Many people wake up in the morning and immediately check their Instagram, Snapchat or Twitter.  On platforms like Instagram, some users feel obligated to edit their posts in order to fit the terms of “attractiveness.” .This is where the people get wrong .Users may begin to compare themselves to others and think, “am I good enough?” “Am I pretty enough?” Social media is a powerful thing. It can help people across the world connect and inspire people to achieve social change. However, we must also realize the negative implications social media has on our lives

.

This study aims to investigate the effect of social media and conventional media, their relative importance, and their interrelatedness features daily media content across various conventional media and social media outlets for 824 public traded firms across 6 industries. Social media outlets include blogs, forums, and Twitter. Conventional media includes major newspapers, television broadcasting companies, and business magazines. We apply the advanced sentiment analysis technique that goes beyond the number of mentions (counts) to analyse the overall sentiment of each media resource toward a specific company on the daily basis

.

Hate speech in recent times has becoming a troubling development. It has different meaning to different people in different cultures any of them posted online creates bullying of individuals. The anonymity and ubiquity of social media provides a breeding ground for hate speech and makes combating it seems like lost battle. However what any constitute a hate speech in a cultural or religious neutral society may not be perceived as such in a polarised multi-cultural and multi-religious society .Therefore defining hate speech may be contextual. Therefore everything perceived along ethnic, religious and political boundaries .The purpose of this project is to check the presence of hate speech in social media platform like Twitter.

The social media being an avenue where people may easily pop-in to make new friends and express their diverse opinion on trending issues all over the world has recently become the avenue where people express their anger and hatred toward other people or the government in power. Sentiments in these media are being expressed in the form of name-calling, insinuations of race or tribal superiority, religious bigotry, abuses, or posting of inciting comments, images and videos, especially on Twitter and What’s App. The one posting the sentiment may call it fighting a cause, while the one who receives it and is offended may call it a hate speech. Some of the problems caused by hate speech may include the promotion of violence, discrimination, disintegration, communal wars and ultimately, loss of lives and properties.

This mechanism can be used by various social media platforms to keep a check of the offensive words being used, that might hurt the sentiment of people, bring down the brand name and may create chaos among people creating a mindset that may sound offensive to them.

**SENTIMENTS AND ITS TYPES**

Hate speech in recent times has become a troubling development. It has different meanings to different people in different cultures. The anonymity and ubiquity of the social media provides a breeding ground for hate speech and makes combating it seems like a lost battle. However, what may constitute a hate speech in a cultural or religious neutral society may not be perceived as such in a polarized multi-cultural and multi-religious society like Nigeria. Defining hate speech, therefore, may be contextual. Hate speech in Nigeria

may be perceived along ethnic, religious and political boundaries. The purpose of this paper is to check for the presence of hate speech in social media platforms like Twitter, and to what degree is hate speech permissible, if available? It also intends to find out what monitoring mechanisms the social media platforms like Facebook and Twitter have put in place to combat hate speech.

**Sentiment analysis** is one of the Natural Language Processing fields, dedicated to the exploration of subjective opinions or feelings collected from various sources about a particular subject. Sentiment analysis is a predominantly classification algorithm aimed at finding an opinionated point of view and its disposition and highlighting the information of particular interest in the process.

Sentiment Analysis deals with the perception of the tweet or text written and understanding of the matter or topic through the lens of sentiment data.

There are many sources of public and private information out of which you can harness an insight into the user’s perception of the certain current issues and general market situation.

To understand how to apply sentiment analysis in the context of survey operation - we need to understand its different types.

In this section, we will look at the main types of sentiment analysis.

* **1st type.** **Fine-grained Sentiment Analysis**

**It i**nvolves determining the polarity of the opinion. It can be a simple binary positive/negative sentiment differentiation. This type can also go into the more high specification (for example, very positive, positive, neutral, negative, very negative), depending on the use case (for example, as in five-star Amazon reviews).

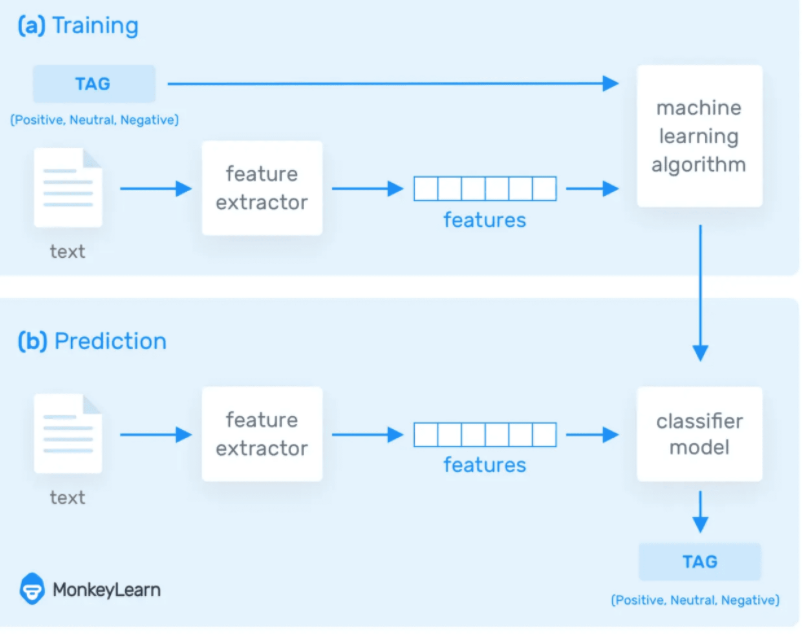
* **2nd type.** **Emotion detection**

**It** is used to identify signs of specific emotional states presented in the text. Usually, there is a combination of lexicons and machine learning algorithms that determine what is what and why.

* **3rd type. Aspect-based sentiment analysis** goes deeper. Its purpose is to identify an opinion regarding a specific element of the product. For example, the brightness of the flashlight in the smartphone. The aspect-based analysis is commonly used in product analytics to keep an eye on how the product is perceived and what are the strong and weak points from the customer point of view.
* **4th type.** **Intent Analysis**

It is all about the action. Its purpose is to determine what kind of intention is expressed in the message. It is commonly used in customer support systems to streamline the workflow.

**SENTIMENT ANALYSIS USING MACHINE LEARNING**



While the popularity of Social Network is raising the field of Social network Analysis has become an important and interesting study in the area. Social Network analysis refers to the process of exploring social structures through the use of network and graphs. The information on the social network is unstructured and there is a need to extract the structured information for making use of the valuable information. Extracting information from the social network is the exploration that empowers the use of such a massive amount of unstructured distributed information in a structured way. Natural language processing is employed to enhance the accuracy in visualizing the structured information that is speckled over the social network. The foremost notion of monitoring is to analyse the meaningful information from texts written by naïve users of social network. It analyses natural language text in order to extract information about different types of entities, relationships or events. The Natural Language methods are being looked closely by means of this research. In this paper researcher attempts to review various text mining systems which is the keystone of Natural Language Processing to analyse social network information.

This is one of the reasons why detecting sentiment from natural language (NLP or natural language processing) is a surprisingly complex task. Any machine learning model that hopes to achieve suitable accuracy needs to be able to determine what textual information is relevant to the prediction at hand, have an understanding of negation, human patterns of speech, idioms, metaphors, etc, and be able to assimilate all of this knowledge into a rational judgment about a quantity as nebulous as “sentiment.”

In fact, when presented with a piece of text, sometimes even humans disagree about its tonality, especially if there’s not a fair deal of informative context provided to help rule out incorrect interpretations. With that said, recent advances in deep learning methods have allowed models to improve to a point that is quickly approaching human precision on this difficult task.

The primary role of machine learning in sentiment analysis is to improve and automate the low-level text analytics functions that sentiment analysis relies on, including Part of Speech tagging. For example, data scientists can train a machine learning model to identify nouns by feeding it a large volume of text documents containing pre-tagged examples. Using supervised and unsupervised machine learning techniques, such as neural networks and deep learning, the model will learn what nouns look like.

Once the model is ready, the same data scientist can apply those training methods towards building new models to identify other parts of speech. The result is quick and reliable Part of Speech tagging that helps the larger text analytics system identify sentiment-bearing phrases more effectively.

Machine learning also helps data analysts solve tricky problems caused by the evolution of language. For example, the phrase “sick burn” can carry many radically different meanings. Creating a sentiment analysis ruleset to account for every potential meaning is impossible. But if you feed a machine learning model with a few thousand pre-tagged examples, it can learn to understand what “sick burn” means in the context of video gaming, versus in the context of healthcare. And you can apply similar training methods to understand other double-meanings as well.

Most importantly, “machine learning” really means “machine teaching.” We know **what** the machine needs to learn, so our task is to create a **learning framework** and provide properly-formatted, relevant, clean **data** for the machine to learn from.

When we talk about a “model,” we’re talking about a mathematical representation. Input is key. A machine learning model is the sum of the learning that has been acquired from its training data. The model changes as more learning is acquired.

Unlike algorithmic programming, a machine learning model is able to generalize and deal with novel cases. If a case resembles something the model has seen before, the model can use this prior “learning” to evaluate the case. The goal is to create a system where the model continuously improves at the task you’ve set it.

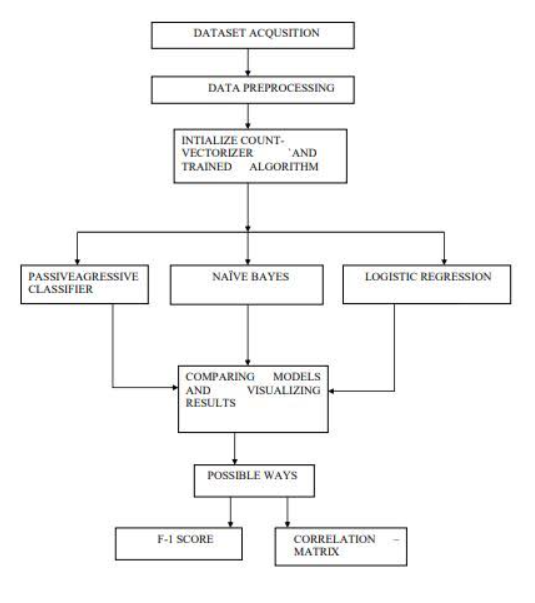
Machine learning for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text. The techniques can be expressed as a model that is then applied to other text, also known as supervised machine learning. It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning. It’s important to understand the difference between supervised and unsupervised learning, and how you can get the best of both in one system.

Machine learning for NLP helps data analysts turn unstructured text into usable data and insights.

Text data requires a special approach to machine learning. This is because text data can have hundreds of thousands of dimensions (words and phrases) but tends to be very sparse. For example, the English language has around 100,000 words in common use. But any given tweet only contains a few dozen of them. This differs from something like video content where you have very high dimensionality, but you have oodles and oodles of data to work with, so, it’s not quite as sparse.

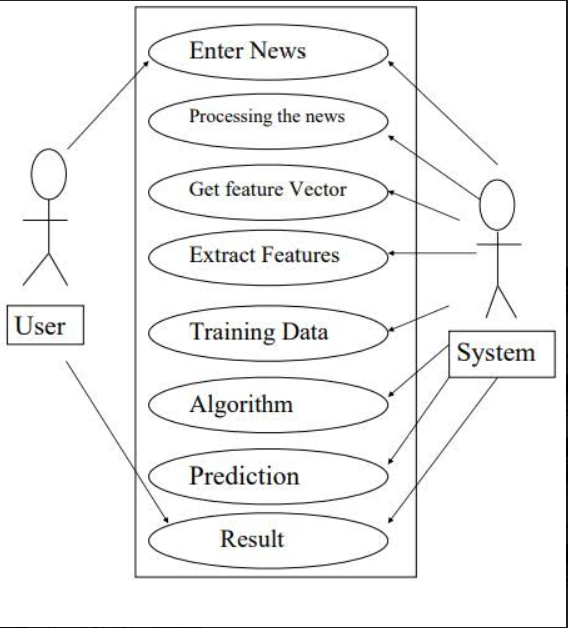
**Supervised Machine Learning for Natural Language Processing and Text Analytics**

## In supervised machine learning, a batch of text documents are tagged or annotated with examples of what the machine should look for and how it should interpret that aspect. These documents are used to “train” a statistical model, which is then given un-tagged text to analyse.Later, you can use larger or better datasets to retrain the model as it learns more about the documents it analyses. For example, you can use supervised learning to train a model to analyse movie reviews, and then later train it to factor in the reviewer’s star rating.

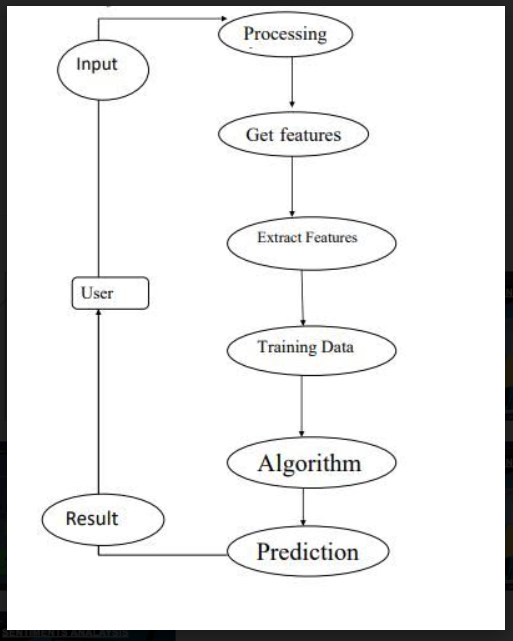


**CHAPTER 3: SOFTWARE DESIGN .**

* **USE-CASE DIAGRAM**



**DATA FLOW DIAGRAM**



**CHAPTER 4: METHODOLOGY .**

Fake news generally carries strong sentiments and thus circulates in no time on social media. Response based technique takes into consideration the collected responses on tweets/posts to determine the credibility of the news.

1) Collection of data from social media platform, Facebook and Twitter

2) Choosing relevant features for classification and Training the Model

3) Evaluation of different model performance based on extracted features

4) Improving performance

5) Discussion and Presentation of results

This project was developed in Python using Sci-kit libraries. Python has a huge set of libraries and extensions, which can be easily used in Machine Learning. Sci-Kit Learn library is the best source for machine learning algorithms where nearly all types of machine learning algorithms are readily available for Python, thus easy and quick evaluation of ML algorithms is possible.

**Machine Learning Workflow**:

There are five core tasks in the common ML workflow:

**1. Get Data:** The first step in the Machine Learning process is getting data. This process depends on your project and data type. For example, are you planning to collect real-time data from an IoT system or static data from an existing database? You can also use data from internet repositories sites such as Kaggle and others.

**2. Clean, Prepare & Manipulate Data**: Real-world data often has unorganized, missing, or noisy elements. Therefore, for Machine Learning success, after we chose our data, we need to clean, prepare, and manipulate the data. This process is a critical step, and people typically spend up to 80% of their time in this stage. Having a clean data set helps with your model’s accuracy down the road. After getting the data to a state you like, you need to convert the data sets into valid formats for your chosen ML platform. For example, you may need to translate the data into a .CSV file. Finally, you split your data into training and test data sets. The training set is used to train the model in the next step, while the test data is used to validate the model in the fourth step. The typical default is a 70/30 split between training and test sets.

**3. Train Model**: This step is where the magic happens! The data set connects to an algorithm, and the algorithm leverages sophisticated mathematical modeling to learn and develop predictions. These algorithms commonly fall into one of three categories: Binary – Classify into two categories Classification – Classify into many categories Regression – Predict a numeric

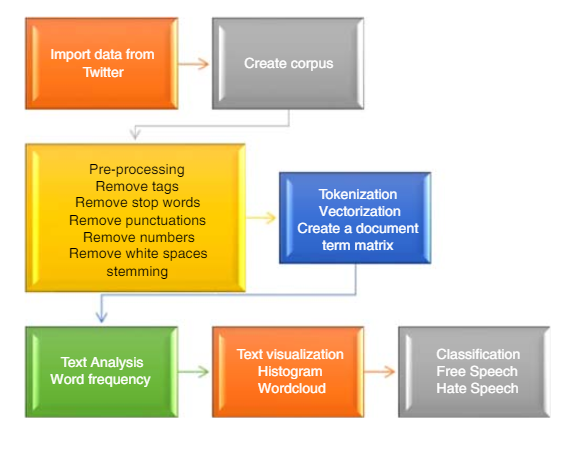
**4. Test Model:** Now, it’s time to validate your trained model. Using the test data from Step 3, we check the model’s accuracy. If the results are not satisfactory, you need to improve and retrain your ML model

**5. Improve:** Practice makes perfect! Here are a few things you can do to refine your model and improve accuracy: Review your model’s results with your business stakeholders. Are there other data elements worth adding to your model to make it more accurate?

Reconsider your algorithm choice. Within each class of algorithm, there are dozens of algorithm choices. A different algorithm may perform better for you

Adjust the parameters of your chosen algorithm to improve performance. Sometimes small adjustments have a significant impact.

**FLOW CHART**



**DATASET EXPLAINATION**

The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets.

The dataset used in our project is taken from well know dataset factory i.e. kaggle.com .Formally, given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist, our objective is to predict the labels on the test dataset.

It has A full training dataset with the following attributes:

• id: unique id for tweets

• tweet: the text of the article; could be incomplete

• label: a label that marks the tweet as potentially racist.

Full tweet texts are provided with their labels for training data.  
Mentioned users' username is replaced with [@user](https://www.kaggle.com/user).

For implementation purpose I have choosed the training set file ,two column from the file for classification that is tweet and label.

Originally there were 5 types of fine grained sentiments:

1. Highly positive is considered as non-racist (1)
2. Positive is considered as non-racist (1)
3. Neutral is considered as non-racist (1)
4. Negative is considered as racist (0)
5. Highly negative is considered as racist (0)

**PREPROCESSING**

After actually getting a hold of your text data, the first step in cleaning up text data is to have a strong idea about what you’re trying to achieve.

1. **Load Data:** The text is small and will load quickly and easily fit into memory. This will not always be the case and you may need to write code to memory map the file. Tools like NLTK will make working with large files much easier.

2. **Split by Whitespace**: Clean text often means a list of words or tokens that we can work with in our machine learning models. This means converting the raw text into a list of words and saving it again. We can also use select words to select the words (use the regex model (re) and split the document into words by selecting for strings of alphanumeric characters (a-z, A-Z, 0-9 and ‘\_’).

3. **Remove Punctuation**: We may want the words, but without the punctuation like commas and quotes. We also want to keep contractions together. One way would be to split the document into words by white space (as in “2. Split by Whitespace“), then use string translation to replace all punctuation with nothing (e.g. remove it).

4. **Normalizing Case**: It is common to convert all words to one case. This means that the vocabulary will shrink in size, but some distinctions are lost (e.g. “Apple” the company vs. “apple” the fruit is a commonly used example).

5. **Filter out Stop Words (and Pipeline):** Stop words are those words that do not contribute to the deeper meaning of the phrase. They are the most common words such as: “the“, “a“, and “is“. For some applications like documentation classification, it may make sense to remove stop words. NLTK provides a list of commonly agreed upon stop words for a variety of languages, such as English, French.

6. **Stemming:** Stemming refers to the process of reducing each word to its root or base. For example “fishing,” “fished,” “fisher” all reduce to the stem “fish.” Some applications, like document classification, may benefit from stemming in order to both reduce the vocabulary and to focus on the sense or sentiment of a document rather than deeper meaning. There are many stemming algorithms, although a popular and longstanding method is the Porter Stemming algorithm. This method is available in NLTK via the Porter Stemmer class.

**DATA VISUALIZATION**

"A picture is worth a thousand words". We are all familiar with this expression. It especially applies when trying to explain the insight obtained from the analysis of increasingly large datasets. Data visualization plays an essential role in the representation of both small and large-scale data. Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed.

Python offers multiple great graphing libraries that come packed with lots of different features. No matter if you want to create interactive, live or highly customized plots python has an excellent library

In our project we have done the data visualization using word cloud.

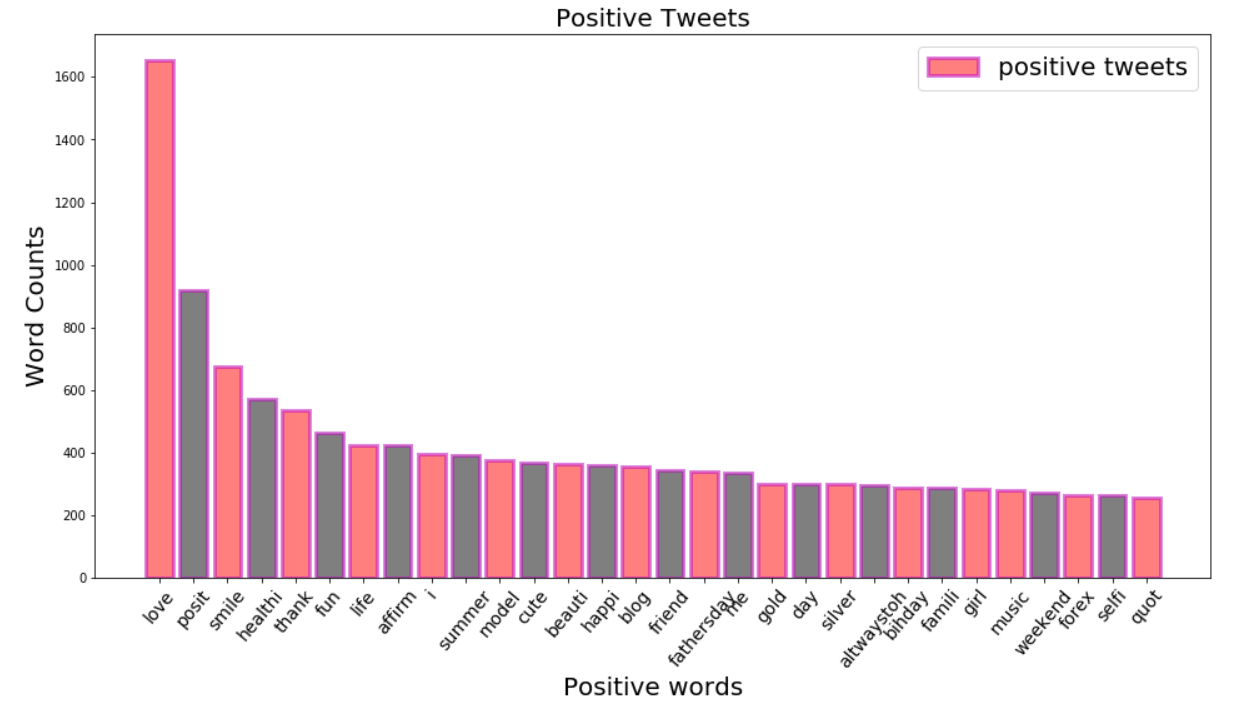
As we are analysing our dataset from the Twitter tweets so the first word cloud is created in the shape of hamming bird.

And the second word cloud is created in the form of squares and the words arranged horizontally and vertically.

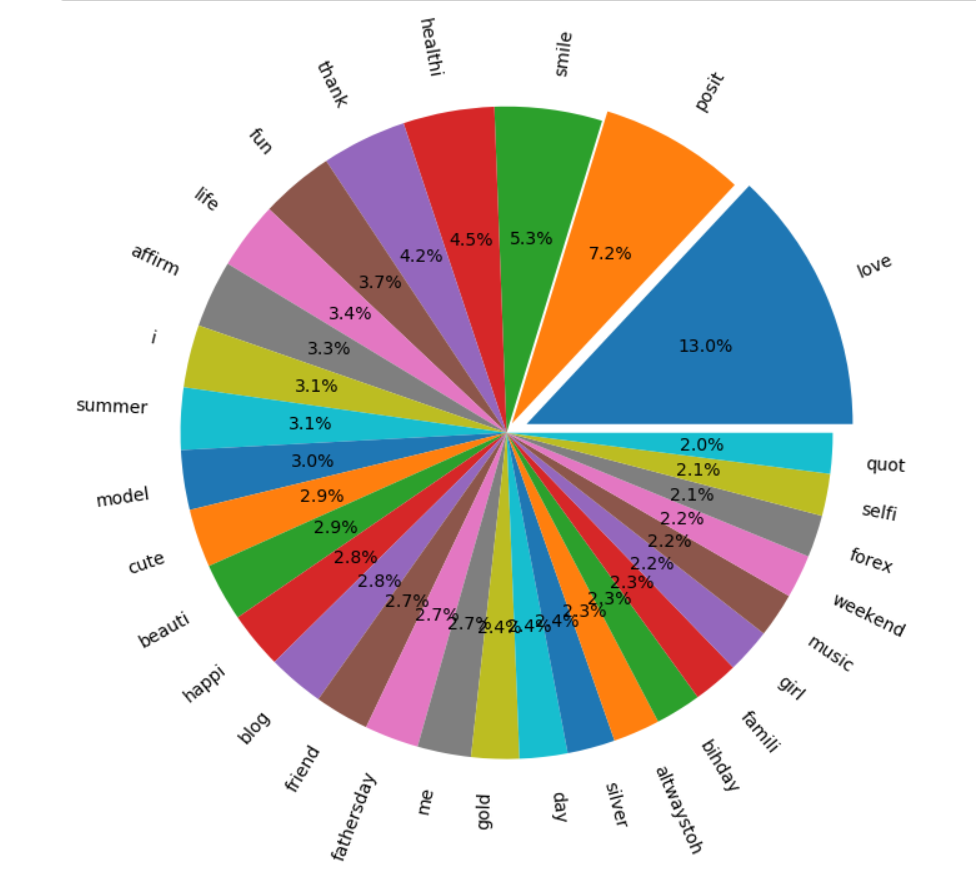




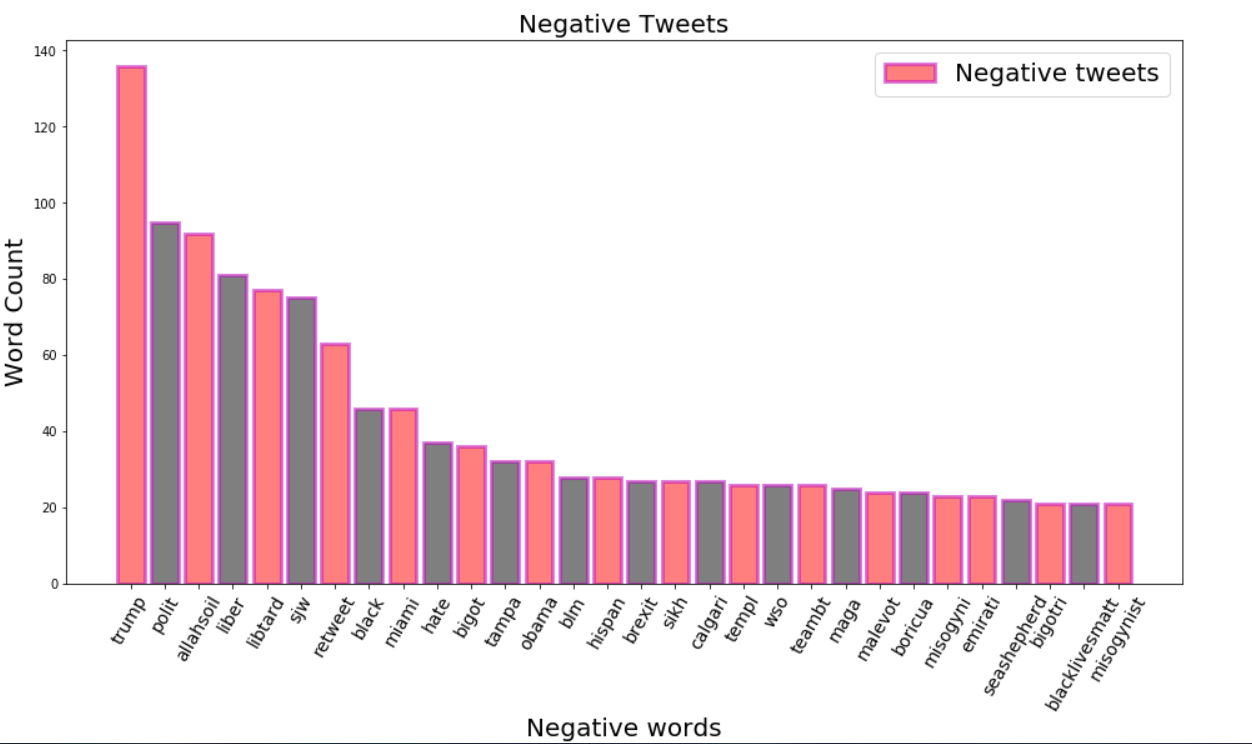
**BAR PLOT FOR POSITIVE WORDS**



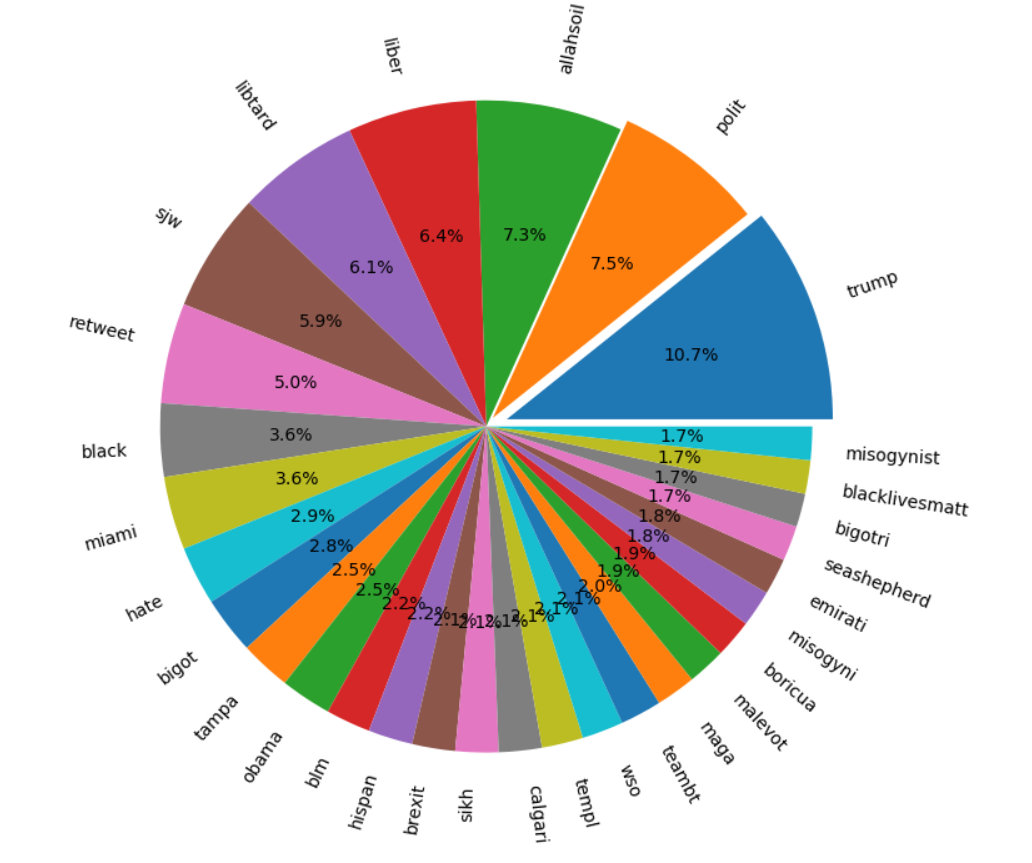
**PIE CHART FOR POSITIVE WORDS**



**BAR PLOT FOR NEGATIVE WORDS**



**PIE CHART FOR NEGATIVE WORDS**



**FEATURE EXTRACTION**

We cannot work with text directly when using machine learning algorithms. Instead, we need to convert the text to numbers. We may want to perform classification of documents, so each document is an “input” and a class label is the “output” for our predictive algorithm. Algorithms take vectors of numbers as input; therefore we need to convert documents to fixed-length vectors of numbers.

To create the vectors of words we use an approach **Bag of Words Algorithms**

**Bag of Words Algorithm**

Bag of words is a way of representing text data when modelling with machine learning algorithm. To convert the text to numbers, in precise vector of numbers, so for extraction of text we use this algorithm. Bag of words describes occurrence of words within the document. And hence creates vocabulary of known words. Here model is concerned with whether the words are known to the documents and their position does not matter. This can be done by assigning each word a unique number. Then any document we see can be encoded as a fixed-length vector with the length of the vocabulary of known words. The value in each position in the vector could be filled with a count or frequency of each word in the encoded document. This is the bag of words model, where we are only concerned with encoding schemes that represent what words are present or the degree to which they are present in encoded documents without any information about order.

**Step 1: Collect Data**

Eg. Let the document be

[Document 1]“ It was the best of time

[Document 2]It was the worst of times

[Document 3]It was the age of wisdom

[Document 4]It was the age of foolishness”

**Step 2: Design Vocabulary**

1. It
2. Was
3. The
4. Best
5. Worst
6. Age
7. Of
8. Wisdom
9. Foolishness
10. Times

**Step 3: Create Document Vectors**

To score words in each document and hence create vector for each document which can be used as input and output to machine learning model

0-----🡪 for absent

1-----🡪for present

Document 1 = [ 1, 1, 1, 1, 0, 0, 1, 0, 0, 1]

Document 4 = [ 1, 1, 1, 0, 0, 1, 1, 1, 0, 0]

**Step 4:Calculate the frequency of the words that how many times a word appears in the document—(Count Vectorizer)**

Count Vector is a matrix notation of the dataset, in which rows represent the documents in the corpus, columns represent a term from the corpus, and cells represent the count of that particular term in a particular document. The dictionary is created using the list of unique tokens or words in the corpus.

**Count Frequency** = Frequency of a word in a document

Number of the words in the vocabulary

Count Frequency counts number of times each word appears in document.

**LIMITATIONS OF BAG OF WORDS ALGORITHM**

* The vocabulary should be carefully designed.
* Sparsity -computational problems +space complexity +time complexity +memory
* Meanings get altered as the word order is not maintained
* As the size of the documents increases the vocabulary size increases, vector representation of elements increases causing storage problems

**TFIDF VECTORIZER**

To solve the problem that occurs scoring the words using Count Vectorizer using frequency is that higher frequent words start to dominate the document and might not contain “information content”.

The approach is to rescale the frequency of words that appear often in all documents. Like the short words “the” can be penalised if not removed. The approach is called term frequency or Inverse Document Frequency.

This can be better understood as :-

Term frequency is defined as scoring of the words normally in current documents

Inverse Document Frequency is the scoring of how rare the word is along the document.

Scores have the effect of highlighting words that are different (useful information container) in document. i.e inverse document frequency of the rare term is higher and inverse document frequency of frequent term is lower.

**ALGORITHM USED**

In this following project of classifying the tweets into racist and non-racist we have used the following classification algorithms after the data pre-processing and feature extraction. The algorithms are:-

* Naïve Bayes Algorithm
* Logistic Regression
* K Nearest Neighbors
* Passive Aggressive Classifier
* Decision Tree Classification
* Random Forest

**NAÏVE BAYES ALGORITHM**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes’ Theorem is stated as:

P (h|d) = (P (d|h) \* P (h)) / P (d)

Where

❖ P (h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.

❖ P (d|h) is the probability of data d given that the hypothesis h was true.

❖ P (h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.

❖ P (d) is the probability of the data (regardless of the hypothesis).

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis is simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P (d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) \* P(d2|H) and so on.

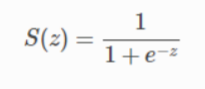
This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold

**LOGISTIC REGRESSION**

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

**Sigmoid activation**

In order to map predicted values to probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



**Decision boundary**

Our current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class (true/false, cat/dog), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.

p≥0.5, class=1

p<0.5, class=0

For example, if our threshold was .5 and our prediction function returned .7, we would classify this observation

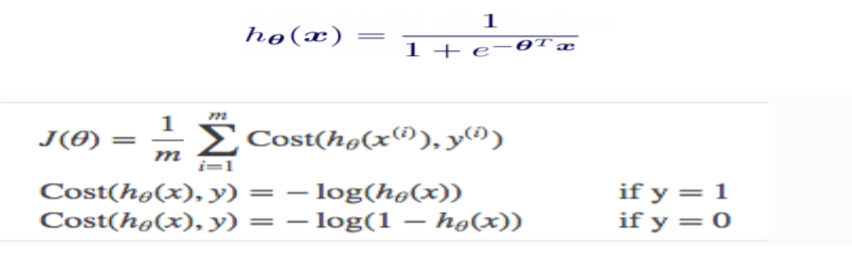
as positive. If our prediction was .2 we would classify the observation as negative. For logistic regression with multiple classes we could select the class with the highest predicted probability.

**Making predictions**

Using our knowledge of sigmoid functions and decision boundaries, we can now write a prediction function. A prediction function in logistic regression returns the probability of our observation being positive, True, or “Yes”. We call this class 1 and its notation is P(class=1)P(class=1). As the probability gets closer to 1, our model is more confident that the observation is in class 1.

**COST FUNCTION**

Since the hypothesis function for logistic regression is sigmoid in nature hence, The First important step is finding the gradient of the sigmoid function. We can see from the derivation below that gradient of the sigmoid function follows a certain pattern.

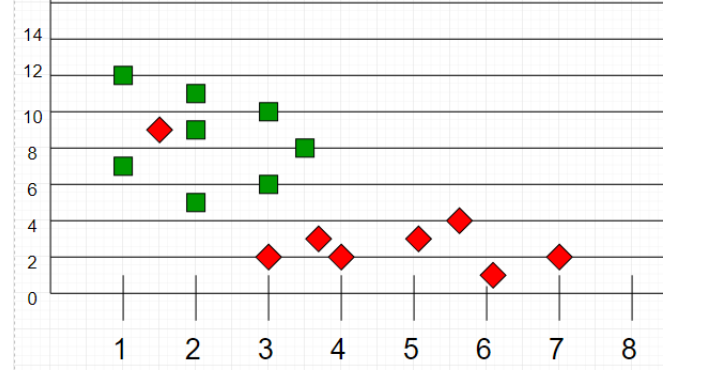


This is the required cost function for the logistic regression.

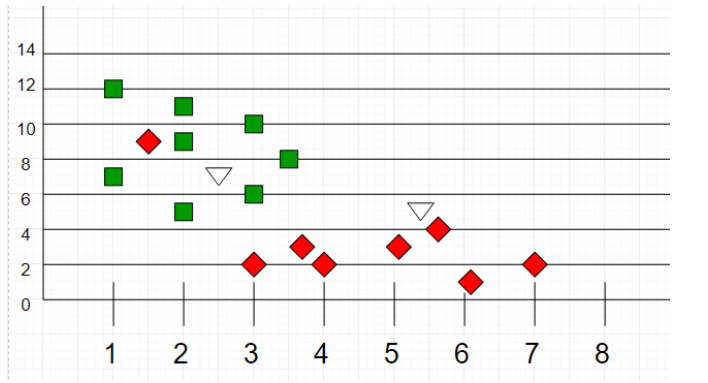
**K NEAREST NEIGHBORS**

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute. As an example, consider the following table of data points containing two features:



Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set. Note that the unclassified points are marked as ‘White’.



If we plot these points on a graph, we may be able to locate some clusters or groups. Now, given an unclassified point, we can assign it to a group by observing what group its nearest neighbors belong to. This means a point close to a cluster of points classified as ‘Red’ has a higher probability of getting classified as ‘Red’.

Intuitively, we can see that the first point (2.5, 7) should be classified as ‘Green’ and the second point (5.5, 4.5) should be classified as ‘Red’.

K can be kept as an odd number so that we can calculate a clear majority in the case where only two groups are possible (e.g. Red/Blue). With increasing K, we get smoother, more defined boundaries across different classifications. Also, the accuracy of the above classifier increases as we increase the number of data points in the training set.

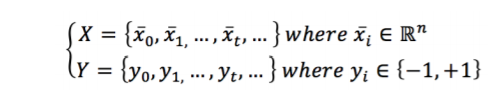
**PASSIVE AGGRESSIVE CLASSIFIER**

Passive Aggressive Classifier: The Passive Aggressive Algorithm is an online algorithm; ideal for classifying massive streams of data (e.g. twitter). It is easy to implement and very fast. It works by taking an example, learning from it and then throwing it away.

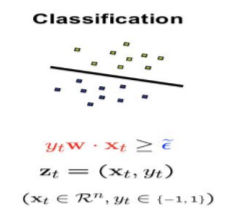
Passive Aggressive Algorithms are a family of online learning algorithms (for both classification and regression) proposed by Crammer at al. The idea is very simple and their performance has been proofed to be superior to many other alternative methods like Online Perception.

Classification

Let’s suppose to have a dataset:

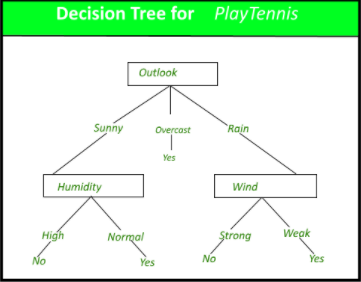


The index t has been chosen to mark the temporal dimension. In this case, in fact, the samples can continue arriving for an indefinite time. Of course, if they are drawn from same data generating distribution, the algorithm will keep learning (probably without large parameter modifications), but if they are drawn from a completely different distribution, the weights will slowly forget the previous one and learn the new distribution. For simplicity, we also assume we’re working with a binary classification based on bipolar labels.



**DECISION TREE CLASSIFIER**

**Decision Tree :**Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



 A classification model is typically used to,

* Predict the class label for a new unlabeled data object
* Provide a descriptive model explaining what features characterize objects in each class

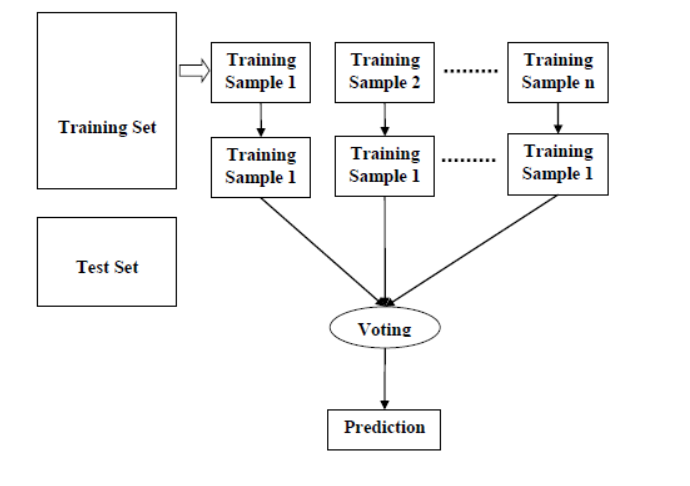
A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

**RANDOM FOREST CLASSIFIER**

The Random Forest (RF) classifiers are suitable for dealing with the high dimensional noisy data in text classification. An RF model comprises a set of decision trees each of which is trained using random subsets of features. Given an instance, the prediction by the RF is obtained via majority voting of the predictions of all the trees in the forest. However, different test instances would have different values for the features used in the trees and the trees should contribute differently to the predictions. This diverse contribution of the trees is not considered in traditional RFs. Many approaches have been proposed to model the diverse contributions by selecting a subset of trees for each instance.

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART models with different samples and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

* **Step 1** − First, start with the selection of random samples from a given dataset.
* **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
* **Step 3** − In this step, voting will be performed for every predicted result.
* **Step 4** − At last, select the most voted prediction result as the final prediction result.

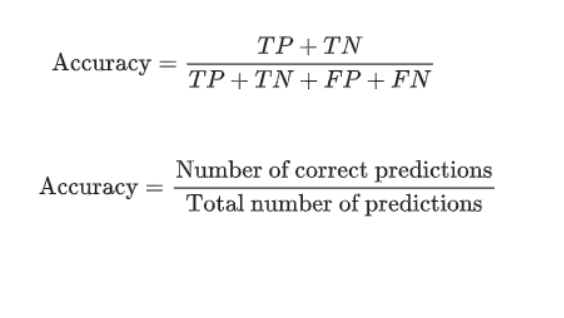


**CHAPTER 5: EXPERIMENTAL AND RESULT ANALYSIS**

In this section we are going to deal with testing, testing is finding out how well something works. In terms of human beings, testing tells what level of knowledge or skill has been acquired. In computer hardware and software development, testing is used at key checkpoints in the overall process to determine whether objectives are being met. There are various techniques for testing the accuracy but we are going to use some of them.

**ACCURACY**

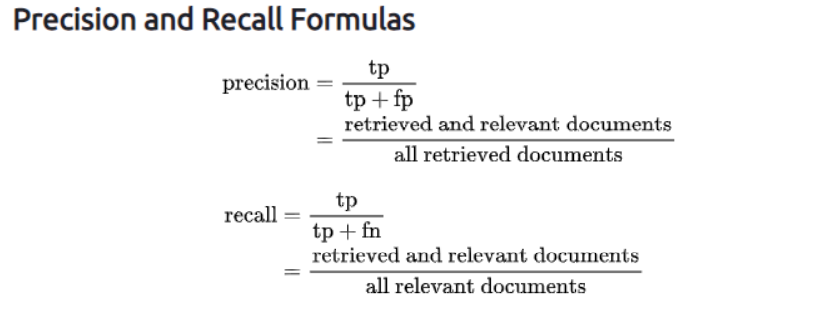
It is the most common evaluation metric for classification problems. It is defined as the number of correct predication as against the number of total predictions. However, this metric alone cannot give enough information to decide whether the model is a good one or not. It is suitable when there are equal numbers of observation in every class.



**F1\_SCORE**

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives. A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified. For example, if the algorithm classified a false data point as true, it would be a false positive. Often, accuracy is used along with precision and recall, which are other metrics that use various ratios of true/false positives/negatives.

Precision and recall are two numbers which together are used to evaluate the performance of classification or information retrieval systems. Precision is defined as the fraction of relevant instances among all retrieved instances. Recall, sometimes referred to as ‘sensitivity, is the fraction of retrieved instances among all relevant instances. Perfect classifiers have precision and recall both equal to 1.

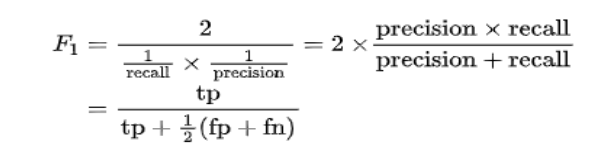


The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall.

The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

The formula for the standard F1-score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1.



**CONFUSION MATRIX**

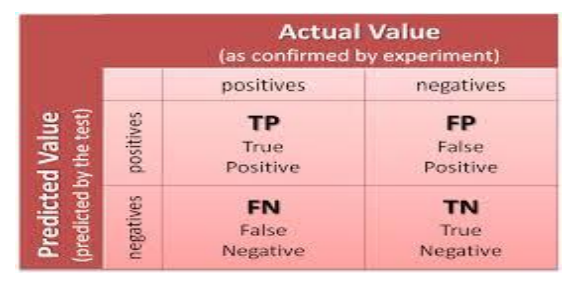
It is also known as Error matrix, which is a table representation that shows the performance of the model. It is special kind of Contingency table 21 having two dimensions- “actual”, labelled on x-axis and “predicted” on y-axis. The cells of the table are the number of predictions made by the algorithm.

True Positives: It is correctly predicted positive values.

True Negatives: It is correctly predicted negative values.

False Positives: It is incorrectly predicted negative values as positive values.

False Negatives: It is incorrectly predicted negative values as positive values.

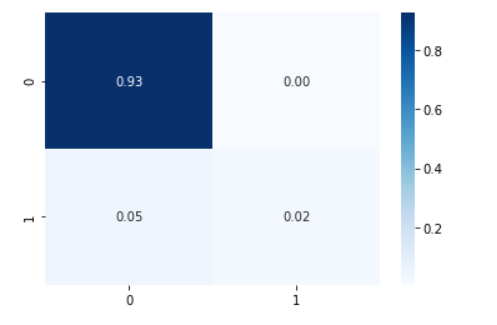


**RESULT ANALYSIS**

We have performed six algorithms on our dataset namely Naïve Bayes algorithm, K Nearest Neighbors, Logistic Regression, Passive Aggressive Classifier, Decision tree and Random forest. And here we have tabulated the results.

|  |  |  |
| --- | --- | --- |
| Name Of The Algorithm | Accuracy using Count Vectorizer | Accuracy using TFIDF |
| NAÏVE BAYES | 94.93% | 94.88% |
| LOGISTIC REGRESSION | 96.02% | 95.35% |
| K NEAREST NEIGHBORS | 93.71% | 94.41% |
| PASSIVE AGRESSIVE | 94.57% | 95.33% |
| DECISION TREE CLASSIFIER | 94.32% | 91.52% |
|  |  |  |

CONFUSION MATRIX FOR NAÏVE BAYES ALGORITHM

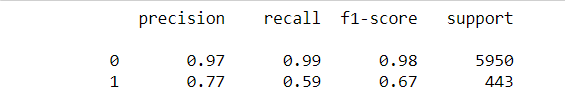


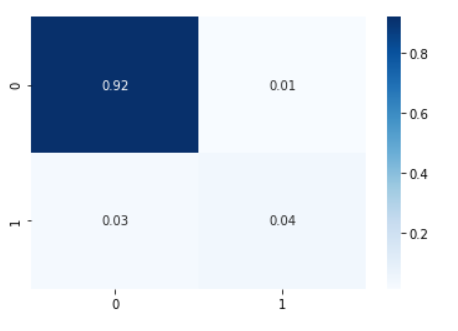
**precision recall f1-score**

**0 0.98 0.97 0.97**

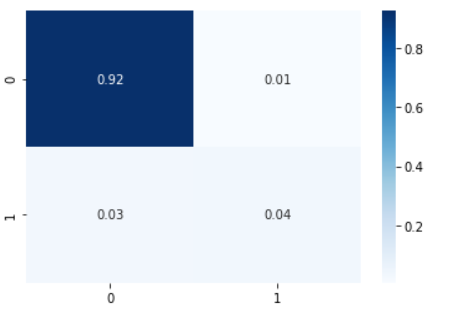
**1 0.62 0.69 0.65**

CONFUSION MATRIX FOR PASSIVE AGGRESSIVE





CONFUSION MATRIX FOR LOGISTIC REGRESSION



**precision recall f1-score**

**0 0.97 0.99 0.98**

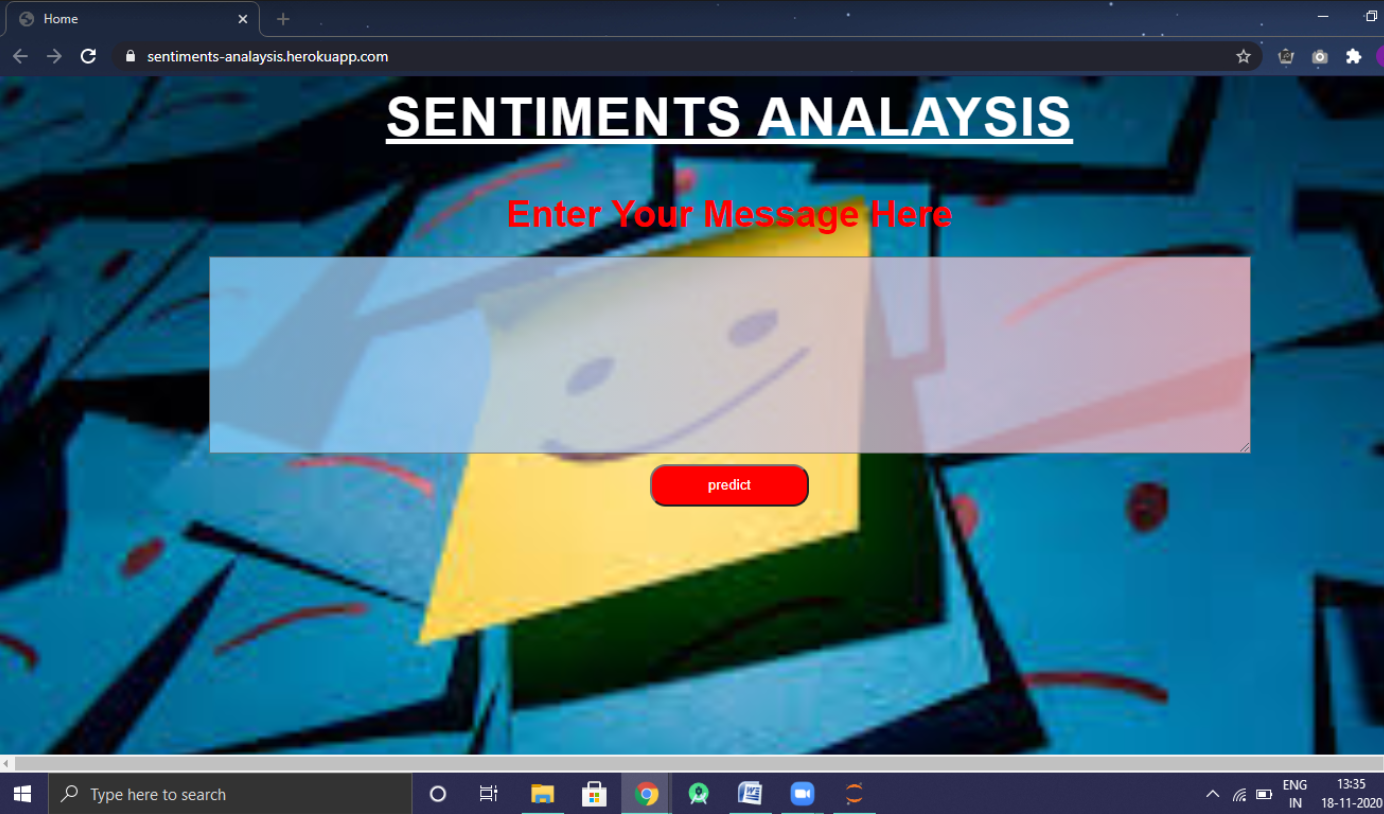
**1 0.84 0.53 0.65**

**CHAPTER 6: USER INTERFACE AND SOFTWARE**

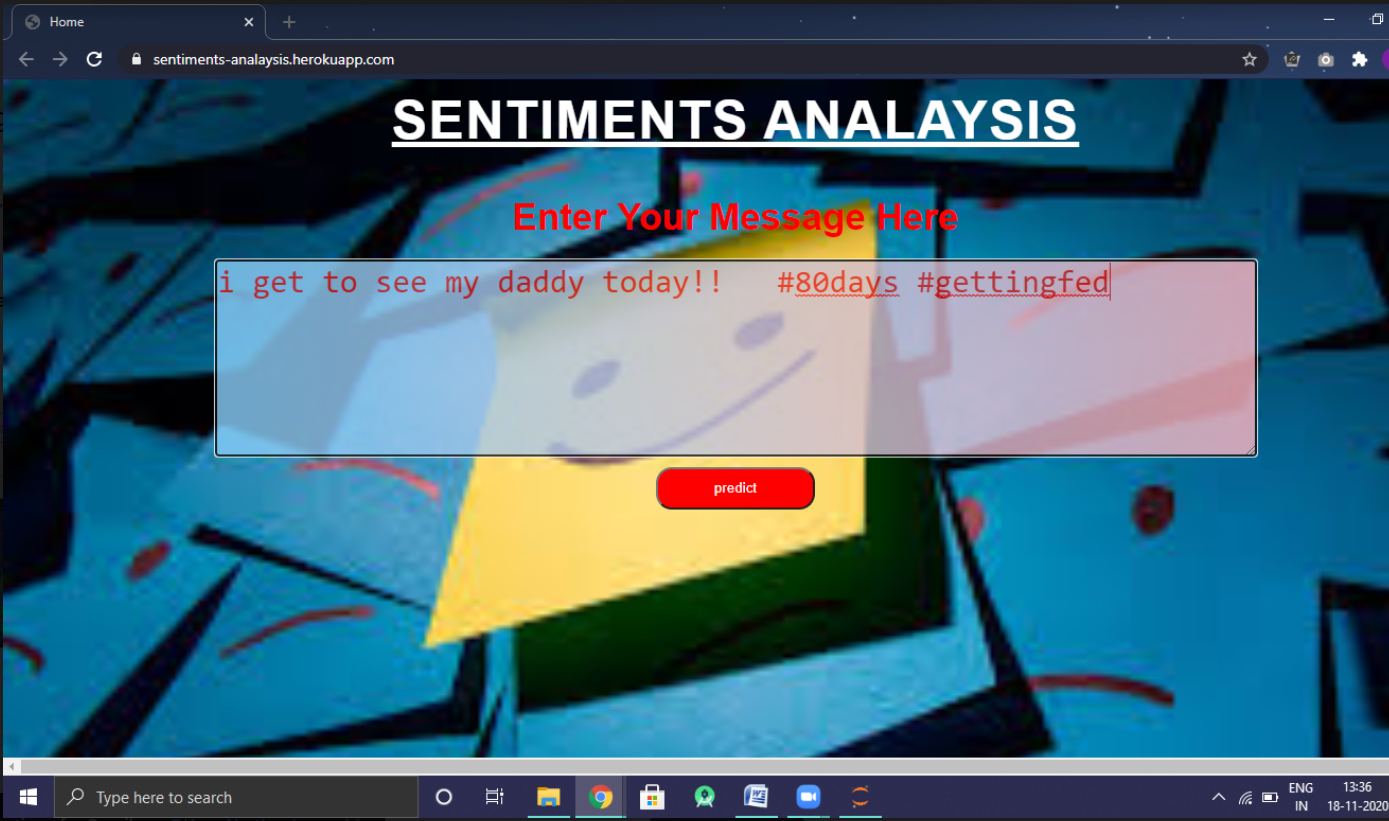
**TESTING**

After developing our machine learning model, training it with the available data-set and testing it with the test dataset we have received satisfactory accuracy

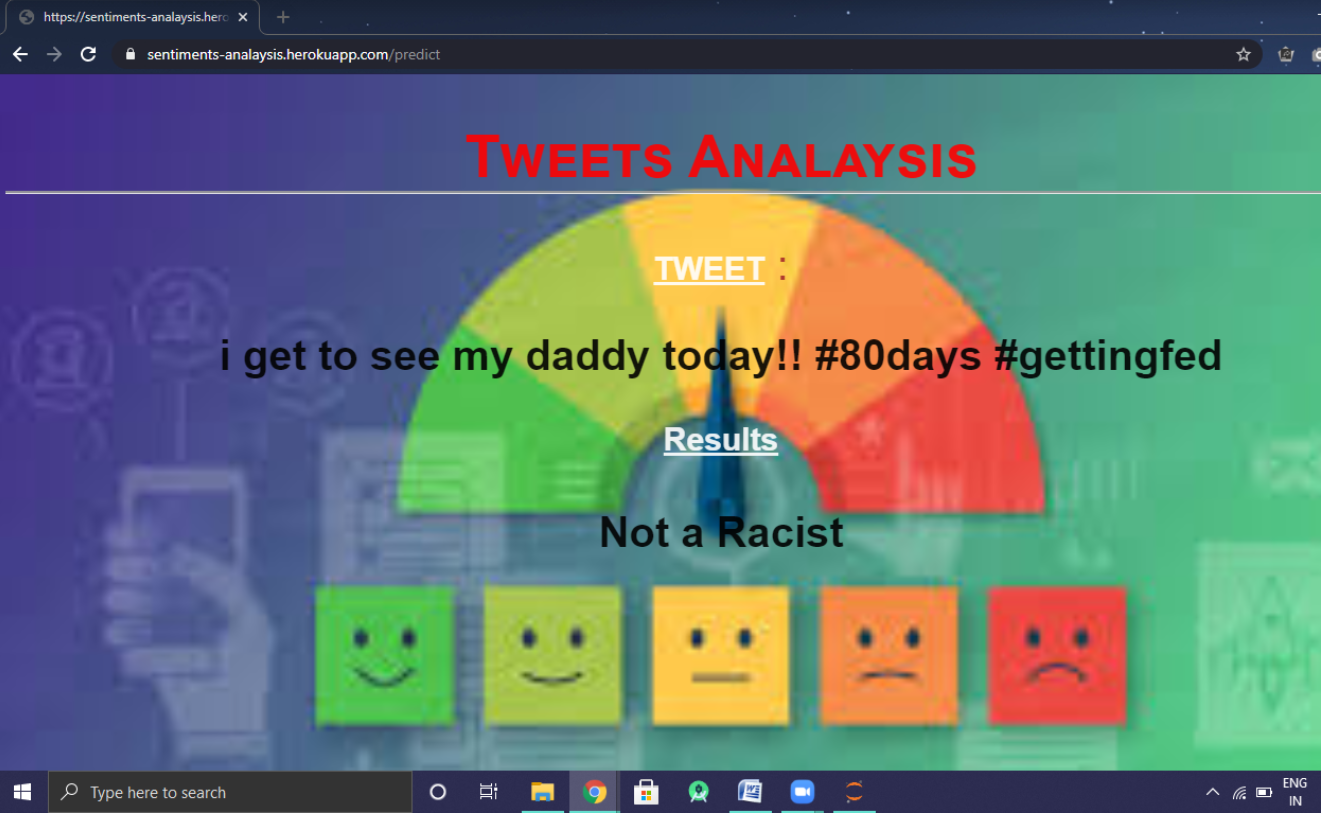
and so to convert the model to be implemented as a application we have used a flask framework to provide the user interface to the model. And deployed the model for use and testing.



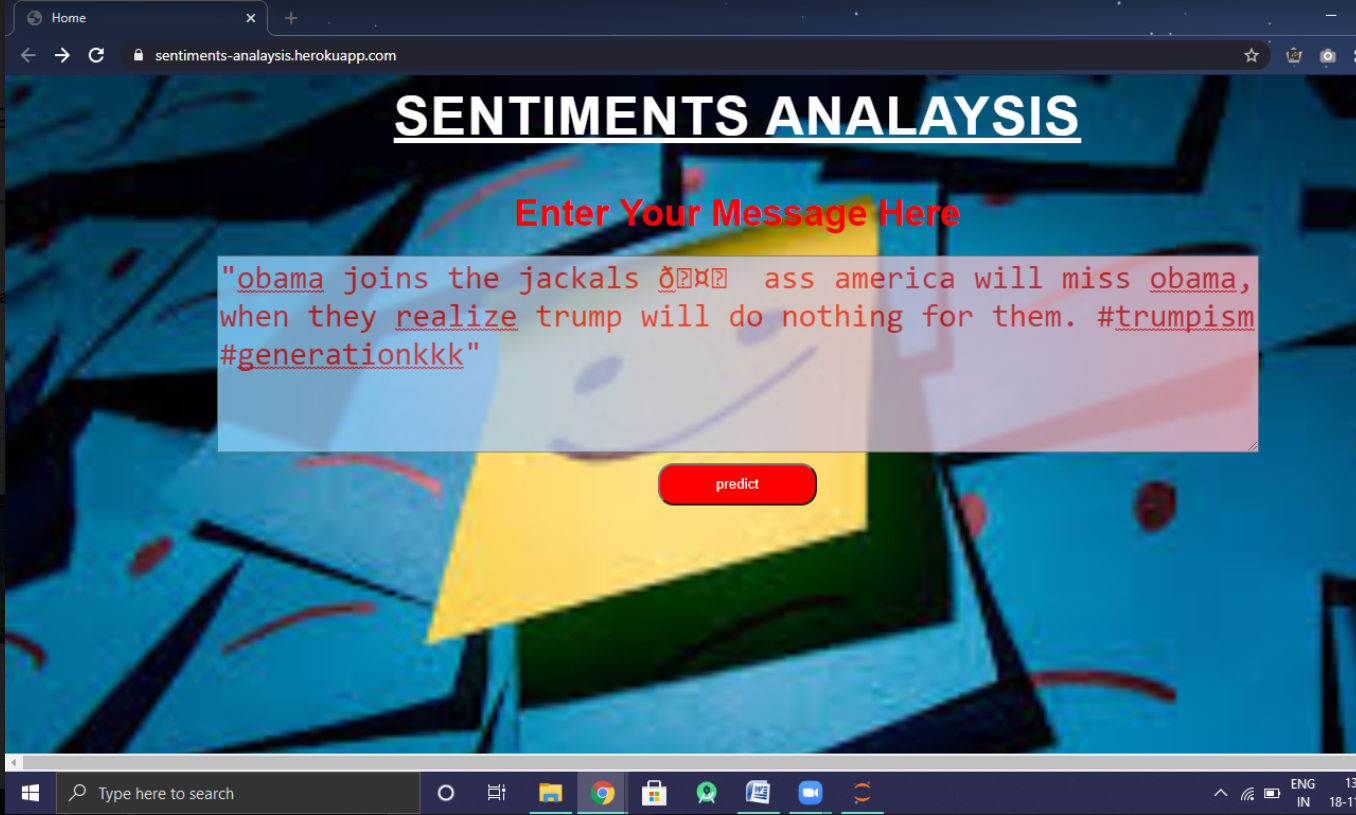
This is the first screen that appears at our deployed model and asks for the input which is to be tested.



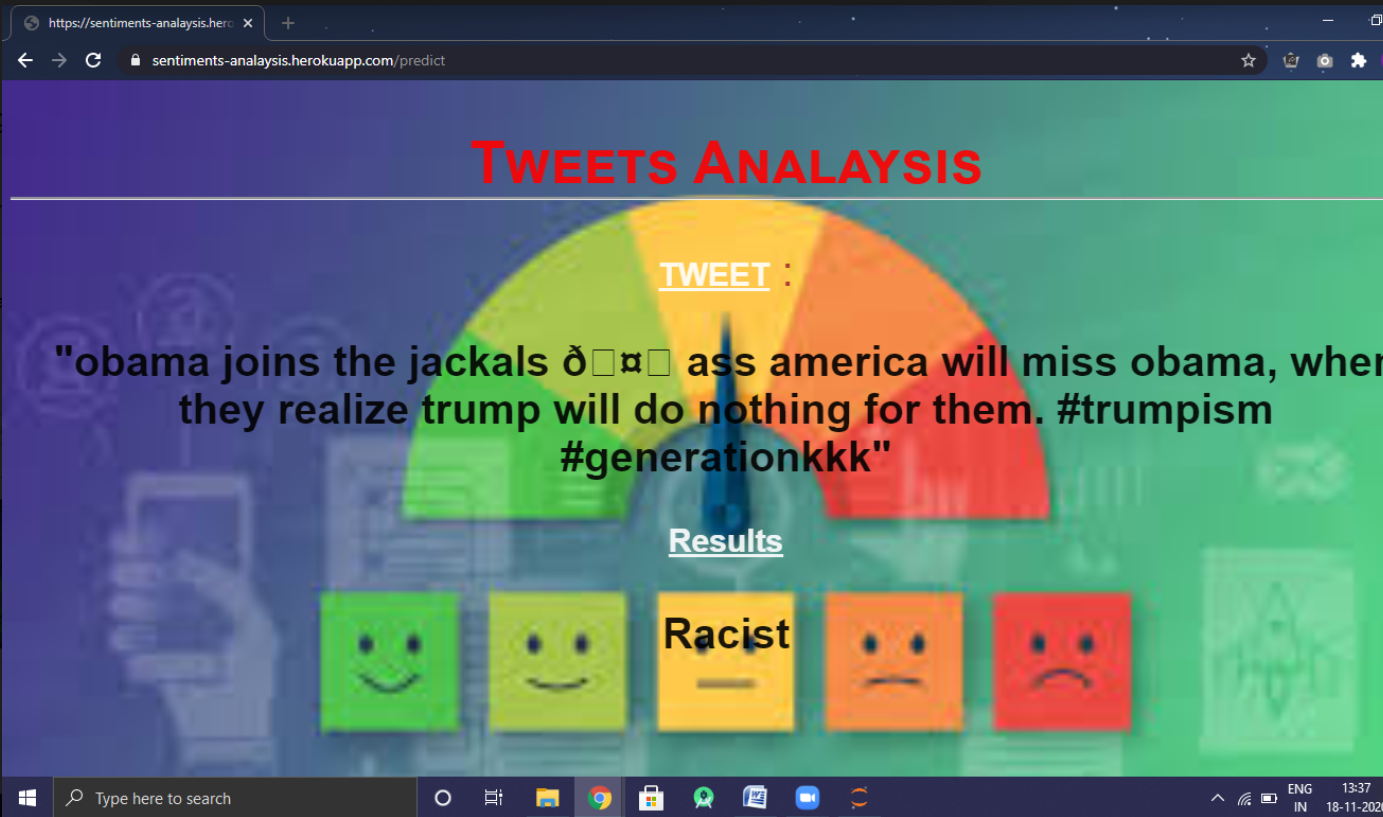
After we enter the tweet inside the box and click on the predict the result appears.



Another example : Tweet Entered



The result will be printed as:-



So this is how the Tweet Analyser works



**CONCLUSION**

Social media sentiment is the attitude and feelings people have about a brand on social media, about a particular event that happened or some incident or a current topic that matters and holds value in the real world. It adds context to all the @-mentions, comments, and shares.

So we have figured out where the people indulge in, stand on the positive/negative spectrum, we need to analyse the conversations of people in order to protect some other from various comments or thinking that might have a negative impact on the society we are living in. So in this project we have analysed the comments and conversation of one of the social media platforms Twitter and classify the tweets among two categories:- Racist and Non-Racist. So this application can be used by various other social media platforms to detect all the racist content and hence prevent the misunderstandings and the grudge that grows amoung the people against the government the individual ,stops bulling on social media and sometimes prevent the spoiling of the brand name as well.

***References***

* [https://hcis-journal.springeropen.com/articles/10.1186/s13673-017-0116-3#Sec1](https://hcis-journal.springeropen.com/articles/10.1186/s13673-017-0116-3%23Sec1)
* <https://www.researchgate.net/publication/336084873_Sentiment_Analysis_Techniques_for_Social_Media_Data_A_Review>
* <https://monkeylearn.com/blog/sentiment-analysis-machine-learning/>
* <https://www.tutorialspoint.com/machine_learning_with_python/classification_introduction.htm>