

CHAPTER 1

INTRODUCTION

Our project is a web application that predicts the likelihood/certainty that a customer will buy a product that he is interested in based on his social media posts such as Twitter tweets and user profile data. This will help the company/business target a particular customer more efficiently and boost their sales.

First, we search for Twitter tweets of potential customers wanting to buy a product. And based on those tweets we estimate/predict the likelihood that the customer will buy the product. We then make a model by gathering tweets from users who have already expressed intention to buy the product using their tweet history and if possible, their web search history as well and then training the text analytical model based on those tweets. Using the model, we input potential customers who have tweeted about the product but have not bought it. And based on the training data the model estimates a prediction/likelihood of whether the customer will buy it or not. We have limited the scope of our work to only mobile phones. Our model predicts the consumer intention for the latest upcoming mobile phones. We will be testing it on the latest iPhone X variants and check with its accuracies.

There have been several research studies for analyzing the insights of online consumers buying behavior. However, only a few have addressed the customers buying intention for products. We want to develop a machine learning approach that will identify potential customers for a product by estimating the purchase intention in measurable terms from tweets on twitter. We have used a text analytical machine learning approach because although text analytics can be performed manually, it is inefficient. By using text mining and natural language processing algorithms it will be much faster and efficient to find patterns and trends. In a way we can say that Purchase Intention detection task is close to the task of identifying wishes in product reviews.

1.1 OBJECTIVES

- To create Web application we need to create various Dashboards.
- To update or to test the data we need to have the annotated Data.
- The annotated data is the collect the twitter data in the form of csv file.
- based on the training data, the model will estimate a prediction/likelihood of whether the customer will buy it or not.
- We are using the following text analytical models:
 1. Support Vector Machine (SVM)
 2. Naive Bayes
 3. Logistic Regression
 4. Decision Tree
 5. Neural Network
- Positive reviews and comments generated by customers on social networking platforms can help improve business sales and profitability.
- This social media data retrieval system enables companies to reach out to new and existing information about their products.
- By erasing the physical and spatial boundaries between people, social networking websites can provide instant reachability.
- Organizations and businesses can use social networking to build a following and review their product globally.

1.2 PROBLEM STATEMENT

➤ Problem

Purchase intentions are frequently measured and used by marketing managers as an input for decisions about new and existing products and services. Up till now many companies still use customer survey forms in which they ask questions like how likely you are to buy a product in a given time frame and using that information they calculate the purchase intention. We want to see if we can use Twitter tweets to train a model to identify tweets which show purchase intention for a product.

➤ Complexity

The complexity of our approach is that we have to calculate how to measure the purchase intention from a tweet. Exploring the different type of text analytical methods and choosing the best one for our task will be quite challenging. Measuring the results of our machine learning model and then deciding the best one will involve a lot of factors which we will have to calculate.

➤ Motivation

We want to develop a machine learning model which can predict the numerical value for the consumer intention for a tweet. By doing this we can prove that we social media such as Twitter is also an important tool which marketers can use when deciding to target a customer. We believe that our work can be valuable to applications focusing on exploiting purchase intentions from social media.

➤ Challenges

The first challenge we faced was that we were not able to find any public dataset regarding purchase intention. We had to scrap the data from Twitter using a web scraper. Secondly, since we ourselves gathered the data we had to manually annotate the tweets. Again, this process was extremely time consuming as we had to go through each tweet and decide the purchase intention. Thirdly, we had limited annotated data because of the lengthy process of manual annotation and time constraint.

CHAPTER 2

LITERATURE REVIEW

[1] OUR INFERENCE OUT OF FIRST PAPER :

- Explains about Social platforms like Twitter and Facebook allow users to share and publish information with their users. In addition to this, users use them to answer very precise and highly contextualized queries, or queries for which the relevant content has not been authored yet, e.g., asking about a conference event using its hashtag on Twitter.
- Even if the profile is private, we can still derive some attributes. Also, when we tried as an external user to extract profiles from different online social network e.g. Twitter and Facebook , either no data or minimal data was made available
- Data extraction is a field that is concerned with grabbing information from different web resources including websites, online databases and services. It is necessary to find tools for data extraction because of the dynamic nature of the World Wide Web
- This creates some difficulties for end users and application programs when it comes to finding useful data.
- They are stricter in terms of having well-formed documents i.e. the documents' contents should conform to their syntax rules. This feature helps the parsers of search engines to interact with the web pages' contents more efficiently.
- The common machine learning algorithms that are used for text analysis are Linear Regression, Random Forest, Naive Bayes and Support Vector Machine. We will be looking at these models later in detail.
- There have been several research studies for analyzing the insights of online consumers buying behavior. However, only a few have addressed the customers buying intention for products. Studies on identification of wishes from texts, specifically Ramanand et al. (Ramanand, Bhavsar, and Pedanekar 2010) consider the task of identifying 'buy' wishes from product reviews

- These wishes include suggestions for a product or a desire to buy a product. They used linguistic rules to detect these two kinds of wishes. Although rule-based approaches for identifying the wishes are effective, but their coverage is not satisfactory, and they can't be extended easily. Purchase Intention detection task is close to the task of identifying wishes in product reviews.
- Here we don't use the rule-based approach, but we present a machine learning approach with generic features extracted from the tweets.
- Past studies have shown that it is possible to apply Natural Language Processing (NLP) and Named Entity Recognition (NER) to tweets. However, applying NER to tweets is very difficult because people often use abbreviations or (deliberate) misspelled words and grammatical errors in tweets.
- These studies merely analyze the sentiment of a tweet about a product after the author has bought it. We will however be extracting features from tweets to find whether the user has shown purchase intention towards the product or not.
- . The first studies used product or movie reviews because these reviews are either positive or negative.

[2] OUR INFERENCE OUT OF SECOND PAPER:

- The common machine learning algorithms that are used for text analysis are Linear Regression, Random Forest, Naive Bayes and Support Vector Machine. Applying Named Entity Recognition (NER) and Natural Language Processing (NLP) to tweets is very difficult because people often use abbreviations or deliberately misspelled words and grammatical errors in tweets.
- Some preprocessing techniques commonly used for twitter data are the sentiment140 API (Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter), the TweetNLP library (a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets), unigrams, bigrams and stemming.
- There are also some dictionary-based approaches such as using the textBlob library (TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent

API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more).

- Further looking at research reports like The Impact of Social Network Marketing on Consumer Purchase Intention in India: Consumer Engagement as a Mediator (Asian Journal of Business and Accounting give us an insight of the impact of social network marketing on consumer purchase intention and how it is affected by the mediating role of consumer engagement. Based on UGT theory (Uses and Gratification Theory).
- Social platforms like Twitter and Facebook allow users to share and publish information with their friends and often with general public. In addition to this, users use them to answer very precise and highly contextualized queries, or queries for which the relevant content has not been authored yet, e.g., asking about a conference event using its hashtag on Twitter. We refer to such a process of finding information as Social search, and we define it as follows.
- Social platforms like Twitter and Facebook allow users to share and publish information with their friends and often with general public. In addition to this, users use them to answer very precise and highly contextualized queries, or queries for which the relevant content has not been authored yet, e.g., asking about a conference event using its hashtag on Twitter. We refer to such a process of finding information as Social search, and we define it as follows.
- The social search is associated to platforms that are defined as search engines specifically dedicated to social data management such as Facebook. The main ingredient to perform a social search is the user interactions.
- Social platforms allow users to provide, publish and spread information, e.g., commenting or tweeting about an event. In such a context, a huge quantity of information is created in social media.
- They do not suppose any additional structure, including the social network that surrounds users. This does not reflect the real behaviors of users, since they normally ask friends for recommendations before acting

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

The popularity of online social networking sites has increased the amount of personal data which is distributed on the net. This is supported by the fact that social networking sites have overtaken email in terms of usage. Online social networking sites contain user profiles which consist of personal data. Those profiles are semi-structured and the profile data or structure may change in an unpredictable way.

This fits in nicely with the way online social networks operate. Social network profiles change all the time not just in structure but content as well. More research needs to be done into the extraction from semi-structured pages in terms of online social networking profiles. The motivation for this paper is that as far as the paper authors' know there has been little research associated with automated extraction methods from semi-structured web pages from online social networks.

Our goal is to extract the relevant profile data so it can be mined in the future to find attributes that can cause the profile owner to be vulnerable to social engineering attacks. In terms of online social networks our research will allow us in the future to investigate the friends of a profile and see if any of the friends have profiles on other online social networks.

This links into the transitivity concept where e.g. A and B are friends on one online social network but B and C are also friends on another online social network. The question is: will A and C become friends and if so how great will the strength of their friendship be. The question posed fits with theory about weak ties and how they can provide an alternative information source to the ones associated with the strong ties.

Data extraction is a field that is concerned with grabbing information from different web resources including websites, online databases and services. It is necessary to find tools for data extraction because of the dynamic nature of the World Wide Web. This creates some difficulties for end users

and application programs when it comes to finding useful data. Automated data extraction is just starting to be used in research about online social networking. Table 1 below illustrates some of the data extraction techniques used to extract attributes from online social networking profiles. It shows some data extraction methods ranging from manual through to automated methods. Our methodology outlines our approach to extracting attributes and a list of top friends from a MySpace profile.

MySpace was chosen because it allows a rich source of data to be derived from profiles without the need to be a member of MySpace. Even if the profile is private, we can still derive some attributes. Also, when we tried as an external user to extract profiles from different online social network e.g. Facebook and Friendster, either no data or minimal data was made available.

These collaborative tasks that make users more active in generating content are among the most important factors for the increasingly growing quantity of available data. In such a context, a crucial problem is to enable users to find relevant information with respect to their interests and needs.

Even if the profile is private, we can still derive some attributes. Also, when we tried as an external user to extract profiles from different online social network e.g. Facebook and Friendster, either no data or minimal data was made available.

Our process will comprise of the following components:

- Stage 1- Data Pre-processing involves analysis of the HTML structure of a given profile. The HTML content is parsed and a vector of tokens is produced. The extracted tokens help in the design of the tables in the repository. The reason of data pre-processing is because MySpace profiles have different structures and therefore different tokens.
- Stage 2- Specify the URL address of a profile. All social network profiles come with a unique profile URL address. The algorithm for extraction of the personal details involved developing and expanding the library provided.
- Stage 3- Check the user's stopping criteria. The user can specify whether they want to stop the extraction by the number of friends extracted (e.g. the first 100 friends) or by the level (e.g. 1 which is just the top friends of the specified profile) extracted

- Stage 5- Extract the relevant personal details from the profile and insert them into the repository. The repository used PostgreSQL 8.1.4. The repository structure has to be designed for the Breadth First Search algorithm.
- Stage 6- Extract list of friends and their profile addresses then insert them into the repository if they have not been stored before. In the case of this research paper, the extracted friends' lists just consist of the top friends who we assumed the user may have a strong affiliation with. The data in the repository can be used for data mining purposes in the future to find patterns.

3.2 PROPOSED SYSTEM

Proposed Approach

In this section, we describe the details of our approach to tackle the problem of purchase intention detection. We will begin by describing our data collection and annotation process. Then we will describe our approach for data preprocessing and transforming the data to train text analytical models.

Data collection and annotation

As there are no annotated Twitter tweets corpora available publicly for detection of purchase intent, we had to create our own. This was done using a web crawler developed by JohnBakerFish which crawled the website to collect the data. We had collected over 100,000 tweets but since they were not annotated, we had to cut down to just 3200 tweets which were randomly selected out of the dataset and we manually annotated them. We used just 3200 tweets out of such a large dataset as we were limited by time.

We defined definition of Purchase Intention as object that is having action word like (buy, want, desire) associated with it. Each tweet was read by 3 people and final class was decided by maximum voting.

Criteria for Labelling of tweets

| | Tweet | Class |
|---|---|-------|
| 1 | Comparing iphone x with other phone and telling other phone are better? | No PI |
| 2 | Talking about good features of iphone x? | PI |
| 3 | Talking about negative features of iphone x? | No PI |
| 4 | liked video on Youtube about iphone x? | PI |

3.2.1 DATA PREPROCESSING

Next, we preprocessed the tweets using these techniques:

1. **LOWERCASE:** So, we started our groundwork by converting our text into lower case, to get case uniformity.
2. **REMOVE PUNC:** Then we passed that lower case text to punctuations and special characters removal function. Text may contain unwanted special characters, spaces, tabs and etcetera which has no significant use in text classification.
3. **STOPWORDS REMOVAL:** Text also contains useless words which are routine part of the sentence and grammar but do not contribute to the meaning of the sentence. Likes of “the”, “a”, “an”, “in” and etcetera are the words mentioned above. So, we do not need these words, and it is better to remove these.
4. **COMMON WORD REMOVAL:** Then there also lots of repetitive words which from their recurrence do not contribute to the meaning in the sentence. This can also be the result of mistake as the data we are analyzing is an informal data where formal sentence norms are not taken into consideration.
5. **RARE WORDS REMOVAL:** We also removed some rare words like names, brand words (not iphone x), left out html tags etc. These are unique words which do not contribute much to interpretation in the model.
6. **SPELLING CORRECTION:** Social media data is full of spelling mistakes. And it is our job to get rid of these mistakes and give our model the correct word as an input.
7. **STEMMING:** Then we stemmed the words to their root. Stemming works like by cutting the end or beginning of the word, considering the common prefixes or suffixes that can be found in that word. For our purpose, we used Porters Stemmer, which is available with NLTK.
8. **LEMMATIZATION:** Then we also performed lemmatization on our text. This analysis is performed in morphological order. A word is traced back to its lemma, and lemma is returned as the output.

After preprocessing the tweets, we are left with about 1300 tweets for training data and remaining for testing.

3.2.2 ALGORITHMS

Once the corpus was ready, we then used different text analytical models to test which one gave the best results. We used the following models:

1. Support Vector Machine (SVM):

Simply put, SVM is a supervised machine learning algorithm which does complex transformation on the data. And then it tries to separate data on classes we have defined on our data.

2. Naive Bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

3. Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

4. Decision Tree:

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

5. Neural Network:

It is deep learning machine algorithm, which is arranged in a layer of neuron. There is an input layer, output layer and hidden layers of neurons. Neuron network is adaptive as neurons in these layers learn from their initial input and subsequent runs.

3.2.3 DOCUMENT VECTOR

1. **TF:** First is the term frequency document vector. We have stored text and its labeled class in data frame. And we have constructed a new data frame with columns as the words and document count as the rows. So, individual frequency of words in a document count is recorded.
2. **IDF:** It is a weighting method to retrieve information from the document. Term frequency and inverse document frequency scores calculated and then product of $TF * IDF$ is called TF-IDF. IDF is important in finding how relevant a word is. Normally words like 'is', 'the', 'and' etc. have greater TF. So IDF calculated a weight to tell how important least occurring words are.
3. **TF-IDF with textblob library:** With the help of textblob library we calculated sentiments of individual word and then multiplied the sentiment score with TF and TF-IDF of that word.

3.2.4 EVALUATION

To evaluate our models, we using the following techniques:

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall
5. F-Measure
6. True Negative Rate

Further we have also considered The True Positive Rate and the shape of the ROC curve for more insights.

CHAPTER 4

DESIGN

4.1 SYSTEM ARCHITECTURE

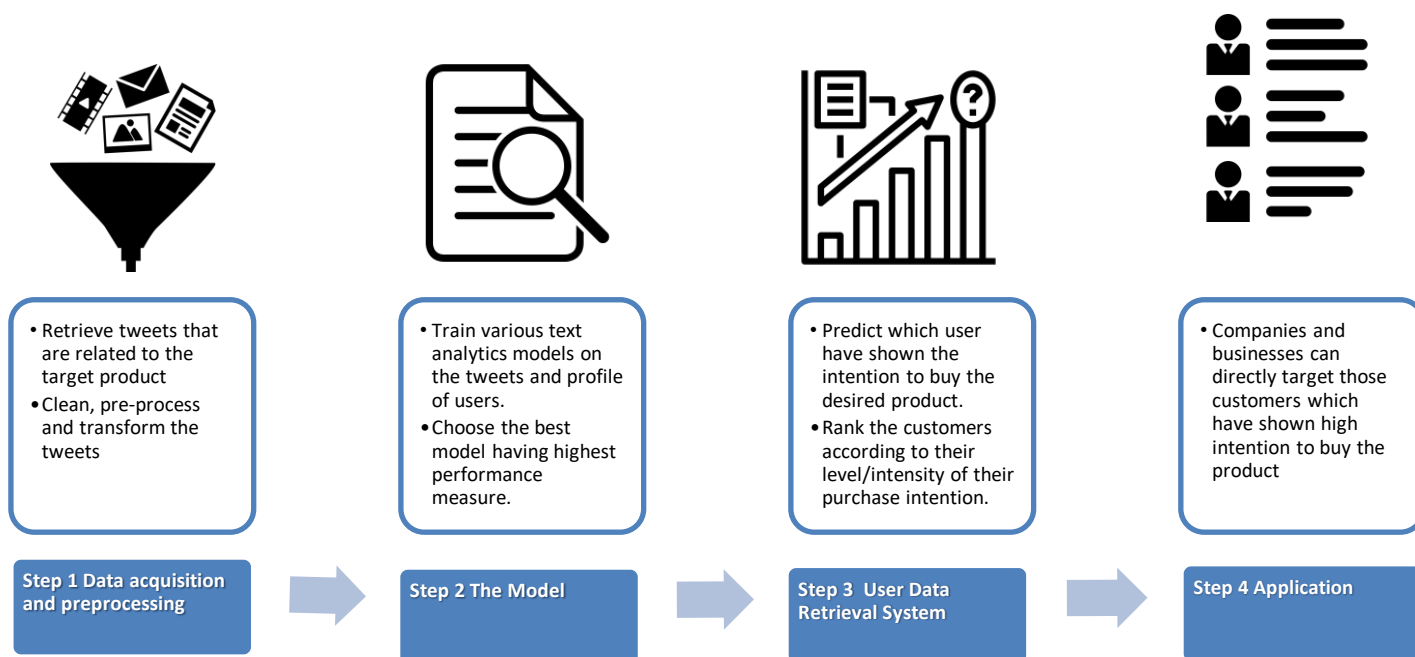


Fig 4.1 : Steps involved in User Interest Based Social Media Data Retrieval System

1. Firstly, the user will open the DASHBOARD and see the status of the product through charts and the relevant tweets for the product in a list.
2. Secondly, the user will press the UPLOAD ANNOTATED DATASET button through which he will be asked to upload the dataset he wants to test for his product and find the relevant customers from that dataset. The output will be in a form of a pie chart and a table showing the different level of customers and the scores assigned to each customer.

- Thirdly, when the user presses the ANALYSIS button, he will be taken to the screen which will show the detailed analysis of our dataset that we have built containing word clouds, positive vs negative tweets and the most used words for that product.

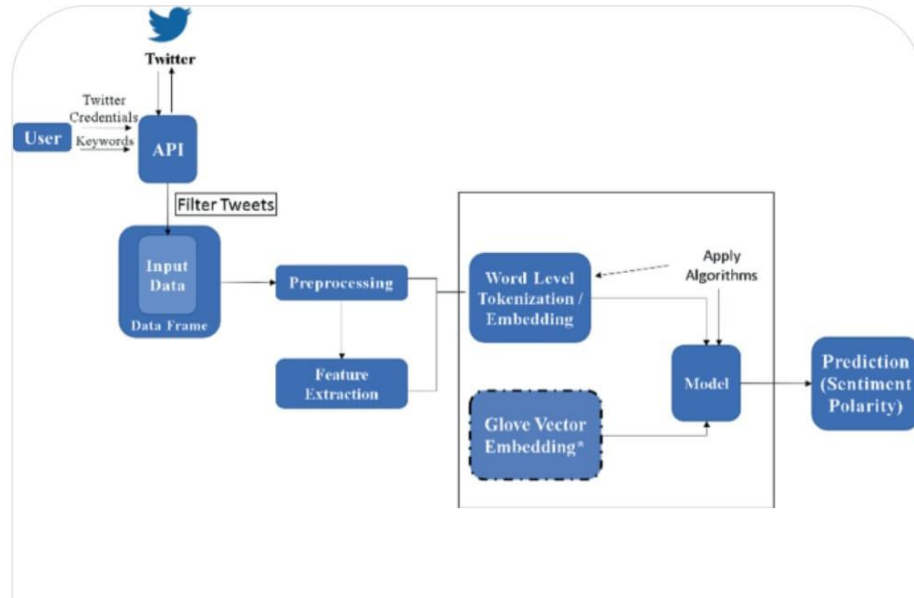


Fig 4.2 Flow Diagram 1

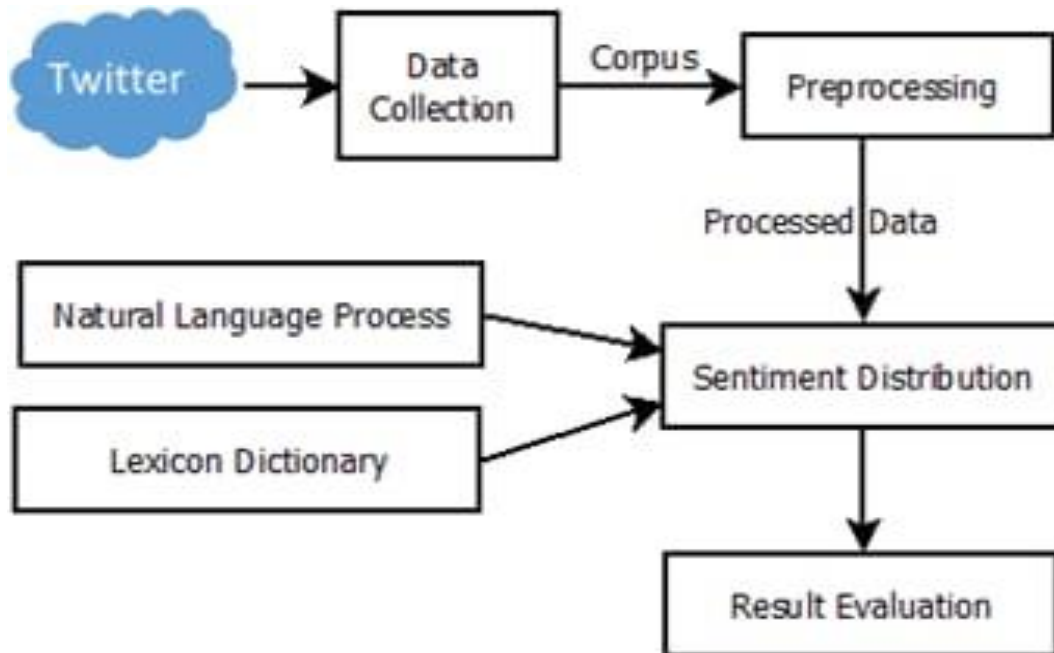


Fig 4.3 Flow Diagram 2

4.2 MACHINE LEARNING

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.

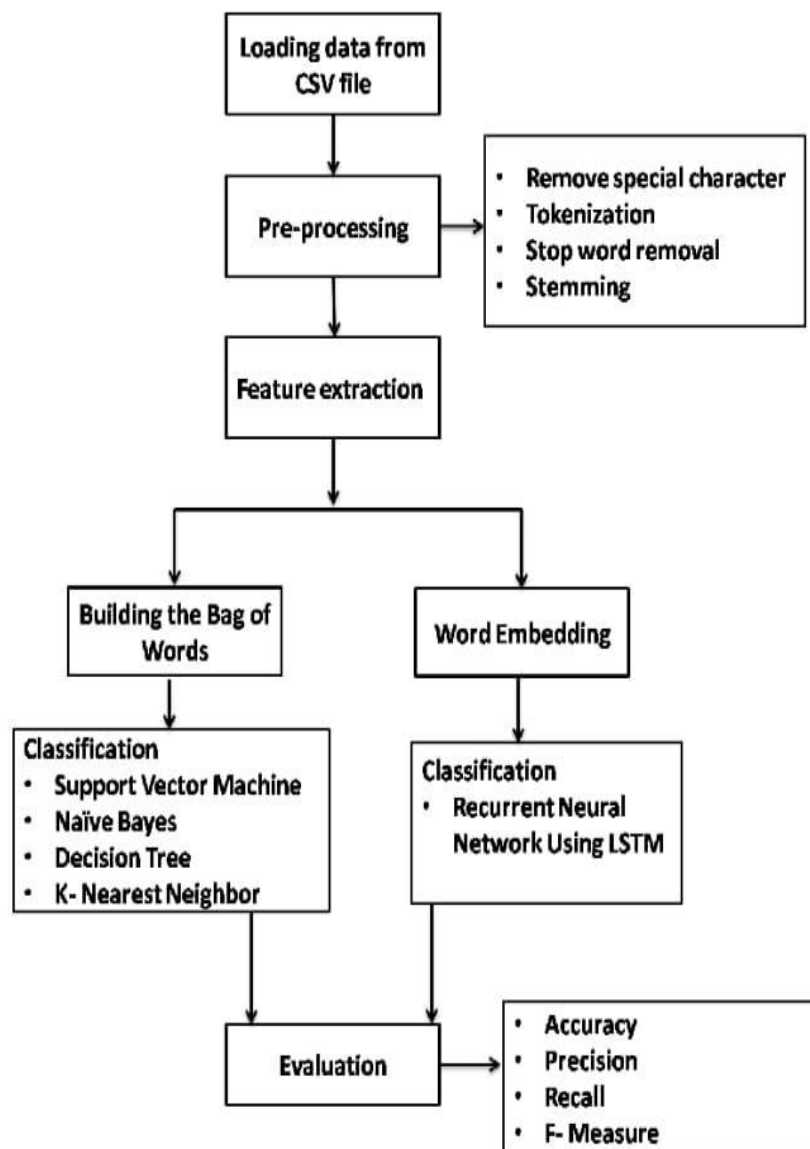


Fig 4.4 Sequential Diagram

➤ **Supervised Learning**

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

➤ **Unsupervised learning**

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar to how a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabelled and uncategorized data which make unsupervised learning more important.

➤ **Reinforcement learning**

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

4.3 ALGORITHMS

4.3.1 SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i$$

subject to the constraints:

$$y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \dots, N$$

where C is the capacity constant, w is the vector of coefficients, b is a constant, and ξ_i represents parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ represents the class labels and x_i represents the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In the 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

Types of SVM:

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where the number of dimensions is greater than the number of samples. ● Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

The followings are important concepts in SVM –

- **Support Vectors** - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.
- **Hyperplane** - As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

- **Margin** - It may be defined as the gap between two lines on the closest data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

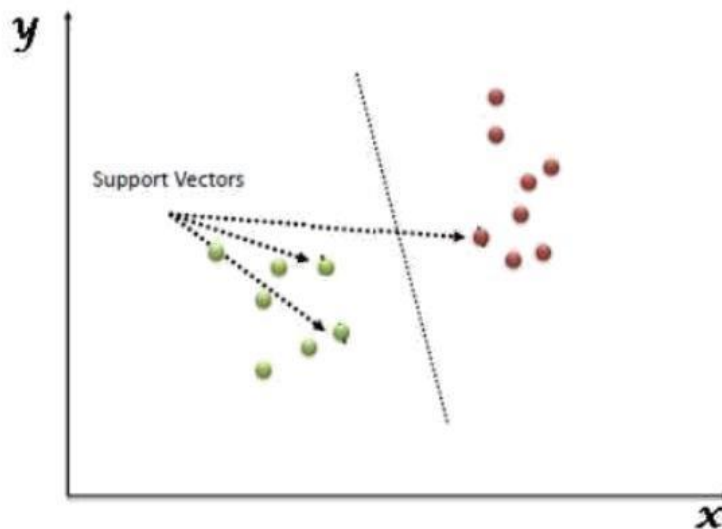


Figure: Support Vector Machine

Fig 4.5 SVM Algorithm

4.3.2 NAIVE BAYES ALGORITHM:

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

The Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. The Naive Bayes algorithm is comprised of two words Naive and Bayes, which can be described as:

- **Naive:** It is called Naive because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes:** It is called Bayes because it depends on the principle of Bayes' Theorem.

Types of Naive Bayes model:

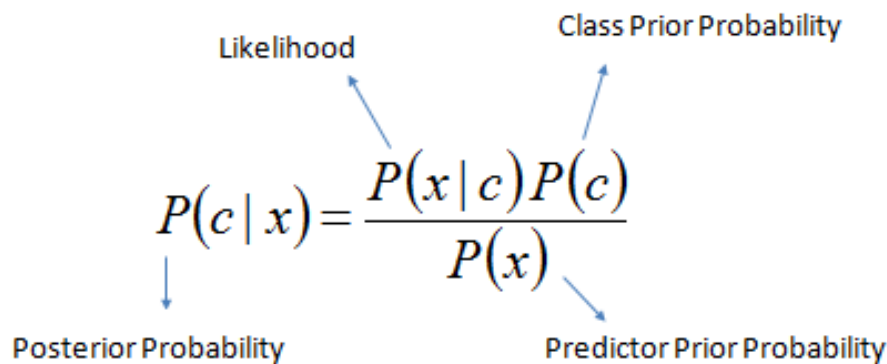
There are three types of Naive Bayes Model, which are given below:

- **Gaussian:** The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
- **Multinomial:** The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems; it means a particular document belongs to which category such as Sports, Politics, education, etc. The classifier uses the frequency of words for the predictors.
- **Bernoulli:** The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

Bayes' theorem:

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:



The diagram shows the Bayes' theorem formula with labels pointing to its components:

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

- Likelihood** points to $P(x | c)$
- Class Prior Probability** points to $P(c)$
- Posterior Probability** points to $P(c | x)$
- Predictor Prior Probability** points to $P(x)$

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

- $P(c|x)$ is the posterior probability of class (target) given predictor (attribute).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

4.3.3 LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between

0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S"-shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).

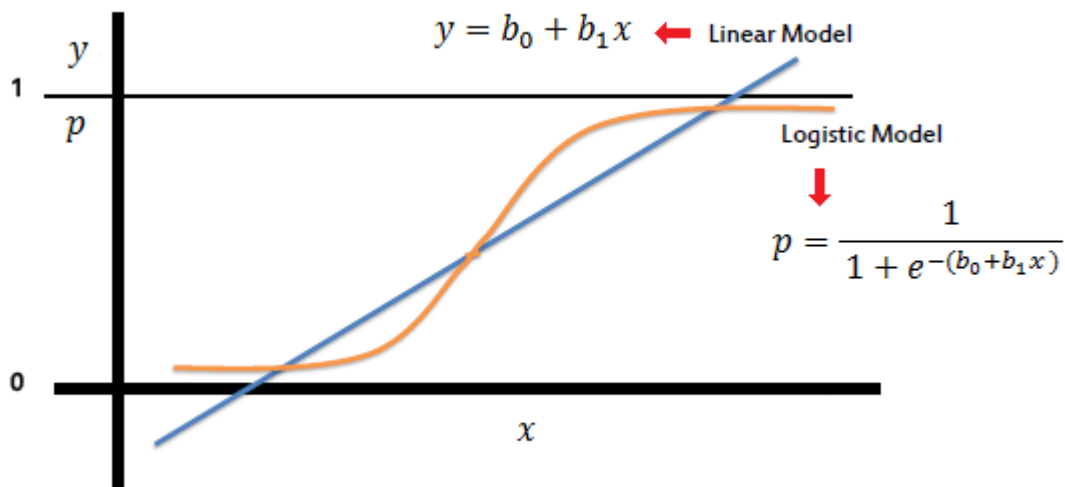


Fig 4.6 Logistic Regression algorithm

In the logistic regression the constant (b_0) moves the curve left and right and the slope (b_1) defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio.

$$\frac{p}{1-p} = \exp(b_0 + b_1 x)$$

Finally, taking the natural log of both sides, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors. The coefficient (b_1) is the amount the logit (log-odds) changes with a one unit change in x .

As mentioned before, logistic regression can handle any number of numerical and/or categorical variables.

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x$$
$$p = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p)}}$$

Advantages:

Logistic Regression is one of the simplest machine learning algorithms and is easy to implement yet provides great training efficiency in some cases. Also due to these reasons, training a model with this algorithm doesn't require high computation power. The predicted parameters (trained weights) give inference about the importance of each feature. The direction of association i.e., positive or negative is also given. So, we can use Logistic Regression to find out the relationship between the features.

This algorithm allows models to be updated easily to reflect new data, unlike Decision Tree or Support Vector Machine. The update can be done using stochastic gradient descent. Logistic Regression outputs well-calibrated probabilities along with classification results. This is an advantage over models that only give the final classification as results. If a training example has a 95% probability for a class, and another has a 55% probability for the same class, we get an inference about which training examples are more accurate for the formulated problem.

Disadvantages:

Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions

on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So, on high dimensional datasets, Regularization techniques should be considered to avoid overfitting (but this makes the model complex). Very high regularization factors may even lead to the model being under-fit on the training data.

Non-linear problems can't be solved with logistic regression since it has a linear decision surface. Linearly separable data is rarely found in real world scenarios. So, the transformation of non-linear features is required which can be done by increasing the number of features such that the data becomes linearly separable in higher dimensions.

Non-Linearly Separable Data: It is difficult to capture complex relationships using logistic regression. More powerful and complex algorithms such as Neural Networks can easily outperform this algorithm.

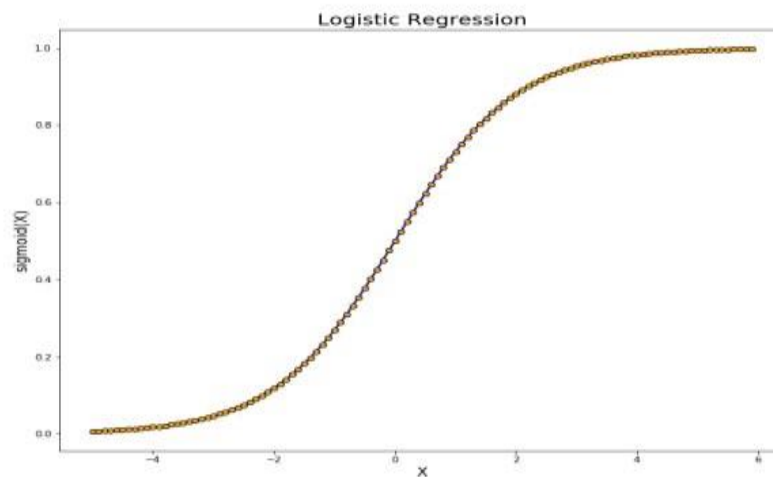


Figure: Logistic Regression

Fig 4.7 Logistic Regression Algorithm

4.3.4 DECISION TREE ALGORITHM

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node.

Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a Decision Tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A Decision Tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. The Decision Tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for a regression problem.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

- a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

$$E(I, X) = \sum_{c \in X} P(c) E(c)$$

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model.

Below are the two reasons for using the Decision Tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a treelike structure.

In Decision Tree the major challenge is to identify the attribute for the root node in each level.

This process is known as attribute selection. We have two popular attribute selection measures:

1. **Information Gain:** When we use a node in a Decision Tree to partition the training instances into smaller subsets, the entropy changes. Information gain is a measure of this change in entropy. Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy the more the information content.
2. **Gini Index:** Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower Gini index should be preferred. Sklearn supports “Gini” criteria for Gini Index and by default, it takes “gini” value.

The most notable types of Decision Tree algorithms are: -

- **IDichotomiser 3 (ID3):** This algorithm uses Information Gain to decide which attribute is to be used to classify the current subset of the data. For each level of the tree, information gain is calculated for the remaining data recursively.
- **C4.5:** This algorithm is the successor of the ID3 algorithm. This algorithm uses either Information gain or Gain ratio to decide upon the classifying attribute. It is a direct improvement from the ID3 algorithm as it can handle both continuous and missing attribute values.
- **Classification and Regression Tree (CART):** It is a dynamic learning algorithm which can produce a regression tree as well as a classification tree depending upon the dependent variable.

Working:

In a Decision Tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the Decision Tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of dataset created in step-3. Continue this process until a stage is reached where you cannot further classify the nodes and call the final node as a leaf node.

4.3.5 NEURAL NETWORK

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature.

Backpropagation algorithms are a family of methods used to efficiently train artificial neural networks (ANNs) following a gradient descent approach that exploits the chain rule. The main feature of backpropagation is its iterative, recursive and efficient method for calculating the weights updates to improve in the network until it is able to perform the task for which it is being trained.

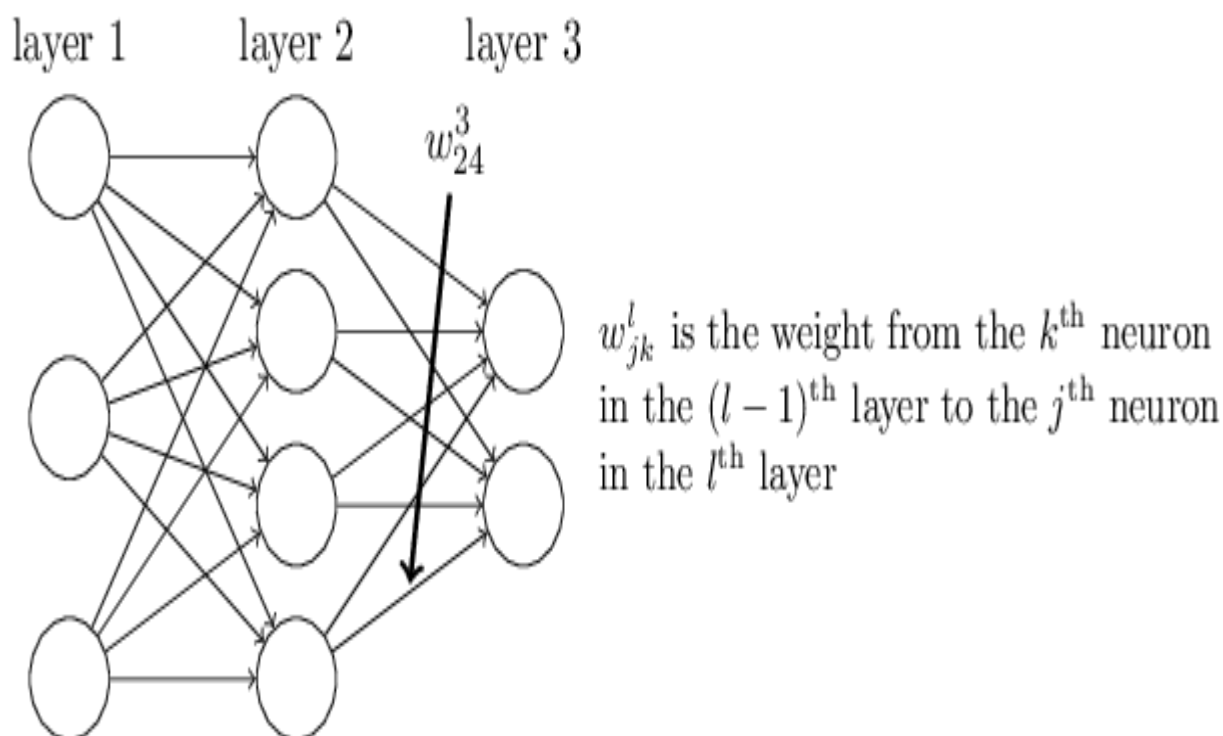


Fig 4.8 Neural Networks Algorithm

TYPES OF NEURAL NETWORKS

➤ **Feed-Forward Neural Networks**

Feed-forward neural networks are one of the more simple types of neural networks. It conveys information in one direction through input nodes; this information continues to be processed in this single direction until it reaches the output mode. Feed-forward neural networks may have hidden layers for functionality, and this type of most often used for facial recognition technologies.

➤ **Recurrent Neural Networks**

A more complex type of neural network, recurrent neural networks take the output of a processing node and transmit the information back into the network. This results in theoretical "learning" and improvement of the network. Each node stores historical processes, and these historical processes are reused in the future during processing.

➤ **Convolutional Neural Networks**

Convolutional neural networks, also called ConvNets or CNNs, have several layers in which data is sorted into categories. These networks have an input layer, an output layer, and a hidden multitude of convolutional layers in between. The layers create feature maps that record areas of an image that are broken down further until they generate valuable outputs. These layers can be pooled or entirely connected, and these networks are especially beneficial for image recognition applications.

➤ **Deconvolutional Neural Networks**

Deconvolutional neural networks simply work in reverse of convolutional neural networks. The application of the network is to detect items that might have been recognized as important under a convolutional neural network. These items would likely have been discarded during the

convolutional neural network execution process. This type of neural network is also widely used for image analysis or processing.

➤ **Modular Neural Networks**

Modular neural networks contain several networks that work independently from one another. These networks do not interact with each other during an analysis process. Instead, these processes are done to allow complex, elaborate computing processes to be done more efficiently. Similar to other modular industries such as modular real estate, the goal of the network independence is to have each module responsible for a particular part of an overall bigger picture.

ADVANTAGES OF NEURAL NETWORKS

- Neural networks that can work continuously and are more efficient than humans or simpler analytical models. Neural networks can also be programmed to learn from prior outputs to determine future outcomes based on the similarity to prior inputs.
- Neural networks that leverage cloud of online services also have the benefit of risk mitigation compared to systems that rely on local technology hardware.
- Last, neural networks are continually being expanded into new applications. While early, theoretical neural networks were very limited to its applicability into different fields, neural networks today are leveraged in medicine, science, finance, agriculture, or security.

DISADVANTAGES OF NEURAL NETWORKS

- Though neural networks may rely on online platforms, there is still a hardware component that is required to create the neural network. This creates a physical risk of the network that relies on complex systems, set-up requirements, and potential physical maintenance.
- Though the complexity of neural networks is a strength, this may mean it takes months (if not longer) to develop a specific algorithm for a specific task. In addition, it may be difficult to spot any errors or deficiencies in the process, especially if the results are estimates or theoretical ranges.

CHAPTER 5

SYSTEM OVERVIEW

We aim to analyze the tweets related to a product and identify the purchase intention in it. In this way we can rank the tweets which have high purchase intention and report the name of the person who tweeted as potential customer of product.

We will make a model by gathering tweets from users who have already expressed intention to buy the product and see their tweet history and if possible, their web search history as well. Using this model, we will input potential customers who have tweeted about the product but have not bought it yet! And based on the training data the model will estimate a prediction/likelihood of whether the customer will buy it or not.

5.1 BACKGROUND AND MOTIVATION

- Currently we have many recommendation systems available which recommend different products to the user, most of which are not efficient. No such effective model for businesses to identify potential customers.
- We want to develop a software that will help the businesses identify potential customers for their products by estimating their purchase intention in measurable terms from their tweets and user profile data on twitter.

5.2 PROJECT GOAL

- We aim to analyze the tweets related to a product and identify the purchase intention in it. In this way we can rank the tweets which have high purchase intention and report the name of the person who tweeted as potential customer of product.
- We will make a model by gathering tweets from users who have already expressed intention to buy the product and see their tweet history and if possible, their web search history as well. Using this model, we will input potential customers who have tweeted about the product but have not bought it. And based on the training data the model will estimate a prediction/likelihood of whether the customer will buy it or not.

5.3 PROJECT REQUIREMENTS

We require data from twitter to analyze purchase intention. For now, we are gathering the data by scraping the tweets of a product through a scraper. We will need a mechanism to save the scrapped tweets in storage for further processing and performing analysis and for that we have decided to use mongo db. We will need to annotate data retrieved from scraper. Once model is trained, we will need to develop a website so that users can easily access our application through the graphical interface of the website. We will show on the website the purchase intention rank of tweets on level of 1 to 5 for the desired product of the user.

1. Functional Requirements

- Website : Website will be required so that users can easily access our application.
- Dashboard : The user will open the dashboard and see the status of the product through charts and the relevant tweets for the product in a list.
- Get Relevant Users : The user will press the Get Relevant Users button through which he will be taken to the next screen which will show the list of users and tweets who have shown intention to buy the product.
- Find Purchase Intention of The Users : The user will then press The Find Purchase Intention Of The Users button through which he will be taken to the final screen which will show the results in a pie chart and list format for all the users who should be targeted and will also show the result in level form meaning that the users will be categorized according the level of intention they have expressed to buy the product.

2. Objectives

- We aim to analyze the tweets related to a product and identify the purchase intention in it.
- We want to rank the tweets according to the level of purchase intention and show a list of potential customers to the user which he can use to directly target the customer.

3. Constraints

- We are scrapping the data not getting data through twitter Search API. This is sometimes not reliable and robust, so we will need to either get the data from some company which sells such data or will need access to the twitter search API.

4. Validation and Acceptance Tests:

- We will test the model accuracy by confusion matrix (Accuracy, Precision, Recall, F-Measure). This will give us a percentage of accuracy achieved by our model.

For our application development, we are going to run unit testing in which individual units of source code, would be tested to determine if they are fit to use. Then, we are going to run integration testing where the individual source codes would be merged and tested as group.

5.4 ERRORS

- Initially we encountered some errors in running the python environment and setting up a separate virtual environment for our project. However, we were soon able to overcome this by reading up on the documentation and following tutorials.
- We also encountered some errors running the Django framework. Setting up a sperate virtual environment and installing the missing dependencies later fixed this problem.
- We also faced some issues in integrating the python code with the website because it was our first time configuring a back-end server to run code scripts and sending the output to the html page and getting it to display correctly.

5.5 TRAILS AND TESTS

- We have tested the model accuracy by confusion matrix (Accuracy, Precision, Recall, F-Measure). Further we have also considered the True Negative Rate, The True Positive Rate and the shape of the ROC curve for more insights. This will give us a percentage of accuracy achieved by our model.

- For our application development, we opted to use unit testing in which individual units of source code, were tested to determine if they are fit to use. Then, we ran integration testing where the individual source was merged and tested as group.
- We also tested the usability of our website by carrying out the tasks and functions of the website in different scenarios and checking if they successfully completed.

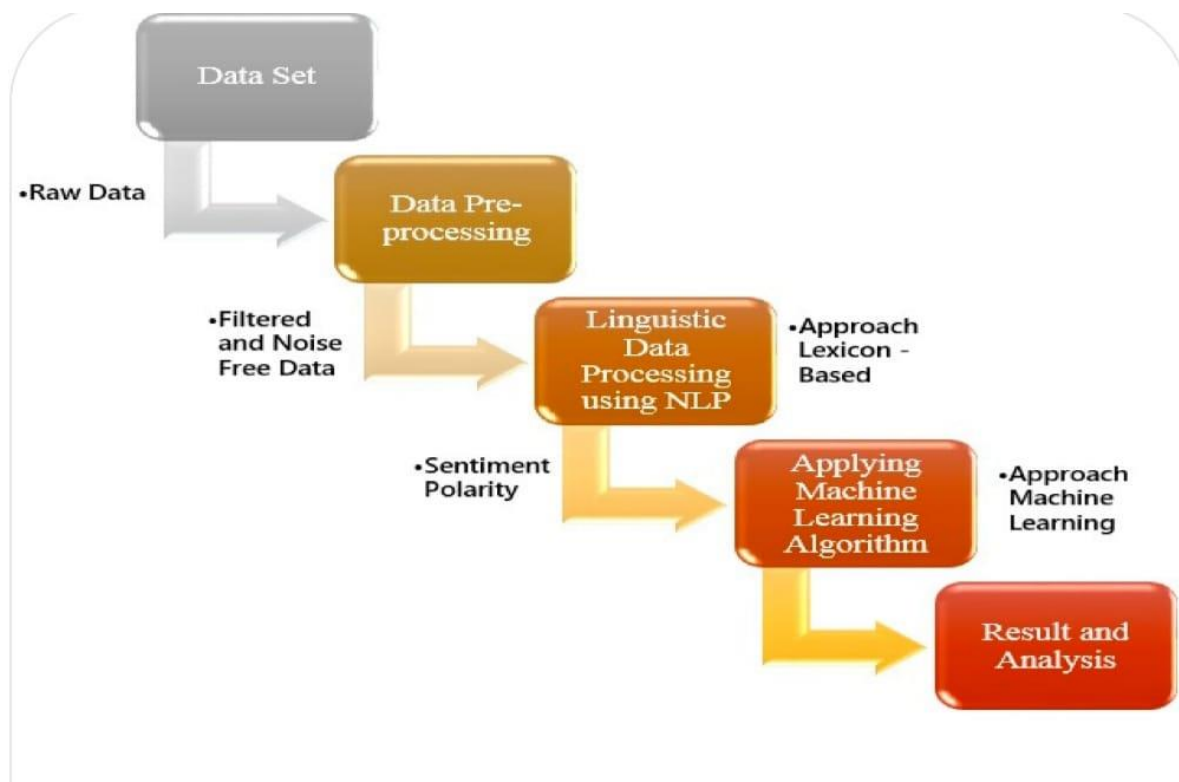


Fig 5.1 Data set Processing

5.6 ADVANTAGES OF USER INTEREST BASED SOCIAL MEDIA DATA RETRIVAL SYSTEM

- Consumer Intention Prediction using Twitter offers several advantages that can benefit businesses and marketers in understanding and targeting their audience. Some of the key advantages include:
- Real-Time Insights: Twitter provides a vast amount of real-time data generated by users worldwide. By analyzing this data, businesses can gain immediate insights into consumer opinions, trends, and intentions. Real-time insights enable businesses to respond quickly to emerging opportunities or challenges, adapt their strategies, and stay ahead of the competition.
- Large and Diverse User Base: Twitter has a diverse user base that spans across different demographics, industries, and geographical locations. This diversity provides businesses with a rich source of data to analyze and understand consumer intentions across various segments. The broad reach of Twitter allows businesses to capture insights from a wide range of consumers and tailor their marketing efforts accordingly.
- Unfiltered Consumer Opinions: Twitter offers a platform where users openly express their opinions, feedback, and intentions. By monitoring and analyzing these conversations, businesses can gain unfiltered insights into consumer preferences, needs, and behaviors. This unfiltered nature of Twitter data can provide businesses with a deeper understanding of their audience, helping them refine their products, services, and marketing strategies.
- Trend Identification: Twitter is known for its ability to drive and amplify trends. By monitoring trending topics, hashtags, and conversations, businesses can identify emerging trends and consumer intentions in real time. This enables businesses to capitalize on these trends, adapt their marketing campaigns, and stay relevant in a rapidly changing landscape.
- Influencer Engagement: Twitter is home to many influential individuals and thought leaders who shape opinions and trends. By analyzing consumer interactions with influencers, businesses can understand the impact of these influencers on consumer intentions. This knowledge can help businesses identify potential collaborations, partnerships, or influencer

marketing strategies to effectively reach their target audience and influence consumer behavior.

- Targeted Advertising Opportunities: Twitter offers robust advertising capabilities that allow businesses to target specific audiences based on demographics, interests, and behaviors. By leveraging consumer intention prediction on Twitter, businesses can optimize their ad targeting, ensuring that their marketing messages reach the right audience at the right time. This can lead to higher engagement, conversions, and return on investment (ROI) for advertising campaigns.
- In summary, Consumer Intention Prediction using Twitter provides businesses with real-time insights, a diverse user base, unfiltered consumer opinions, trend identification, influencer engagement, and targeted advertising opportunities. Leveraging these advantages can help businesses better understand their audience, tailor their marketing efforts, and drive business success.

CHAPTER 6

EXPERIMENTAL ANALYSIS

6.1 HARDWARE REQUIREMENTS

- Processor : > i3
- Ram : 4GB.
- Hard Disk : 500 GB.
- Input device : Standard Keyboard and Mouse.
- Compact Disk : 650 Mb.
- Output device : High Resolution Monitor.

6.2 SOFTWARE REQUIREMENTS

- Operating system : macOS, Windows XP/7 or higher version.
- Coding language : Python 3.6.
- Framework : Python Django Framework.
- Storage : Mongo DB. (if needed)
- Libraries : scikit-learn library for Python.
- Internet browser : Google Chrome.
- Code Editor : visual studio code.

6.3 TOOLS AND TECHNOLOGY DETAILS

6.3.1 PYTHON



Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

6.3.2 VISUAL STUDIO CODE



Visual Studio is an integrated development environment (IDE) created by Microsoft. It provides developers with a comprehensive set of tools and features to build software applications for various platforms, including Windows, macOS, Android, and iOS.

Visual Studio supports multiple programming languages, such as C#, C++, JavaScript, Python, and more, allowing developers to work on a wide range of projects. It offers a rich set of code editing and debugging capabilities, including intelligent code completion, syntax highlighting, and advanced debugging tools. The IDE includes a powerful built-in compiler and a wide range of libraries and frameworks to facilitate application development. It also provides integrated version control, project management, and collaboration features to enhance team productivity.

Visual Studio offers a customizable and extensible environment, allowing developers to tailor their workflow to their specific needs. It supports extensions and plugins from the Visual Studio Marketplace, enabling developers to add additional functionality and tools to the IDE. Overall, Visual Studio is a versatile and feature-rich IDE that streamlines the development process and empowers developers to create high-quality software applications efficiently. Visual Studio is a comprehensive integrated development environment (IDE) developed by Microsoft. It is widely used by software developers to build applications for various platforms, including Windows, macOS, Android, and iOS.

6.3.3 DJANGO FRAMEWORK



Django is a Python framework that makes it easier to create web sites using Python. Django takes care of the difficult stuff so that you can concentrate on building your web applications. Django emphasizes reusability of components, also referred to as DRY (Don't Repeat Yourself), and comes with ready-to-use features like login system, database connection and CRUD operations (Create Read Update Delete).

Django follows the MVT design pattern (Model View Template).

- Model - The data you want to present, usually data from a database.
- View - A request handler that returns the relevant template and content - based on the request from the user.
- Template - A text file (like an HTML file) containing the layout of the web page, with logic on how to display the data.

Model

The model provides data from the database. In Django, the data is delivered as an Object Relational Mapping (ORM), which is a technique designed to make it easier to work with databases. The most common way to extract data from a database is SQL. One problem with SQL is that you have to have a pretty good understanding of the database structure to be able to work with it. Django, with ORM, makes it easier to communicate with the database, without having to write complex SQL statements.

- The models are usually located in a file called models.py.

View

A view is a function or method that takes http requests as arguments, imports the relevant model(s), and finds out what data to send to the template, and returns the final result.

- The views are usually located in a file called `views.py`.

Template

A template is a file where you describe how the result should be represented. Templates are often `.html` files, with HTML code describing the layout of a web page, but it can also be in other file formats to present other results, but we will concentrate on `.html` files. Django uses standard HTML to describe the layout,

- The templates of an application is located in a folder named `templates`.

URLs

Django also provides a way to navigate around the different pages in a website. When a user requests a URL, Django decides which *view* it will send it to.

- This is done in a file called `urls.py`.

6.3.4 MACHINE LEARNING TECHNIQUES

Machine learning techniques refer to a variety of algorithms and approaches used to train computer systems to automatically learn patterns and make predictions or decisions based on data. Some common machine learning techniques include:

- **Supervised Learning:** This technique involves training a model using labeled data, where the desired output is known. The model learns to map inputs to outputs and can make predictions on new, unseen data.

- **Unsupervised Learning:** In unsupervised learning, the model is trained on unlabeled data without any specific desired output. The goal is to discover patterns or structures in the data, such as clustering similar data points or finding hidden correlations.
- **Reinforcement Learning:** This technique involves training an agent to interact with an environment and learn from feedback in the form of rewards or penalties. The agent learns to take actions that maximize cumulative rewards, making it suitable for tasks requiring sequential decision-making.
- **Deep Learning:** Deep learning is a subfield of machine learning that uses artificial neural networks with multiple layers to learn hierarchical representations of data. It has been particularly successful in tasks such as image and speech recognition.
- **Transfer Learning:** Transfer learning involves leveraging knowledge learned from one task or domain to improve performance on another related task or domain. It allows models to generalize better and requires less training data.
- **Ensemble Learning:** Ensemble learning combines predictions from multiple individual models to make more accurate predictions. Techniques like bagging (bootstrap aggregating) and boosting are commonly used in ensemble learning.
- **Dimensionality Reduction:** Dimensionality reduction techniques aim to reduce the number of variables or features in a dataset while preserving important information. Principal Component Analysis (PCA) and t-SNE are examples of dimensionality reduction techniques.

6.3.5 DATA ANALYSIS

Data analysis is the process of analyzing the raw data so that the processed/analyzed data can be used in a system or a method/process. It majorly involves three steps data acquisition, data preprocessing and exploratory data analysis. Data acquisition is collecting the data from various sources like agencies, etc. for further analysis. While acquiring the data it is important to collect data which is relevant to the system or the process. Data preprocessing is a methodology in data mining that is used to convert the raw data into meaningful and efficient format. Many unrelated and may be present in the results. Software cleaning is done to tackle the portion. This includes managing details which are incomplete, noisy information etc. and hence the process of data preprocessing is performed.

6.4 RESULTS

6.4.1 FORMATION OF DOCUMENT VECTOR

We made 3 types of document vectors for the purpose of experimentation. First, is the term frequency document vector. We have stored text and its labeled class in data frame. And we have constructed a new data frame with columns as the words and document count as the rows. So, individual frequency of words in a document count is recorded. Second, is the inverse document frequency vector which is a weighting method to retrieve information from the document. Term frequency and inverse document frequency scores calculated and then product of $TF \times IDF$ is called TF-IDF. IDF is important in finding how relevant a word is. Normally words like 'is', 'the', 'and' etc. have greater TF. So IDF calculated a weight to tell how important least occurring words are. Lastly, we also used the textblob library to help create the document vector. With the help of textblob library we calculated sentiments of individual word and then multiplied the sentiment score with TF and TF-IDF of that word.

6.4.2 MODELLING

At this stage, the data preparation was complete, and we were ready to build our model. As discussed above we chose these 5 text analytical algorithms; Support Vector Machine, Naïve Bayes, Logistic Regression, Decision Tree and Artificial Neural Network, because they are the most used by researchers in this field.

To split our dataset for training and testing we first used the simple split of 70-30. However, since our dataset was limited, and we also had an imbalance class problem we also used the k-fold technique with $k=5$.

1. For the first algorithm, the multinomial Naive Bayes classifier, we configured it as follows:
 - Used Laplace smoothing for features not present in the learning samples to prevent zero probabilities in testing data.
 - Also considered prior probability of the features rather than using a uniform prior probability.
2. For the next algorithm, the Support Vector Machine classifier, we configured it as follows:
 - The algorithm we used was the linear SVM.
 - The penalty of an error was set to 1.
 - Considered probability estimates.
3. The next algorithm we used was Logistic Regression with the following configuration:
 - The inverse of regularization strength coefficient was set to 1 for stronger regularization.
 - Maximum number of iterations to converge was set to 100.

- For optimization we used the liblinear algorithm as it is best suited for small datasets.
4. We also tested the Decision Tree classifier with the following configuration:
 - The function to measure the quality of a split was ‘gini’
 - At least 7 samples were required to split an internal node as this was giving the highest accuracy.
 5. Finally, we also used the Artificial Neural Network algorithm with the following configurations:
 - ‘Relu’, the rectified linear unit function was used as the activation function for the hidden layer.
 - ‘lbfgs’, an optimizer in the family of quasi-Newton methods, was the method used as the solver for weight optimization because for small datasets, ‘lbfgs’ can converge faster and perform better.
 - The learning rate schedule for weight updates was kept to constant.
 - The hidden layers were kept as follows 50, 20, 10, 5.
 - The input layer was the number of features.
 - The output layer were the 2 classes.

Once the models were configured, we used the training data to train our models and then test our data. The results are discussed in the next section.

6.4.3 EXPERIMENTATION AND RESULTS

We built our models based on the training dataset and then experimented with the testing dataset on the models. To evaluate our models, we used the following techniques based on the Confusion Matrix (A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known):

1. Accuracy: $(TP + TN) / (TP + TN + FP + FN)$
2. Precision: $TP / (TP + FP)$
3. Recall: $TP / (TP + FN)$
4. F-Measure: $(2 * Precision * Recall) / (Precision + Recall)$
5. True Negative Rate: $TN / (TN + FN)$ (for imbalance class analysis)

Since we are using a machine learning (artificial intelligence) based approach we needed to set an accuracy standard for our model and evaluate the results by matching the desired standard.

To evaluate our models, we used the following techniques:

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall
5. F-Measure

Further we have also considered the True Negative Rate, The True Positive Rate and the shape of the ROC curve for more insights.

After evaluating our model here are the following results that we have gotten:

For our first attempt this is the results that we got:

| Accuracy Table | | | | | | |
|------------------------|-------------|--------------------|-----------------------|---------------|--------------------------|--------------------|
| | Naive Bayes | Logistic Regressio | Support Vector Machin | Decision Tree | Artificial Neural Networ | Naive Bayes CUSTOM |
| TF | 78.2 | 80.2 | 80.5 | 69.3 | 76 | 79.4 |
| TF-IDF | 65.6 | 78.2 | 78.2 | 72.3 | 77.6 | 76.7 |
| binary doc | 77.5 | 80.8 | 80.2 | 72.6 | 78.9 | 79.4 |
| text-blob + TF | | 79.5 | 78.5 | 66 | 75.2 | 72.7 |
| text-blob + TF-IDF | | 78.9 | 76.9 | 69.6 | 75.6 | 70.75 |
| text-blob + binary doc | | 79.5 | 78.5 | 72.3 | 79.2 | 73.12 |

| True Negative Rate | | | | | | |
|------------------------|-------------|--------------------|-----------------------|---------------|--------------------------|--------------------|
| | Naive Bayes | Logistic Regressio | Support Vector Machin | Decision Tree | Artificial Neural Networ | Naive Bayes CUSTOM |
| TF | 32.8 | 29.7 | 42.2 | 45.3 | 43.8 | 46 |
| TF-IDF | 48.4 | 37.5 | 48.4 | 46.9 | 46.9 | 52 |
| binary doc | 28.1 | 32.8 | 45.3 | 48.4 | 46.9 | 46 |
| text-blob + TF | | 31.2 | 39.5 | 54.7 | 40.6 | 48 |
| text-blob + TF-IDF | | 40.6 | 43.7 | 51.6 | 50 | 54 |
| text-blob + binary doc | | 31.2 | 39 | 48.4 | 32.8 | 50 |

| Precision | | | | | | |
|------------------------|-------------|---------------------|-----------------------|---------------|--------------------------|--------------------|
| | Naive Bayes | Logistic Regression | Support Vector Machin | Decision Tree | Artificial Neural Networ | Naive Bayes CUSTOM |
| TF | 83.4 | 83.2 | 85.4 | 83.8 | 84.9 | 86.8 |
| TF-IDF | 83.5 | 84.2 | 86.2 | 84.7 | 85.8 | 87.5 |
| binary doc | 82.5 | 83.8 | 85.9 | 85.1 | 86 | 86.8 |
| text-blob + TF | | 83.4 | 83.9 | 85 | 84.2 | 86 |
| text-blob + TF-IDF | | 84.8 | 85 | 85.2 | 86 | 86.85 |
| text-blob + binary doc | | 83.4 | 84.5 | 85 | 83.6 | 86.48 |

| Recall | | | | | | |
|------------------------|-------------|---------------------|-----------------------|---------------|--------------------------|--------------------|
| | Naive Bayes | Logistic Regression | Support Vector Machin | Decision Tree | Artificial Neural Networ | Naive Bayes CUSTOM |
| TF | 90.3 | 93.7 | 90.8 | 75.7 | 84.5 | 87.7 |
| TF-IDF | 70.3 | 89.1 | 86.2 | 79.1 | 85.8 | 82.8 |
| binary doc | 90.7 | 93.7 | 89.5 | 79.1 | 87.5 | 87.7 |
| text-blob + TF | | 92.5 | 89.9 | 69 | 84.5 | 78.8 |
| text-blob + TF-IDF | | 89.1 | 85.8 | 74.5 | 82.4 | 74.87 |
| text-blob + binary doc | | 92.4 | 89.1 | 78.6 | 91.6 | 78.81 |

For our second attempt after reorganizing the data preprocessing steps and adding some customized steps specific to our data, we got these results:

| Accuracy Table | | | | | |
|---|-------------|---------------------|------------------------|---------------|---------------------------|
| | Naive Bayes | Logistic Regression | Support Vector Machine | Decision Tree | Artificial Neural Network |
| TF + neg handling + kfold | 75.2 | 76.9 | 74 | 69 | 74.2 |
| TF-IDF + neg handling + kfold | 70.2 | 74.4 | 77.7 | 70.4 | 67.8 |
| TF + neg handling + lemmatization + kfold | 75.4 | 77.4 | 74.4 | 70.9 | 72.7 |
| TF-IDF + neg handling + lemmatization + kfold | 69.6 | 72.8 | 75.9 | 70.4 | 73.7 |
| TF + lemmatization | 75.6 | 76.9 | 73.6 | 73.6 | 71.3 |
| TF-IDF + lemmatization | 73.9 | 74.2 | 79.2 | 69.3 | 73.6 |

| True Negative Rate | | | | | |
|---|-------------|---------------------|------------------------|---------------|---------------------------|
| | Naive Bayes | Logistic Regression | Support Vector Machine | Decision Tree | Artificial Neural Network |
| TF + neg handling + kfold | 45.6 | 47 | 48.6 | 48.6 | 51 |
| TF-IDF + neg handling + kfold | 11.4 | 26.9 | 49.1 | 46.2 | 0 |
| TF + neg handling + lemmatization + kfold | 43.3 | 47.6 | 48.3 | 51.3 | 51 |
| TF-IDF + neg handling + lemmatization + kfold | 11.4 | 24.9 | 46 | 52.7 | 49.3 |
| TF + lemmatization | 49.4 | 46 | 47.1 | 57.5 | 51.7 |
| TF-IDF + lemmatization | 13.8 | 24.1 | 46 | 47.1 | 52.9 |

6.5 SNAPSHOTS

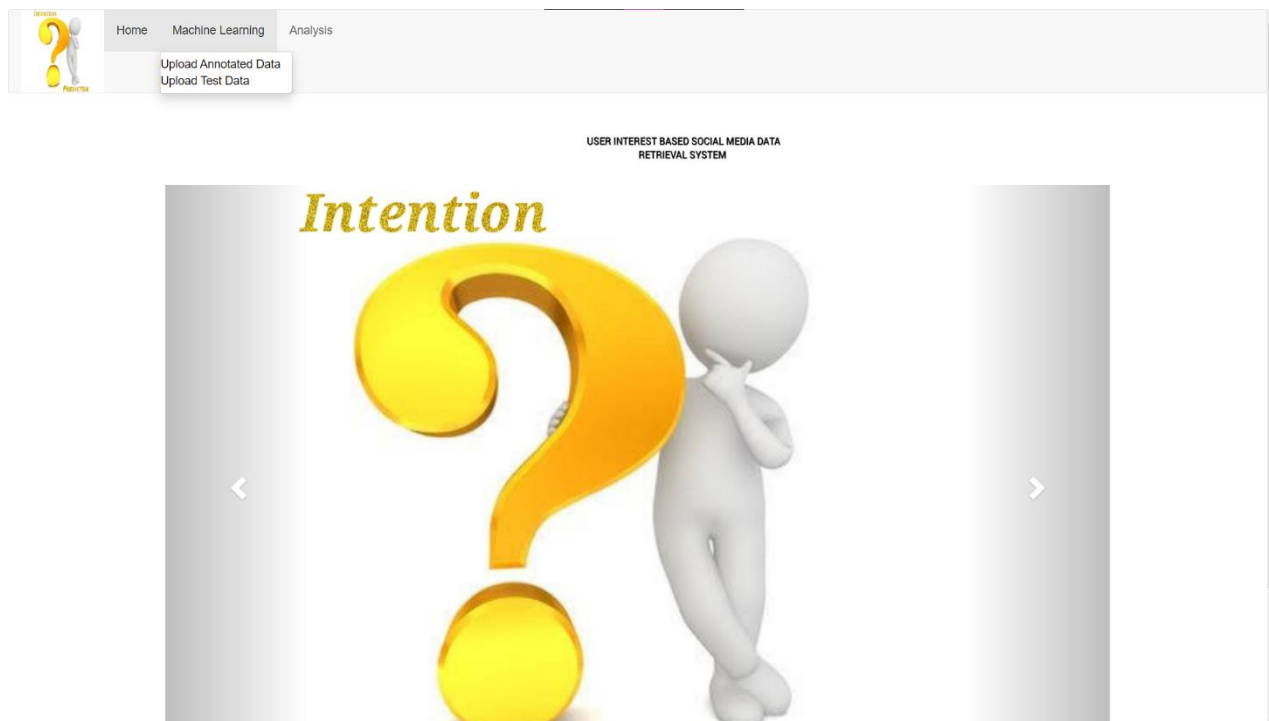


Fig 6.1 : home page 1

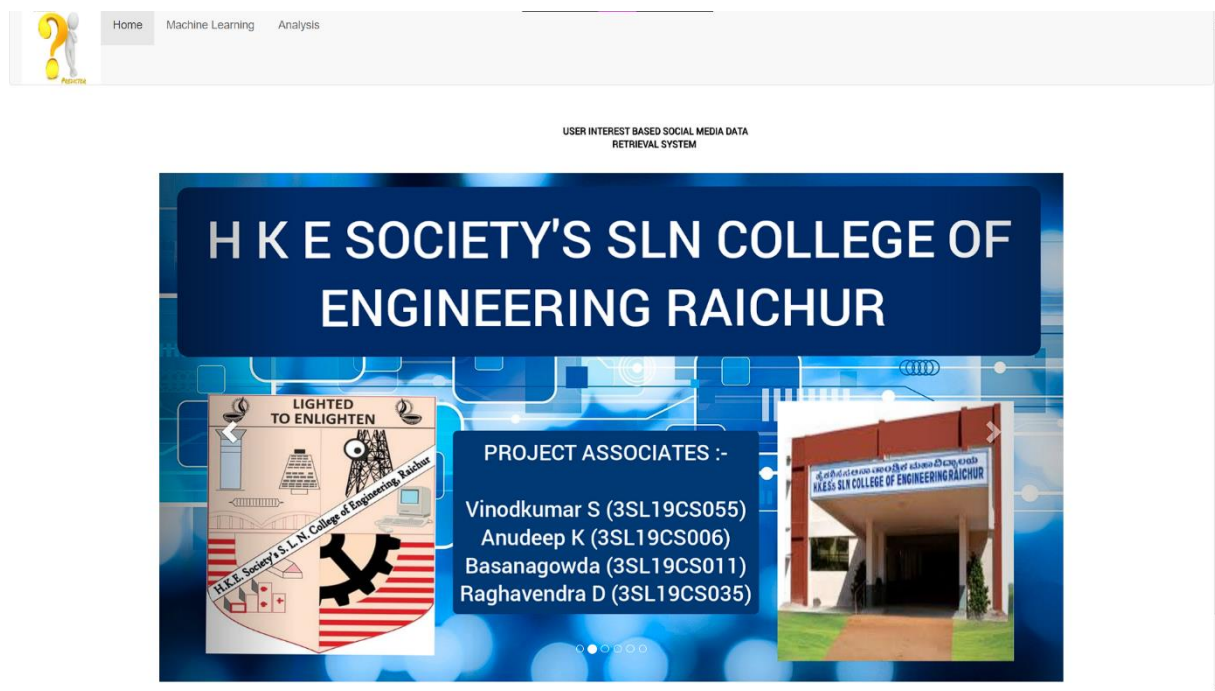


Fig 6.2 : home page 2

The screenshot shows a web application titled "USER INTEREST BASED SOCIAL MEDIA DATA RETRIEVAL SYSTEM". The navigation bar includes "Home", "Machine Learning", and "Analysis". The main heading is "Upload Annotated Data". Below it, a subtext states: "Accepts csv file with columns class and text." The interface includes a "File:" label with a "Choose File" button and "No file chosen" text. An "Upload" button is present. Below these are three dropdown menus: "Select Machine model for training and testing" (set to "Neural Network"), "Select csv for training and testing" (set to "AnnotatedData3.csv"), and "Select Document Vector" (set to "TF-IDF"). A "RUN" button is at the bottom. A footer note says: "This website is done by S L N college of Engineering students". The Windows taskbar at the bottom shows the date as 22-May-23 and time as 11:08 AM.

Home Machine Learning Analysis

USER INTEREST BASED SOCIAL MEDIA DATA
RETRIEVAL SYSTEM

Upload Annotated Data

Accepts csv file with columns class and text.

File: No file chosen

Select Machine model for training and testing

Neural Network

Select csv for training and testing

AnnotatedData3.csv

Select Document Vector

TF-IDF

This website is done by S L N college of Engineering students

Fig 6.3 : upload of Data

The screenshot shows the same web application as Fig 6.3, but with the heading "Test Your Data". The "File:" section remains the same. The "Select Machine model for training and testing" dropdown is now set to "Logistic Regression". The "Select csv for training" dropdown is set to "Annotated4.csv". A new "Select csv for testing" dropdown is added, also set to "Annotated4.csv". The "Select Document Vector" dropdown remains set to "TF". Two new dropdown menus are added: "Level 1" set to "90-100" and "Level 2" set to "70-80". The Windows taskbar at the bottom shows the date as 22-May-23 and time as 11:08 AM.

Home Machine Learning Analysis

USER INTEREST BASED SOCIAL MEDIA DATA
RETRIEVAL SYSTEM

Test Your Data

File: No file chosen

Select Machine model for training and testing

Logistic Regression

Select csv for training

Annotated4.csv

Select csv for testing

Annotated4.csv

Select Document Vector

TF

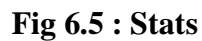
Level 1

90-100

Level 2

70-80

Fig 6.4 Test of Data



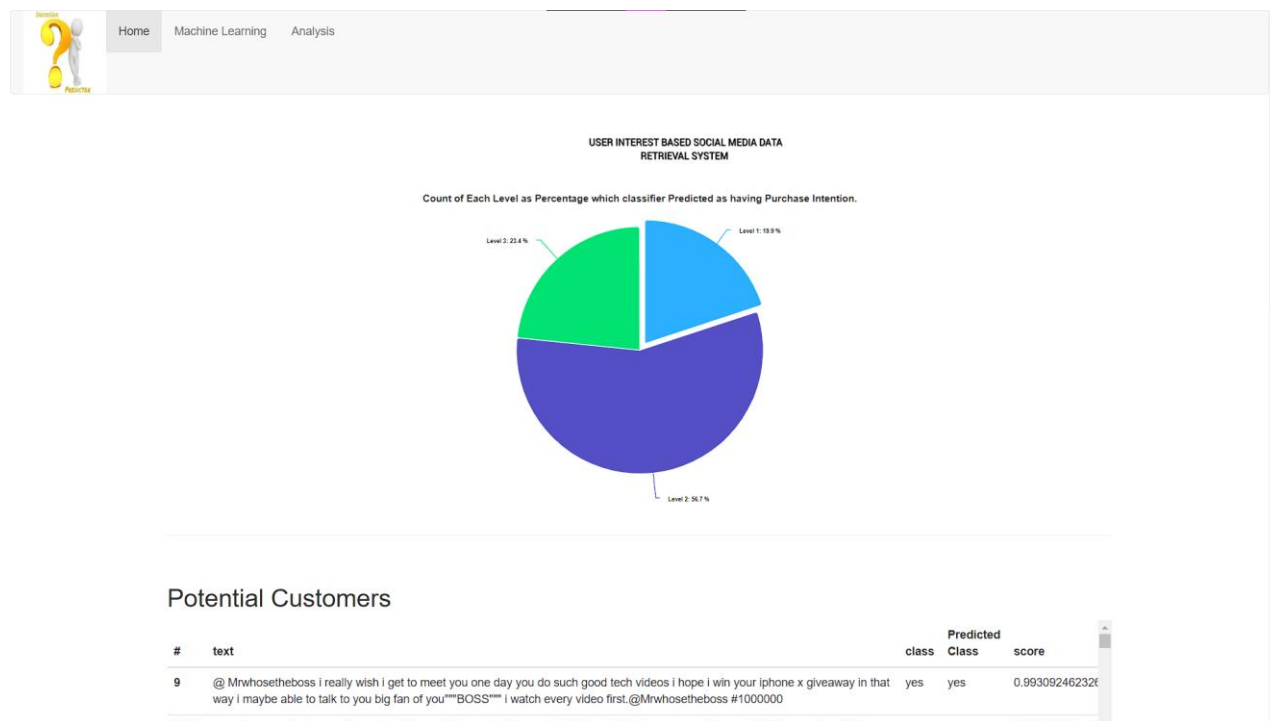


Fig 6.7 : Analysis

Potential Customers

| # | text | class | Predicted Class | score |
|------|---|-------|-----------------|---------------|
| 9 | @Mrwhosetheboss i really wish i get to meet you one day you do such good tech videos i hope i win your iphone x giveaway in that way i maybe able to talk to you big fan of you""BOSS"" i watch every video first.@Mrwhosetheboss #1000000 | yes | yes | 0.99309246232 |
| 1270 | I went to see about upgrading to the iPhone X & I could either pay \$276 & wait till february to get it or wait till february & get it free | yes | yes | 0.99259217991 |
| 367 | Want to win SENA Cases iPhone Xs Giveaway Enter to win: 1st Prize: iPhone X ? I just entered to win and you can too. http://gwwy.io/emu3ld0 | yes | yes | 0.99218679001 |
| 1989 | thats it. I hate spending money but fuck it imma just buy myself an iPhone X & maybe in a couple of weeks Ill get myself an Apple Watch | yes | yes | 0.99200141247 |
| 403 | Yeah, so far Apple has done a good job with OLED, I haven't read stuff about iPhone X or earlier Apple Watch having UI elements stuck. I just hope to see Promotion elsewhere (it might get tough in the Mac with most of them using iGPUs though). | yes | yes | 0.99018370959 |
| 264 | We finally got to Cali and I had to go go to Best Buy so we are in San Francisco and a homeless man with a stack of at least 5k runs into the Best Buy and day look I have money and throws it on the table and says I need an iPhone X | yes | yes | 0.98518663340 |
| 763 | When you purchase the iPhone X but it doesn't come to that store till December 1st so you have to use and 8 til then but idk I kinda like the 8... might have to just get the plus and refund that X | yes | yes | 0.98240678553 |
| 806 | I really need to go to sprint and look at the iphone 8 and iphone x and play w / them to see which one i really want cause ive been hearing mixed reviews | yes | yes | 0.97971309727 |
| 675 | I need a sugar daddy to buy me the new iPhone X for Christmas. Fuck independent woman for now! | yes | yes | 0.97947424163 |


Fig 6.8 : Potential Customers

Test Data Prediction Results

| # | text | class | Predicted Class | score |
|----|--|-------|-----------------|----------------|
| 0 | Some dude in FB selling the iPhone X 64 gb for \$1100 like nigga no one is gonna buy that shit when they can get the 256 gb for that price | no | no | 0.196277532506 |
| 1 | Home dab emote man today and I get hopped on by two full diamonds 5 minutes after pvp w tap too thanks for the same here in my bedroom with my brand new iPhone x x rated sexiest thing | yes | yes | 0.949915010804 |
| 2 | Buy an iPhone X | yes | yes | 0.697449303211 |
| 3 | I hate iOS 11. My iPhone 6+ works 10X slower now. Thanks @Apple subtle way of forcing me to purchase an iPhone 8 or X that I don't need . | no | no | 0.246851437884 |
| 6 | Bo-go sale tomorrow at T-Mobile buy one get one free. Phones and tablets. iPhone X is NOT included holla at me if you need more info. | no | no | 0.480727245986 |
| 7 | I can get the iPhone X here..... If I'm willing to buy at \$1500. | yes | yes | 0.670851630695 |
| 8 | When you buy an iPhone X with Unlimited everything | yes | yes | 0.625651914667 |
| 9 | @Mrwhosetheboss I really wish I get to meet you one day you do such good tech videos I hope I win your iPhone X giveaway in that way I maybe able to talk to you big fan of you""BOSS"" I watch every video first.@Mrwhosetheboss #1000000 | yes | yes | 0.993092462321 |
| 10 | @sprint @sprintcare if I switch to @TMobile I can get the iPhoneX because they'd buy out most of my contract | no | no | 0.377896778894 |
| 11 | Oneplus 5t has faster face recognition than iPhone X . | no | no | 0.171792239796 |
| 12 | Don't buy an iPhone X if you're ugly. I need time off work to recover from these selfies. | no | no | 0.270254508404 |
| 13 | Could buy 2x 5T with the price I paid for the iPhone X . Nice | no | yes | 0.552067797371 |

This website is done by S L N college of Engineering students

Fig 6.9 Potential Data Results


[Home](#)
[Machine Learning](#)
[Analysis](#)

USER INTEREST BASED SOCIAL MEDIA DATA RETRIEVAL SYSTEM

Upload Annotated Data

Accepts csv file with columns class and text.

File: No file chosen

Select Machine model for training and testing

Logistic Regression

Logistic Regression

SVM

Naive Bayes

Neural Network

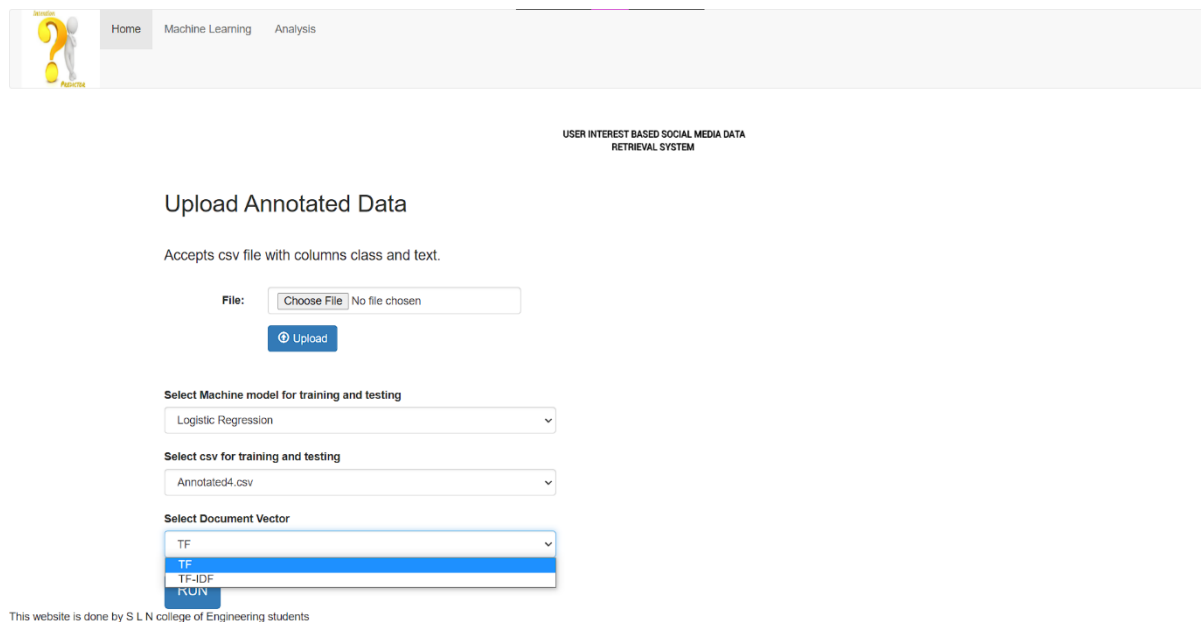
Decision Tree

Select Document Vector

TF

This website is done by S L N college of Engineering students

Fig 6.10 Select Machine Mode



USER INTEREST BASED SOCIAL MEDIA DATA RETRIEVAL SYSTEM

Upload Annotated Data

Accepts csv file with columns class and text.

File: No file chosen

Select Machine model for training and testing

Logistic Regression

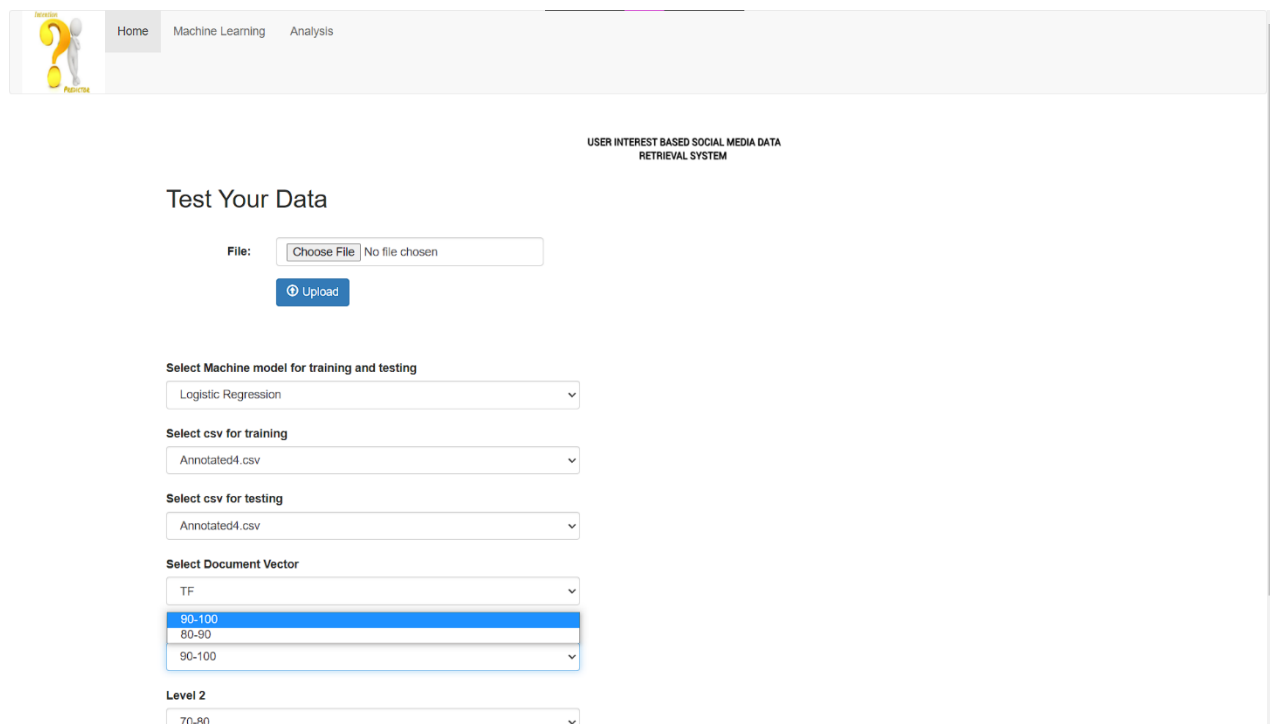
Select csv for training and testing

Annotated4.csv

Select Document Vector

TF
TF-IDF
RGIN

This website is done by S L N college of Engineering students

Fig 6.11 : Select the Document Vector

USER INTEREST BASED SOCIAL MEDIA DATA RETRIEVAL SYSTEM

Test Your Data

File: No file chosen

Select Machine model for training and testing

Logistic Regression

Select csv for training

Annotated4.csv

Select csv for testing

Annotated4.csv

Select Document Vector

TF
90-100
80-90
90-100

Level 2

70-80

Fig 6.12 : Select Level



Fig 6.13 : Positive and Negative Tweets

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

7.1 CONCLUSION

Our results were quite promising since we had created our own dataset and were building the model from scratch. We had to create our own dataset because there does not exist a publicly available dataset for purchase intention based on twitter tweets.

The 2 major problems that we faced were:

- i The imbalance class problem: Since our dataset was manually annotated by us, we had about 2000 positive tweets and 1200 negative tweets. Due to this we were getting a very low True Negative Rate and our model was not accurately predicting the negative class.
- ii Limited annotated data: Since we had to manual annotate each tweet in the dataset and this process takes a lot of time, we were only able to annotate about 3200 tweets.

Looking at the other researches that are done in the similar field, our project also stands apart since we have implemented 5 different models and after evaluating them, we choose the best one customized to the product data.

We were not able to get more than 80% accuracy because of the two problems highlighted above. To achieve even 80% accuracy with an imbalance class data and such a small dataset is a victory.

After showing the website to a few potential clients we have received positive remarks about our product and people seem interest in this new approach to customer identification and targeted marketing.

7.2 FUTURE SCOPE

To continue our work forward, it is worth trying out the dataset on deep learning models such as RNNs (recurrent neural networks), convolutional NN, and deep belief networks. Further, we can also use the dataset to find the intention shown towards specific features of the product rather than the product as a whole and target the user towards the specific feature of the product to increase the likeliness to purchase the product.

REFERENCES

1. Data Retrieval from Online Social Network Profiles for Social Engineering Applications, Sophia Alim, Ruquya Abdul-Rahman, Daniel Neagu and Mick Ridley Department of Computing, University of Bradford, BD7 1DP {S.Alim, R.S.H Abdul-Rahman, D.Neagu, M.J.Ridley}@bradford.ac.uk.
2. Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms, Information Systems 56(2016)1–18.