A PROJECT ON
"MOVIE REVIEW SENTIMENT ANALYSIS"

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE COURSE OF DIPLOMA IN BIG DATA ANALYTICS FROM CDAC

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY

'Sunbeam', Plot No. R/2, Market Yard Road,

Behind Hotel Flora, Gultekdi Pune-411037

SUBMITTED BY

NARENDRA MEWADA

UNDER THE GUIDENCE OF:

Mr. Shubham More

Faculty Member

Sunbeam Institute of Information Technology, Pune

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Pradnya Dindorkar (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Shubham More

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

NARENDRA MEWADA

DBDA February2019 Batch,

SIIT Pune

INDEX

1.Introduction and Background

People’s opinion has become one of the extremely important sources for various services in ever-growing popular social networks. In particular, online opinions have turned into a kind of virtual currency for businesses looking to market their products, identify new opportunities, and manage their reputations. In general, Rating Systems are defined as the supporting systems which help users to rate products, or services (such as books, movies, music, digital products, websites, and TV programs) by aggregating and analyzing suggestions from other users, which means reviews from various authorities, and user attributes.
After viewing such reviews, they take their decisions. So, such reviews must be correct and proper. Generally, the reviews are generated in graphical format that is in star ratings. Users just have to see the ratings which are generated by analyzing the ratings given by other users to that product and have to take his/her decisions. Such ratings are easily understandable by any user. They are helpful only in the scenario where if any product is excellent or very poor.
The scenario where product is average, star ratings prove bit confuse for any user. They don’t get clear views of what the other users think of that product. If the reviews are in simple graphical format it would be easy for any user to understand the feelings of the other users too, about the product. Also, star ratings will be there for his/her help. So, the review about any product will give clear idea to any user so that he can easily take his/her decisions in such confusing scenario too.

- His or her judgment or evaluation.
- Affective state (that is to say, the emotional state of the author when writing).
- The intended emotional communication (that is to say, the emotional effect the author

In last decade there is a rise of social media such as blogs and social networks, which has fuelled the interest in sentiment analysis. Online opinion has turned into a kind of virtual currency with the proliferation of reviews, ratings, recommendations and other forms of online expression, for businesses that are looking to market their products, identify new opportunities and manage their reputations. In order to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and following appropriate actions, many are now looking to the field of sentiment analysis. The problem of most sentiment analysis algorithms is that they use simple terms to express sentiment about a product or service. However, cultural factors, sentence negation, sarcasm, terseness, language ambiguity and differing contexts make it extremely difficult to turn a string of written text into a simple pro or con statement.

1.1 Statement of the problem:

A fundamental task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect. It focuses on whether the expressed opinion in a document, a sentence or a feature/aspect is positive, negative, or neutral. Sometimes it goes beyond polarity and looks at emotional states such as "angry," "sad," and "happy."
Another task in sentiment analysis is subjectivity/objectivity identification where it focuses on classifying a given text (usually a sentence) into one of the two classes (objective or subjective). As the subjectivity of words and phrases may depend on their context and an objective document may contain subjective sentences (a news article quoting people's opinions), this problem can sometimes be more difficult than polarity classification.

1.2 Technical Analysis:

Dataset:

The dataset used for this task was collected from Large Movie Review Dataset which was used by the AI department of Stanford University . The dataset contains 50,000 training examples collected from IMDb where each review is labelled with the rating of the movie on scale of 1-10. As sentiments are usually bipolar like good/bad or happy/sad or like/dislike, we categorized these ratings as either 1 (like) or 0 (dislike) based on the ratings. If the rating was equal or above 7, we deduced that the person liked the movie and if the rating was equal or below 4 we deduced that the person disliked the movie. Initially the dataset was divided into two subsets containing 25,000 examples each for training and testing.

1.3 Goal:

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. 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.

1.4 Project Goals and Scope:

Information presented in cyberspace is now more diverse and the media used means in the process of information diffusion are growing. One of the main media used in the process diffusing information in cyber media is text or document media. The ability in order to extract information from documents is absolutely necessary. The method of extracting information from data in the form of documents is known as text mining. Over 80% of information is currently stored in the form of text, so that text mining is believed to have high commercial value potential , describes the steps taken in text mining: tokenizing, filtering, stemming,.

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to extract and identify subjective information in source materials. It aims to determine the attitude of a speaker or a writer with respect to some topic .The attitude can be,

- His or her judgement or evaluation,
- Affective state (Emotional state of the writer),
- The intended emotional communication.

Sentiment analysis is one of the new branches in the domain of text mining or data extraction in the form of text, consisting of processing and extracting textual data automatically in order to obtain information . Sentiment analysis can be utilized as a tool in seeing the public response of a particular event, either positive or negative response, so that the next strategic steps can be undergone immediately. An example of sentiment data in the form of document is movie review from various sites on the internet. Reviews obtained from movie reviews sites can be used a reference for movie fans to know recommended and also a medium for movie producers to know the public responses towards the movie released. Movie review can be divided into a number of categories based on the sentiments contained in the document.

1.5 Benefits:

Sentiment analysis can be utilized as a tool in seeing the public response of a particular event, either positive or negative response, so that the next strategic steps can be undergone immediately.

Sentiment analysis is the aim to uncover the attitude of the user on a particular movie from the written text. It is the process of extracting, identifying, analysing, and characterizing the sentiments or opinion in the form of textual information. It identifies the sentiment holders and the entity about which sentiment is expressed. Sentiment analysis is a well-known task in the realm of natural language processing. Given a set of texts; the objective is to determine the polarity of that text. Provides a comprehensive survey of various methods, benchmarks, and resources of sentiment analysis and opinion mining. The sentiments can consist of different classes. A movie review is positive or negative. Where they also employ a novel similarity measure. Sentiment analysis can regroup the opinions of the reviewers and estimate ratings on certain aspects of the product. Another utility of sentiment analysis is for companies that want to know the opinion of customers on their products. They can then improve the aspects that the customers found unsatisfying. Sentiment analysis can also determine which aspects are more important for the customers.

2.Overview and Summary:

2.1 Purpose:

The aim of this project is :

- To summarize the film so that people have more information with which to gauge their interest in the film.
- To analyse the film in terms of its artistic value and technical skill. This includes things like acting, direction, cinematography, writing, and other aspects.
- To analyse the film in terms of its messages and cultural representations.
- To help those with limited money and opportunity to go to the movies decide which film is most worth their time and money.

2.2 System Architecture:

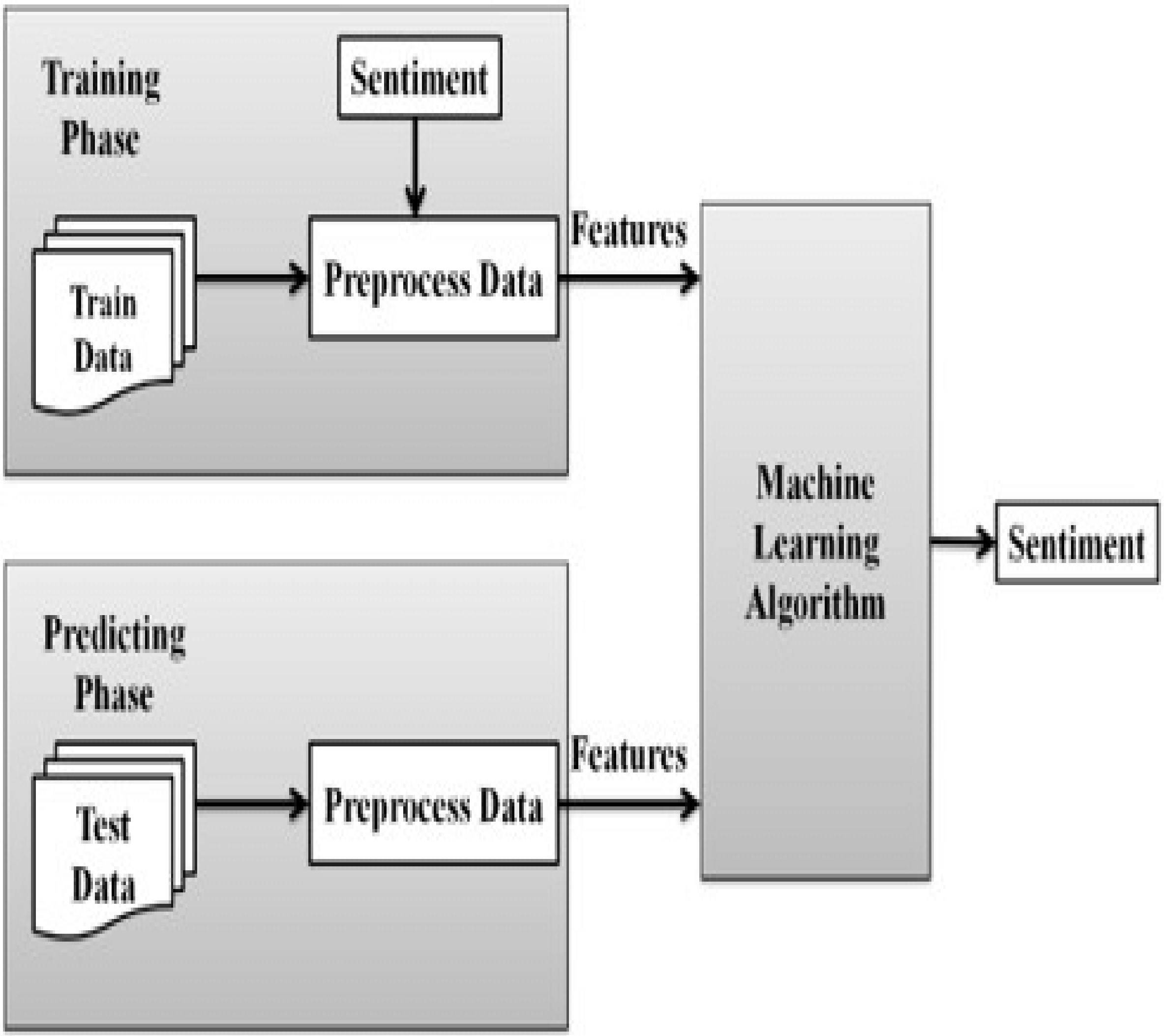
Data preparation involves collecting and pre-processing user reviews for the subsequent analysis. Different pre-processing steps may be required depending on the data sources. A user review is likely to be a semi structured document, containing some structured headers and an unstructured text body. Sentiment analysis algorithms usually do not use information other than the comments and the original ratings given by the users. The review analysis step includes several tasks that help identifying interesting information in reviews, which are unstructured, natural language texts. Some words have negation effects on other words, and negation tagging aims at identifying such words and reflecting their effects when determining the reviews for example, “good” and “not good” obviously represent opposite sentiments. Feature generalization, or metadata substitution is about generalizing features that may be overly specific. This task can be performed when attributes of domain items are available. For the movie reviews domain, for example, a sentence “Toy Story is pleasant and fun.” in which “Toy Story” is the name of the movie being reviewed, is generalized to “MOVIE is pleasant and fun.” On the basis of review database construct an opinion dictionary. An opinion dictionary contains opinion words, their estimated sentiment oriented and the strengths of their sentiment oriented. Determining the SO and strengths of opinion words is done by answering the question: “Given a certain opinion word, how likely is it to be a positive sentiment, and how likely is it to be a negative one.

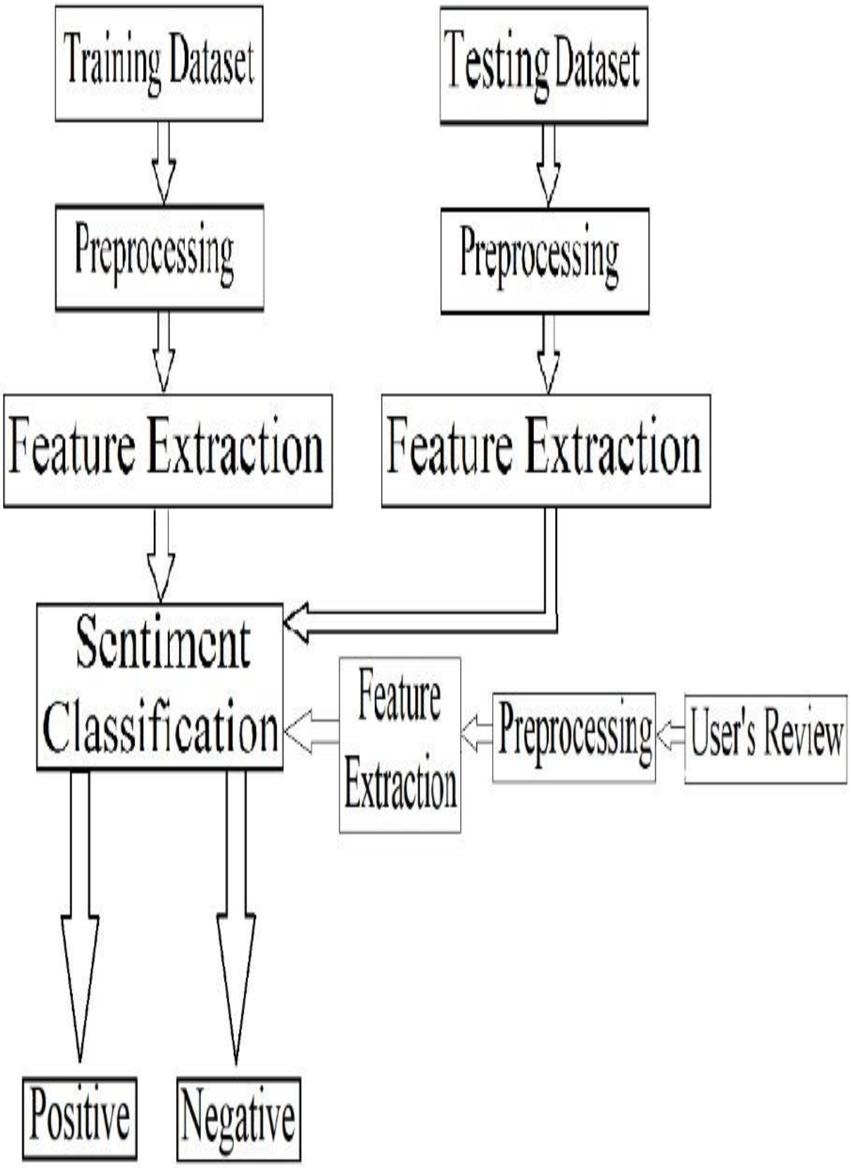
Data pre-processing:

A typical review text looks like this:

> I'm a fan of TV movies in general and this was one of the good ones. The cast performances throughout were pretty solid and there were twists I didn't see coming before each commercial.<br /><br /> Did anyone else think that in certain lights, the daughter looked like a young Nicole Kidman? Are they related in any way? I'd definitely watch it agin or rent it. <br /><br />Dedee was great. Haven't seen in her in a lot of things and she did her job very convincingly. <br /><br />If you're into to TV mystery movies, check this one out if you have a chance.

As seen above, one necessary pre-processing step prior to feature extraction was removal of HTML tags like “<br>”. We used simple regular expressions matching to remove these HTML tags from the text. Another important step was to make the text case-insensitive as that would help us count the word occurrences across all reviews and prune unimportant words. We also removed all the punctuation marks like ‘!’, ‘?’, etc as they do not provide any substantial information and are used by different people with varying connotations. We also removed stopwords from the text for some of our feature extraction tasks. stemming of words as a word to its root form.


Training Phase
Sentiment
Train Data
Preprocess Data
Features
Machine Learning Algorithm
Sentiment
Predicting Phase
Test Data
Preprocess Data
Features



2.3 Overview:

For this analysis we'll be using a dataset of 50,000 movie reviews taken IMDB site. The data is split evenly with 25k reviews intended for training and 25k for testing your classifier.

The raw text is pretty messy for these reviews so before we can do any analytics we need to clean things up. Understanding and being able to use regular expressions is a prerequisite for doing any Natural Language Processing task.

2.3.1 Vectorizer:

In order for this data to make sense to our machine learning algorithm we'll need to convert each review to a numeric representation, which we call *vectorization*.

The simplest form of this is to create one very large matrix with one column for every unique word in your corpus (where the corpus is all 50k reviews in our case). Then we transform each review into one row containing 0s and 1s, where 1 means that the word in the *corpus* corresponding to that column appears in that *review*. That being said, each row of the matrix will be very sparse (mostly zeros). This process is also known as *one hot encoding*.

2.3.2 Text Processing:

Stemming/Lemmatizing to convert different forms of each word into one.

Stemming : A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form, for example,

connection
connections
connective ---> connect
connected
connecting

Lemmatization : Lemmatization is similar to word stemming but it does not require to produce a stem of the word but to replace the suffix of a word, appearing in fre e text, with a (typically) different word suffix to get the normalized word form. For instance, the suffixes of words working, works, worked would change to get the normalized form work standing for the infinitive: work; in this case, both the normalized word form and the word stem are equal. Sometimes the normalized form may be different than the stem of the word. For example, the words computes, computing, computed would be stemmed to compute, but their normalized form is the infinitive of the verb: compute.

Tokenization:

One common task in NLP (Natural Language Processing) is tokenization. "Tokens" are usually individual words (at least in languages like English) and "tokenization" is taking a text or set of text and breaking it up into its individual words. These tokens are then used as the input for other types of analysis or tasks, like parsing (automatically tagging the syntactic relationship between words).

n-grams: Instead of just single-word tokens (1-gram/unigram) we can also include word pairs.
To perform N-Gram-Based Text Categorization we need to compute N-grams (with N=1 to 5) for each word - and apostrophes - found in the text, doing something like (being the word “TEXT”):

- bi-grams: _T, TE, EX, XT, T_
- tri-grams: _TE, TEX, EXT, XT_, T_ _
- quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _

Representations: Instead of simple, binary vectors we can use word counts or *TF-IDF* to transform those counts.

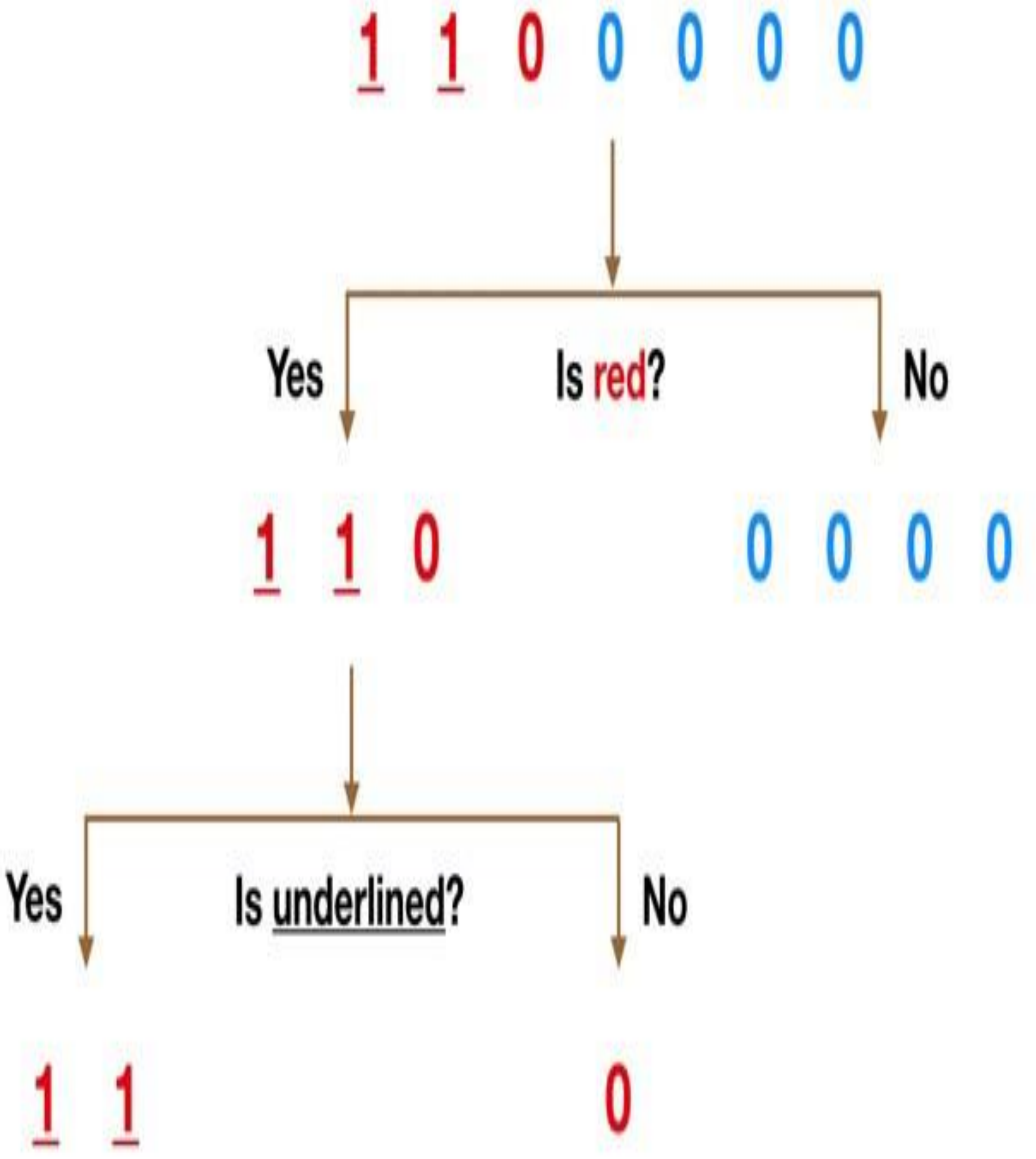
TF-IDF :

tf-idf stands for *Term frequency-inverse document frequency*. The tf-idf weight is a weight often used in information retrieval and text mining. Variations of the tf-idf weighting scheme are often used by search engines in scoring and ranking a document's relevance given a query. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus (data-set).

3. Algorithm:

3.1 Random Forest Classifier:

Random forest Decision tree learning is one of the most popular techniques for classification. Its classification accuracy is comparable with other classification methods, and it is very efficient. ID3 presented by Quinlan (1986), C4.5 presented by Quinlan (1993) and CART presented by Breiman et al (1984) are decision tree learning algorithms. Details can be found in article of Han et al (2006). Random forest belongs to the category of ensemble learning algorithms. It uses decision tree as the base learner of the ensemble. The idea of ensemble learning is that a single classifier is not sufficient for determining class of test data. Reason being, based on sample data, classifier is not able to distinguish between noise and pattern. So it performs sampling with replacement such that given n trees to be learnt are based on these data set samples. Also in our experiments, each tree is learnt using 3 features selected randomly. After creation of n trees, when testing data is used, the decision which majority of trees comes up with is considered as the final output. This also avoids problem of over-fitting .



Imagine that our dataset consists of the numbers at the top of the figure to the left. We have two 1s and five 0s (1s and 0s are our classes) and desire to separate the classes using their features. The features are colour (red vs. blue) and whether the observation is underlined or not. So how can we do this? Colour seems like a pretty obvious feature to split by as all but one of the 0s are blue. So we can use the question, “Is it red?” to split our first node. You can think of a node in a tree as the point where the path splits into two — observations that meet the criteria go down the Yes branch and ones that don’t go down the No branch.

The No branch (the blues) is all 0s now so we are done there, but our Yes branch can still be split further. Now we can use the second feature and ask, “Is it underlined?” to make a second split. The two 1s that are underlined go down the Yes sub-branch and the 0 that is not underlined goes down the right sub-branch and we are all done. Our decision tree was able to use the two features to split up the data perfectly. Victory! Obviously in real life our data will not be this clean but the logic that a decision tree employs remains the same.

3.2 Logistic Regression:

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

Types of logistic regression:

1. Binary logistic regression:
   The categorical response has only two 2 possible outcomes. Example: Spam or Not.
2. Multinomial logistic regression:
   Three or more categories without ordering. Example: Predicting which food is preferred more (Vegetarian, Non-Vegetarian, Vegan).
3. Ordinal logistic regression:
   Three or more categories with ordering. Example: Movie rating from 1 to 5.

Decision Boundary :

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes.
Say, if predicted_value ≥ 0.5, then classify email as spam else as not spam.
Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary.

3.3Naive Bayes:

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis. These classifiers are widely used for machine learning because they are simple to implement.

Naive Bayes is also known as simple Bayes or independence Bayes.

A naive Bayes classifier uses probability theory to classify data. Naive Bayes classifier algorithms make use of Bayes' theorem. The key insight of Bayes' theorem is that the probability of an event can be adjusted as new data is introduced.

What makes a naive Bayes classifier naive is its assumption that all attributes of a data point under consideration are independent of each other. A classifier sorting fruits into apples and oranges would know that apples are red, round and are a certain size, but would not assume all these things at once. Oranges are round too, after all.

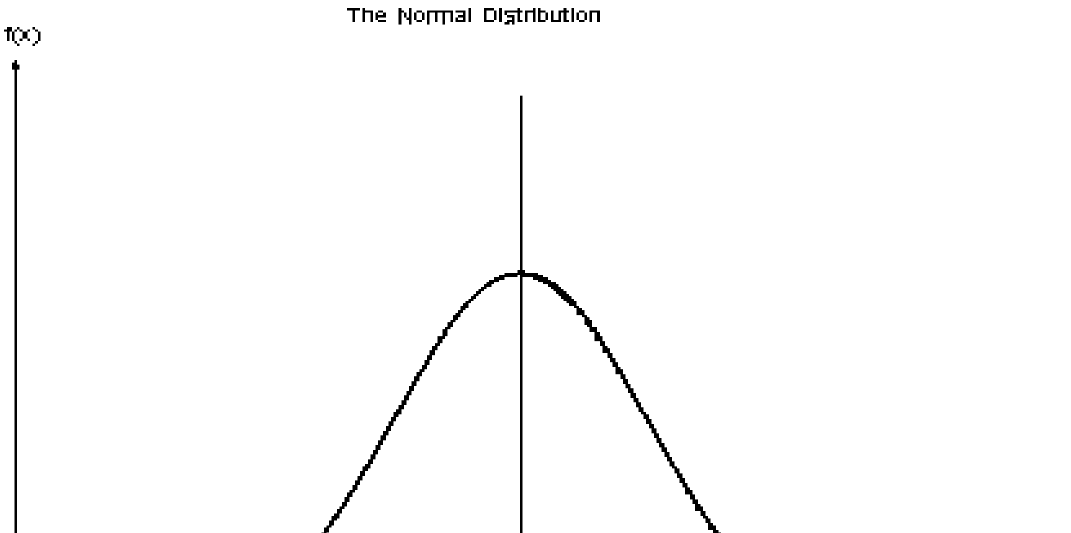
A naive Bayes classifier is not a single algorithm, but a family of machine learning algorithms that make uses of statistical independence. These algorithms are relatively easy to write and run more efficiently than more complex Bayes algorithms.

The most popular application is spam filters. A spam filter looks at email messages for certain key words and puts them in a spam folder if they match.

Despite the name, the more data it gets, the more accurate a naive Bayes classifier becomes, such as from a user flagging email messages in an inbox for spam.

Types of Naive Bayes Classifier:

1. Multinomial Naive Bayes:
   This is mostly used for document classification problem, i.e., whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.
2. Bernoulli Naive Bayes:
   This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.
3. Gaussian Naive Bayes:
   When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.



Gaussian Distribution(Normal Distribution)

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real life cases, the predictors are dependent, this hinders the performance of the classifier.

3.4 Accuracy Table

ALGORITHM	ACCURACY
Linear Regression	77.25%
Random Forest	79.07%
Logistic Regression	83.92%
Multinomial Naive Bayes	84.03%

4.Conclusion:

From the results above, we can infer that for our problem statement, Logistic Regression Model with feature set using mixture of Unigrams and Bigrams is best. Apart from this, one can also use a Naïve Bayes' Classifier or a SGD classifier as they also provide good accuracy percentage. One peculiar thing to note is low accuracy with Random Forest classifier. This might be because of over-fitting of decision trees to the training data. Also, low accuracy of kNN Classifiers shows us that people have varied writing styles and kNN Models are not suited to data with high variance. One of the major improvements that can be incorporated as we move ahead in this project is to merge words with similar meanings before training the classifiers. Another point of improvement can be to model this problem as a multi-class classification problem where we classify the sentiments of reviewer in more than binary fashion like “Happy”, “Bored”, “Afraid”, etc This problem can be further remodeled as a regression problem where we can predict the degree of affinity for the movie instead of complete like/dislike.

Sentiment analysis appeal is tremendously developing. The researchers can be faced with problems, oppositions and challenges; though semantic analysis is emerging but this field is still new to them .Problem related to the nature of classification is a possible challenge. Classification techniques that create two or three groups at most, there is a limit in the extraction of groups and subgroups. Text based data that is valid in specific places at specific times are usually context specific and domain dependent. Sentiment observation has a broad type of approaches in information systems, which includes categorizing reviews, encapsulating analysis (review) and different real time applications. There are probably to be various distinct applications particularly not examined. It is constituted that sentiment words (classifiers) are acutely based on fields or areas or topics. Sentiment analysis can be tried or auditioned for new applications. Even though the algorithms and methods used for Sentiment analysis (SA) are proceeding fast, but, a lot of issues in this area of study prevail unsolved. The main demanding details exist in hold of other languages, trading with negation proclamations(expressions); construct or yield a summary of beliefs build on product features(attributes), difficulty of sentence(document), managing of implied(indirect) product attributes etc. More research could be committed to these problems (challenges) in time to come.