TEXT MINING AND TS APPLICATION: SENTIMENTAL ANALYSIS

Abstract

The project goals to extract the opinions, emotions, attitudes of public towards different object of interest. Sentiment analysis is a form of shallow semantic analysis of texts. In the project we used the automatic approach that involves supervised machine learning and text mining classification algorithms which includes the sentimental analysis in various applications. Various fields like twitter tweets, movie review tweets, election result tweets, digital libraries, life sciences, social media tweets and various other fields have been analyzed and the algorithms like line regression, Support vector machine (SVM), Naïve bayes are used in every content and a result is brought out through the means of various types of graphs. More specifically, MOVIE REVIEW TWEETS have been considered and the scrapping reviews of a particular movie after the predicting the sentiment of each review and on that basis, we decide whether the movie or not. It is predicted that the users generalize the other applications using the specific content discussed above.

Keywords: Text mining, sentiment analysis, line regression, naïve bayes, Support vector machine, lexicon-based approach, supervised and unsupervised machine learning.

I. INTRODUCTION

Sentiment analysis is the process of using natural language processing, text analysis, and statistics to analyse customer sentiment. The best businesses understand the sentiment of their customers—what people are saying, how they’re saying it, and what they mean. Customer sentiment can be found in tweets, comments, reviews, or other places where people mention your brand. Sentiment Analysis is the domain of understanding these emotions with software, and it’s a must-understand for developers and business leaders in a modern workplace. As with many other fields, advances in deep learning have brought sentiment analysis into the foreground of cutting-edge algorithms. Today we use natural language processing, statistics, and text analysis to extract, and identify the sentiment of words into positive, negative, or neutral categories. It is used for brand monitoring, customer service, market research and analysis, analysing reviews, etc

II. BACKGROUND

In most of the papers we analyzed they used different algorithms like Naïve Bias, Logistic regression, Support vector machine, etc. using different dataset having various topics. But many of these algorithms take more time due to their several iterations and data mining functionalities used. And, also due to this much amount of time these methods have to suffer from degradation in accuracy. Whereas in our project we are going to compare the Naïve Bayes Classifiers, Support Vector Machine and Logistic Regression for our Phrase review dataset and find out finally the classifier with the higher accuracy and confusion matrices.

III. Literature Survey

[1]. The paper is all about the detection of hate speeches in the application of twitter. In this paper, they experimented with multiple classier such as Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs) and Deep Neural Net- works (DNNs). The paper has done investigation on the application of deep neutral network architectures for the task of hate speech detection. It was found to significantly outperform the ancient or existing methods. It has been stated that the embeddings learnt from deep neural network models when are combined with gradient boosted decision trees, they led to best accuracy values.

[2]. The paper talks about the figures of the vulgarity used in the social media which we use in day to day lives. In this paper, the Annotators evaluated tweet sentiment on a five-point scale: (1) very negative, (2) somewhat negative, (3) neutral, (4) somewhat positive, and (5) very positive. These five-point scales have been given the scaling and Annotations from users with a Spearman correlation coefficient less than 0.3 were removed from computing consensus Labels. Calculation of the demographics, frequency of vulgarity and sentiment perception is shown as a part of filling the data parameters and tables. Experiments like evaluation of mean absolute error, quantitative analysis are stated and overall, the paper is based on the five-point scale analysis.

[3]. In this paper, the context deals with the release to the community to do high performance and high-coverage lexica, targeting English and Italian languages and to extensively benchmark different setup decisions affecting the construction of the two resources, and further evaluate the performance obtained on several datasets/tasks exhibiting a wide diversity in terms of domain, languages, settings and task. Furthermore, it has been shown that how simple techniques can be used, both in supervised and unsupervised experimental settings, to boost performances on datasets and tasks of varying degree of domain-specificity. Works like Sentiment Lexica, Emotion Lexica, unsupervised regression experiments, supervised regression experiments, supervised classification experiments are majorly focused for detecting the lexicon of the users.

[4]. In this paper, the context deals with the release to the community to do high performance and high-coverage lexica, targeting English and Italian languages and to extensively benchmark different setup decisions affecting the construction of the two resources, and further evaluate the performance obtained on several datasets/tasks exhibiting a wide diversity in terms of domain, languages, settings and task. Furthermore, it has been shown that how simple techniques can be used, both in supervised and unsupervised experimental settings, to boost performances on datasets and tasks of varying degree of domain-specificity. Works like Sentiment Lexica, Emotion Lexica, unsupervised regression experiments, supervised regression experiments, supervised classification experiments are majorly focused for detecting the lexicon of the users.

[5]. The paper is all about categorization of sentiment polarity based on sentiment analysis. When specified with a text we can categorize them into being positive, negative or being neutral. There are three levels in sentiment polarity categorization. The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence’s sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions. Total Sentiment Index can be used to categorize a token, whether it is a positively classified token or whether it is negative. TSI index can be calculated using a generalized formula. Some of the methods of analysis for the document includes scikit-learn, an open-source machine learning software package in Python. The classification models selected for categorization are: Naïve Bayesian, Random Forest, and Support Vector Machine.

[6]. Accessibility of social media platforms empowered the internet users to express and share their opinions on different kinds of components based on their life experience, including products and services that they enjoy. Sentiment Analysis has been a burgeoning technology that taps into customer demands based on Natural Language Processing. This motivation is usually used to properly understand what customers want, when, why and how they want it, retailers need to pivot toward sentiment analysis, hence avoid doing the same mistakes and choosing the right decisions based on comments or reviews. As part of e-commerce, online shopping is a good example of how products or services are sold over the Internet. Big name distributors like Amazon and Alibaba along with tiny distributors out there certainly had disappointing outcomes, one of the primary factors for their slow sales was poor product assortment. Consumer understanding has always been high on the to-do list of distributors and the use of sentiment analysis to monitor those emotions was the main motive for businesses to understand how diverse and thorough the opinion mining on the clients' reviews can be. Some of the techniques that can be used are: In sentiment classification, there are two main study fields such as Machine Learning and Lexicon, and each field has its own subdivision.

[7]. The study focuses on classifying reviews according to the scale determined by the authors. Original evaluations will be used as complements for evaluating the consistency between the classification inferred from the text and the one assigned by the reviewer. The consistency evaluation between the written review and the reviewers' score is proposed as a practical application of sentiment classification. For these reasons, the classifier used in this study was trained according to manual data tagging, not the reviewer's original classification. This allows revising the consistency between what the review states and what the reviewer says about the paper acceptance or rejection. Some of the tools and methods used in this review are : Python programming language, Scikit-learn library, Stanford POS Tagger library, SentiWordNet 3.0 lexical ontology. Algorithms that can be used are: SVM and Naive Bayes.

[8]. In this study we collect, label and thus create a dataset of Persian-English code-mixed tweets. We then proceed to introduce a model which uses BERT pretrained embeddings as well as translation models to automatically learn the polarity scores of these Tweets. Our model outperforms the baseline models that use Naïve Bayes and Random Forest methods. We aim to create a vectorized representation of the textual input in order to be able to fit the data into our machine learning model.

[9]. In this paper, it has been stated regarding the introducing the MELD, a multimodal multi-party conversational emotion recognition dataset. The process of building the dataset and providing with the results obtained with strong baselines methods applied on dataset is displayed. The dataset is used as a training corpus for both conversational emotion recognition and multimodal empathetic response generation. Experiments like feature extraction, usage of baseline models which includes the text-CNN, bcLSTM, Dialogue RNN are described and accurate results have been tabulated. Various related dataset applications have been discussed for future works even.

[10]. In this paper, the uses of sentiment analysis by highlighting the reasons why the domain uses sentiment analysis, the most commonly used techniques in the domain, the prospective opportunities of sentiment in the medical context, and outlined the challenges faced in opinion mining. The main technique used in sentiment analysis is the classification technique which is machine learning based is well described here. Analysis like performance quantification, In-depth insight, staff's motivation and its techniques are prescribed. Classification based on machine learning algorithms are promoted for the analysis in health care. Besides many challenges about sentiment analysis in the healthcare domain such as negated expressions, irony and sarcasm, and co-reference resolution, there is a scope in it in future.

[11]. In this paper, the information is provided on studies on sentiment analysis in social media. The contribution of the paper is to show what is the method used in analyzing sentiment in social media. The most common method uses in Lexicon based method is SentiWordnet and TF-IDF while for machine learning is Naïve Bayes and SVM. It also deals with identifying what is the common type of social media site to extract information for sentiment analysis. Third, it is demonstrated about the application of sentimental analysis in social media. Investigation to develop a universal model of sentiment analysis that can be applied to a different type of data, explores other potential social networking sites to obtain users opinion and expanding the context of sentiment analysis application is the further step described here.

[12]. In this paper the implementation of BERT for the financial domain by further pre-training it on a financial corpus and fine-tuning it for sentiment analysis (FinBERT) is done.
In addition to BERT, also the implementation of other pre-training language models like ELMo and ULMFit are used for comparison purposes. Conducting extensive experiments with BERT, investigating the effects of further pre-training and several training strategies, such kind of methodology is even prescribed. Not only that but also applying an LSTM neural network to ad-hoc company announcements to predict stock-market movements and

show that method to be more accurate than traditional machine learning approaches is also sentimental. FinBERT approaches were found in basic aspect that are to be implemented along with the natural language processor.

[13]. In this paper, a bibliometrics research study of sentiment analysis using the Scopus database is displayed. The most obvious one analysis that has stated is the field has experienced exponential growth. Cooccurrence maps were used in making the system helpful in finding similarities and patterns between keywords and authors. Thus, it helped researchers find areas where they can contribute the most to the scientific community. Secondly, to gain insights beyond the keywords assigned by authors and editors, Latent Dirichlet Allocation was applied on the titles and abstracts of the publications. LDA clustered the publications into topics similar to those found by the keyword co-occurrence maps, which confirms the focal points of the sentiment research.

[14]. In this paper, the proposal is of a novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus and one domain-independent corpus. It discusses existing techniques and approaches for feature extraction in sentiment analysis and opinion mining. The methodology applied is prescribed as extracting the feature, modifier and opinion from the dataset and then using clustering mechanism and dividing them into discrete clusters by user’s opinion. Mining techniques like supervised learning techniques, unsupervised learning techniques, case-based reasoning are applied in the context system. Tools like WEKA, NLTK, Red Opal are used for purposes like data pre-processing, network analysis visualization and other purposes.

[15]. In this paper, the study focuses on the online reviews generally shared by the hotel customers which reflects the experience and the satisfaction via the sentiments such as positive, negative, or neutral. Here, a hybrid approach combining Lexicon-based and Machine Learning is proposed to predict the sentiments more accurately. The methodology used here is breaking the component into collecting of data and other is for analysing the data. Acquisition, pre-processing, polarity detection, sentimental classification are the approaches to be achieved in the system using the mentioned methodology.

[16]. In this paper, the main role is of the human annotators which specify the multilingual sentiment classifications. Here the nature and proper formalization of sentimental classification, selection of the most appropriate classifier, the acceptable levels of the annotator’s agreements are well described and discussed. Classification model’s performance includes the variants of Support Vector Machine SVM which further extends into multi class and regression classifiers. Naïve Bayes, SVM algorithms are stated in the prescription with the five extensions of SVM: Neutral Zone SVM, Two Planes SVM, Two Planes SVM bin, Cascading SVM, Three Planes SVM. Finally, the Friedman Nemenyi Test is applied as a part of ending to the sentiment classification.

[17]. In this paper, the AutoML, one of the highly focused research areas in Mining and machine learning. A considerable growth and interest in doing data mining applications is presented in the paper. However, the AutoML is a newly introduced evolving technology and lots of researches are being conducted in this particular area hence has several pros and cons that are identified in each AutoML library. Moreover, using AutoML libraries for each sector such as Image Classification, Time Series based Predictions, or Sentiment analysis, those libraries performance are varying in all sites. They investigate AutoML tools like TPot, Auto weak, and several other tools. Moreover, the evaluation of AutoML platforms in clouds such as H2O AutoML, Google AutoML.

[18]. In this paper, the whole analysis task is finished by utilizing Rule based and Machine learning based systems and found that machine learning procedure is more precise and accurate in anticipating the conclusion of a sentence or finding sentiment associated with sentence. Methodology like creating a dataset, rule-based mechanisms, sentiment Vader, sent word net, LDA analysis on Naïve Bayes are applied in every analysis aspect and proceeded further.

[19]. In this paper, it is made clear that text mining is an emerging sphere of data mining used to receive the knowledge of the huge mass of data. Much research has been carried on to mine the opinions in the contour of a document, sentence and feature level sentiment analysis. It has been examined that now the opinion mining trend is proceeding to the sentimental reviews of social media data, comments used in social media on pictures, videos or social media status. The sentimental analysis approach of text Mining in detail with the techniques like data setting supervised learning techniques, unsupervised learning techniques, and tools like review seer tool, web fountain, red opal are used.

[20]. Twitter’s strengths real-time and each one the tweets are publicly available and are easily accessible with their geo-tagged locations. Natural language Processing (NLP)is the sub-branch of knowledge science and it's assumed to be vital neighborhood of data science. It generally teaches machines to read and interpret human readable texts. It converts information from computer databases or sentiment intents into readable human language. Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Sentiment Analysis of text identifies and extracts subjective information in the source material, and helping a business to know the social sentiment of their brand, product, or service while monitoring online conversations There are two kinds of machine learning techniques like supervised machine learning algorithms like maximum entropy, SVM, Naïve Bayes, KNN, etc. and unsupervised machine learning algorithms like Neural network, Principal Component Analysis, ICA, SVD, etc.

[21]. This research aims to analyses political orientation of twitter users as positive or negative. Different machine learning algorithms are adopted to identify the user point of view on Ajodhya issue. Then the efficiencies of these built models are contrasted with one another to discover the best machine learning algorithm on text categorization. The result is analyzed based on the measurement metrics namely accuracy, precision, recall, F1- score measures for each sentiment class Author implemented Naïve Bayes and Support Vector Machine classifiers in order to group the twitter data into positive and negative tweets. At that point, the locations are put into categories and sentiment mapping is done which helped in analyzing the opinions of individual Indian states separately. Important features were considered for classification such as latitude, longitude, kilometers and number of tweets to distinguish the opinions based on region. Importance of dataset pre-processing is also demonstrated in the paper. However, the author recognized state-wise people reactions to demonetization, because of population polarization, the general feeling of the citizens couldn't be caught.

[22]. Opinion mining and sentiment analysis in Online Learning Community can truly reflect the students’ learning situation, which provides the necessary theoretical basis for following revision of teaching plans. To improve the accuracy of topic-sentiment analysis, a novel model for topic sentiment analysis is proposed that outperforms other state-of-art models.

Following methods have been used: 1) Precision contrast between different methods based on SVM. 2) Recall contrast between different methods based on SVM. 3) Measure contrast between different methods based on SVM 4) MAE contrast between different methods based on SVM. This paper designs a model for online sentiment analysis of various topics in OLC. The model obtains the topic-terminology hybrid matrix and the document topic hybrid matrix by selecting the real user's comment information on the basis of LDA topic detection approach. Afterwards, a topic clustering concept lattice based on FCA model is constructed, where the topic sentiment can be identified by measuring their sentiment scores.

[23]. In this paper, the fast and in memory computation framework 'Apache Spark' to extract live tweets and perform sentiment analysis. The primary aim is to provide a method for analyzing sentiment score in noisy twitter streams. This paper reports on the design of a sentiment analysis, extracting vast number of tweets. Results classify user's perception via tweets into positive and negative. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. Apache Spark is an open-source lightning-fast cluster computing platform to retrieve streaming data and forwarding to storage system like Database Server. Apache spark is an in-memory fast processing system used for large scale data processing the conducted experiments through sentiment classification algorithms have shown the performance measures of precision, recall and accuracy. They have applied NB and SVM classifiers. These classifiers provide a useful perspective for understanding and evaluating many learning algorithms.

[24]. Humans frequently are able to read and interpret emotions of others by directly taking verbal and non-verbal signals in human-to-human communication into account or to infer or even experience emotions from mediated stories. For computers, however, emotion recognition is a complex problem: Thoughts and feelings are the roots of many behavioral responses and they are deeply entangled with neurophysiological changes within humans. As such, emotions are very subjective, often are expressed in a subtle manner, and are highly depending on context. For example, machine learning approaches for text-based sentiment analysis often rely on incorporating sentiment lexicons or language models to capture the contextual meaning. This paper explores if and how we further can enhance sentiment analysis using biofeedback of humans which are experiencing emotions while reading texts. Specifically, we record the heart rate and brain waves of readers that are presented with short texts which have been annotated with the emotions they induce. We use these physiological signals to improve the performance of a lexicon-based sentiment classifier. We find that the combination of several bio signals can improve the ability of a text-based classifier to detect the presence of a sentiment in a text on a per-sentence level.

[25]. This paper demonstrates state-of-the-art text sentiment analysis tools while developing a new time-series measure of economic sentiment derived from economic and financial newspaper articles from January 1980 to April 2015. We compare the predictive accuracy of a large set of sentiment analysis models using a sample of articles that have been rated by humans on a positivity/negativity scale. The results highlight the gains from combining existing lexicons and from accounting for negation. We also generate our own sentiment-scoring model, which includes a new lexicon built specifically to capture the sentiment in economic news articles. This model is shown to have better predictive accuracy than existing, "off-the-shelf", models. Lastly, we provide two applications to the economic research on sentiment. We estimate the impulse responses of macroeconomic variables to sentiment shocks, finding that positive sentiment shocks increase consumption, output, and interest rates and dampen inflation. Various kind of estimation techniques and scores have been allotted to estimate the scores of various newspapers.

[26]. Predicting the stock market remains a challenging task due to the numerous influencing factors such as investor sentiment, firm performance, economic factors and social media sentiments. However, the profitability and economic advantage associated with accurate prediction of stock price draw the interest of academicians, economic, and financial analyst into researching in this field. Despite the improvement in stock prediction accuracy, the literature argues that prediction accuracy can be further improved beyond its current measure by looking for newer information sources particularly on the Internet. Using web news, financial tweets posted on Twitter, Google trends and forum discussions, the current study examines the association between public sentiments and the predictability of future stock price movement using Artificial Neural Network (ANN). A sentiment analysis of news for predicting stock price movement using SVM enhanced with Particle Swarm Optimization (PSO) technique was proposed in. The study confirmed a correlation between web news and stock price movement volatility increased in strength during financial crisis.

[27]. Open-source software has become increasingly popular with companies looking to create business value through collaboration with distributed communities of organizations and software developers who rely on mailing lists to review code and share their feedback. OSS has undergone a transformation distancing itself from its free software antecedent, leveraging open-source communities to increase development productivity and increased functionality. The sentiment model used works by breaking an email body into lists of sentences, and further breaking these sentences into lists of words. Using the Natural Language Toolkit (NLTK) library, the words are tagged with parts of speech tags (subject, verb, noun etc.). The model tags each word into positive or negative words using an updated custom library with DPDK software terminology being added to the respective dictionaries. Finally, a score is generated based on the count of positive words and the count of negative words.

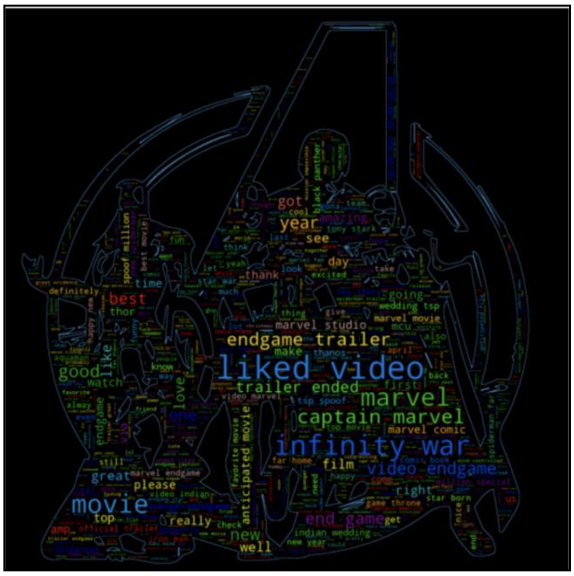
[28]. Drug reviews play a very significant role in providing crucial medical care information for both healthcare professionals and consumers. Customers are utilizing online review sites to voice opinions and express sentiments about experienced drugs. However, a potential buyer typically finds it very hard to go through all comments before making a purchase decision. Another big challenge is the unstructured and textual nature of the reviews, which makes it difficult for readers to classify comments into meaningful insights. For these reasons, this paper primarily aims to classify the side effect level and effectiveness level of prescribed drugs by using text analytics and predictive models within SAS® Enterprise. Additionally, the paper explores specific effectiveness and potential side effects of each prescription drug through sentiment analysis and text mining within SAS® Visual Text Analytics. These models are further validated by using a transfer learning algorithm to evaluate performance and generalization. The results can be used to develop practical guidelines and useful references to facilitate prospective patients in making better informed purchase decisions. Regression, Decision Tree and Neural Networks are been used.

[29]. In this paper, we present a general framework that uses natural language processing (NLP) techniques, including sentiment analysis, text data mining, and clustering techniques, to obtain new scores based on consumer sentiments for different product features. The main contribution of our proposal is the combination of price and the aforementioned scores to define a new global score for the product, which allows us to obtain a ranking according to product features. Furthermore, the products can be classified according to their positive, neutral, or negative features (visualized on dashboards), helping consumers with their sustainable purchasing behavior. After the experimentation, we could conclude that our work is able to improve recommender systems by using positive, neutral, and negative customer opinions and by classifying customers based on their comments.

[30]. Social sites for example Facebook and Twitter are that, where characters put their status or sentiments. People comment on their Facebook account concerning any correct subject of their consideration. Sentiment analysis can be seen as a utilization of content order. The primary occupation of content gathering is how to stamp writings with a predefined set of gatherings. Content gathering has been helpful in different zones for example, article ordering and content cleaning. Huge quantities of comments or surveys are posted by people in general every day. So, to distinguish the assessment of open towards a particular post is by physically analyze and discover each comment. People use very awkward words to express their feelings & most of the people use shortcuts e.g., ohm for awesome, lol for laughing out loud & many more, so this is sometime creating difficulty for the person who is not familiar with these words and cannot recognize the sentiments of the person. Process of data mining or knowledge discovery in database: Support vector machine, Decision tree, A decision tree, Regressions, Prediction.

IV. DATASET DESCRIPTION & SAMPLE DATA

According to the experiment performed in a sample data, we chose one of the data set and achieved the following description in the content of the movie review tweets.



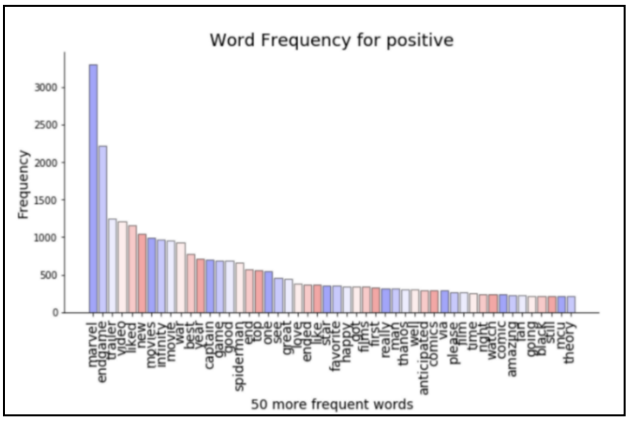
Word Cloud for positive review tweets

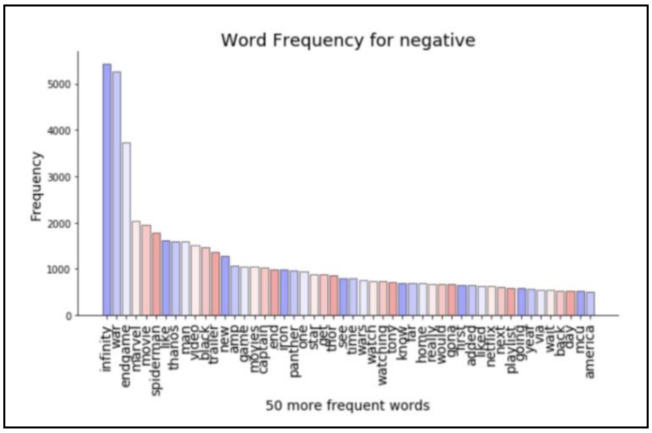
infinity war
movie
endgame
black panther
captain marvel
captain america
added video
age ultron
thano
liked video
video playlist
iron man
spiderman far
avengersendgame
endgame trailer
star war
hugh jackman
year
get
watch
first
need
going

Word Cloud for negative review tweets

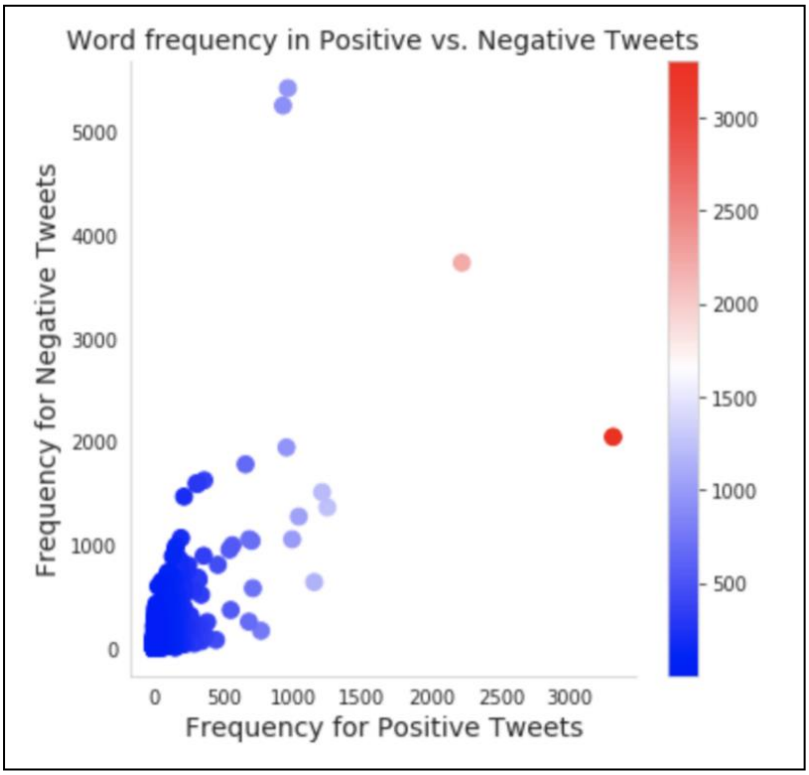
By using Word Cloud the representation of the data of tweets is presented of the form.
Python has a Word Cloud library that allows us to apply a mask using an image that we upload from our hard drive, select the background, the word colormap, the maximum words, font size, among other characteristics of the graph.

The dataset description represented via graphical way for the positive and negative word frequency brought a great idea about the experiment we were goanna make.

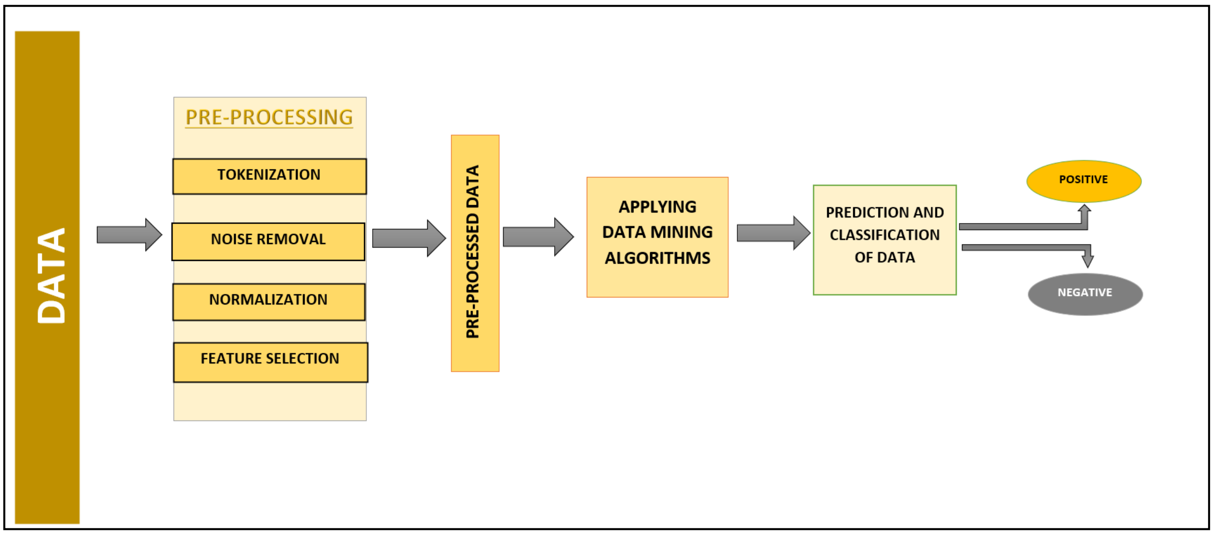




The data description via correlation between the frequency terms has been stated via the replot graph function in the python, on understanding the graph we come to know about the frequency for positive and negative tweet reviews.



V. PROPOSED ALGORITHM WITH FLOWCHART



In essence, the automatic approach involves supervised machine learning and text mining classification algorithms. The sentiment analysis is one of the more sophisticated examples of how to use classification to maximum effect through the text mining. In addition to that, unsupervised machine learning algorithms are used to explore data.

Overall, Sentiment analysis may involve the following types of classification algorithms:

- **Naive Bayes**

Naive Bayes is a fairly simple group of probabilistic algorithms that, for sentiment analysis classification, assigns a probability that a given word or phrase should be considered positive or negative.

Essentially, this is how Bayes' theorem works. The probability of A, if B is true, is equal to the probability of B, if A is true, times the probability of A being true, divided by the probability of B being true:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

The first step for performing classification is to understand the problem and identify the potential features and label. It has two phases, the learning phase and evaluation phase. In the learning phase, classifier trains its model on a given dataset and in the evaluation phase, it tests the classifier performance. Performance is evaluated on the basis of various parameters such as accuracy, error, precision, and recall.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

Naive Bayes classifier calculates the probability of an event in the following steps:

Step 1: Calculate the prior probability for given class labels

Step 2: Find Likelihood probability with each attribute for each class

Step 3: Put these values in Bayes Formula and calculate posterior probability.

Step 4: See which class has a higher probability, given the input belongs to the higher probability class.

Basically, Naive Bayes calculates words against each other. So, with machine learning models trained for word polarity, we can calculate the likelihood that a word, phrase, or text is positive or negative. When techniques like lemmatization, stop word removal, and TF-IDF are implemented, Naive Bayes becomes more and more predictively accurate.

- **Linear Regression**

Linear regression is a statistical algorithm used to predict a *Y* value, given *X* features. Using machine learning and opinion mining, the data sets are examined to show a relationship. The relationships are then placed along the *X/Y* axis, with a straight line running through them to predict further relationships.

Linear regression calculates how the *X* input (words and phrases) relates to the *Y* output (polarity). This determines where words and phrases fall on a scale of polarity from “really positive” to “really negative” and everywhere in between.

Linear regression consists of 3 stages majorly:

1. Analysing the correlation and directionality of the data
    - A scatter plot is drawn to analyse the data and check for directionality and correlation of data. The first scatter plot drawn indicates a positive relationship between the two variables. The data is fit to run a regression analysis.
    - It enables the researcher to formulate the model, i.e. that variable X has a causal influence on variable Y and that their relationship is linear.
2. Estimating the model, i.e., fitting the line
    - The second scatter plot is drawn to have an inverse U-shape this indicates that a regression line might not be the best way to explain the data, even if a correlation analysis establishes a positive link between the two variables.
    - The second step of regression analysis is to fit the regression line. Mathematically least square estimation is used to minimize the unexplained residual. The basic idea behind this concept is illustrated in the following graph. In our example we want to model the relationship between age and job satisfaction. The research team has gathered several observations of self-reported job satisfaction and the age of the participant
3. Evaluating the validity and usefulness of the model.
    - Most often data contains quite a large amount of variability in these cases it is up for decision how to best proceed with the data. This is analysed here in the last step of linear regression.
    - The last step for the linear regression analysis is the test of significance. Linear regression uses two tests to test whether the found

model and the estimated coefficients can be found in the general population the sample was drawn from.

- **Support Vector Machines**

Support vector machines are a set of supervised learning methods used for classification, regression, and outlier’s detection. All of these are common tasks in machine learning. You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model. There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC). The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible

A simple linear SVM classifier works by making a straight line between two classes. The data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This states that there can be an infinite number of lines to choose from. SVM tries to make a decision boundary in such a way that the separation between the two classes (that street) is as wide as possible. The decision boundary doesn't have to be a line. It's also referred to as a hyperplane because you can find the decision boundary with any number of features, not just two.

Sentimental Analysis using Logistic Regression, Support Vector Machine and Naive Bayes

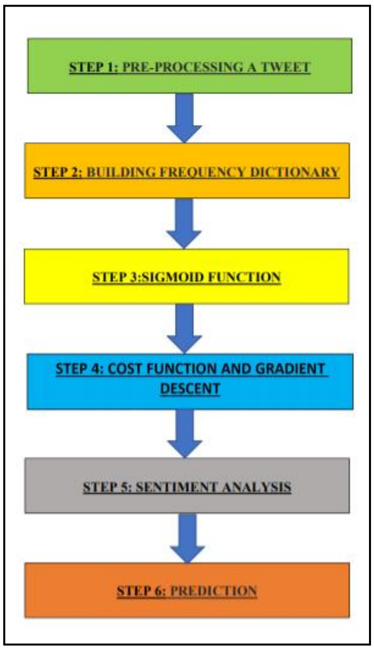
STEP 1: PRE-PROCESSING A TWEET

When pre-processing, you have to perform the following:

→Eliminate handles and URLs

→Tokenize the string into words.

→Remove stop words like “and, is, a, on, etc.”


STEP 1: PRE-PROCESSING A TWEET
STEP 2: BUILDING FREQUENCY DICTIONARY
STEP 3:SIGMOID FUNCTION
STEP 4: COST FUNCTION AND GRADIENT DESCENT
STEP 5: SENTIMENT ANALYSIS
STEP 6: PREDICTION

→Stemming- or convert every word to its stem. Like a dancer, dancing, danced, becomes ‘dance’. You can use porter stemmer to take care of this.

→Convert all your words to lower case

STEP 2: BUILDING FREQUENCY DICTIONARY

Now, we will create a function that will take tweets and their labels as input, go through every tweet, pre-process them, count the occurrence of every word in the data set and create a frequency dictionary.

The squeeze function is necessary or the list ends up with one element. The required functions for processing tweets are ready, now let's build our logistic regression model.

STEP 3: SIGMOID FUNCTION

Logistic regression makes use of the sigmoid function which outputs a probability between 0 and 1. The sigmoid function with some weight parameter θ and some input x^{(i)}x(i) is defined as follows: -

$h(x^{(i)}, \theta) = 1/(1 + e^{(-\theta^T * x^{(i)})}).$

STEP 4: COST FUNCTION AND GRADIENT DESCENT

The logistic regression cost function is defined as

$J(\theta)=(-1/m)*\sum i=1 \text{ to } m[y(i)\log(h(x(i),\theta)+(1-y(i))\log(1-h(x(i),\theta))]$

We aim to reduce cost by improving the theta using the following equation:

$\theta j:=\theta j-\alpha*\partial J(\theta)/\theta j$

On testing the model using the test data set we get an accuracy of 99.5%

STEP 5: SENTIMENT ANALYSIS USING NAIVE BAYES

Naive Bayes algorithm is based on the Bayes rule, which can be represented as follows:

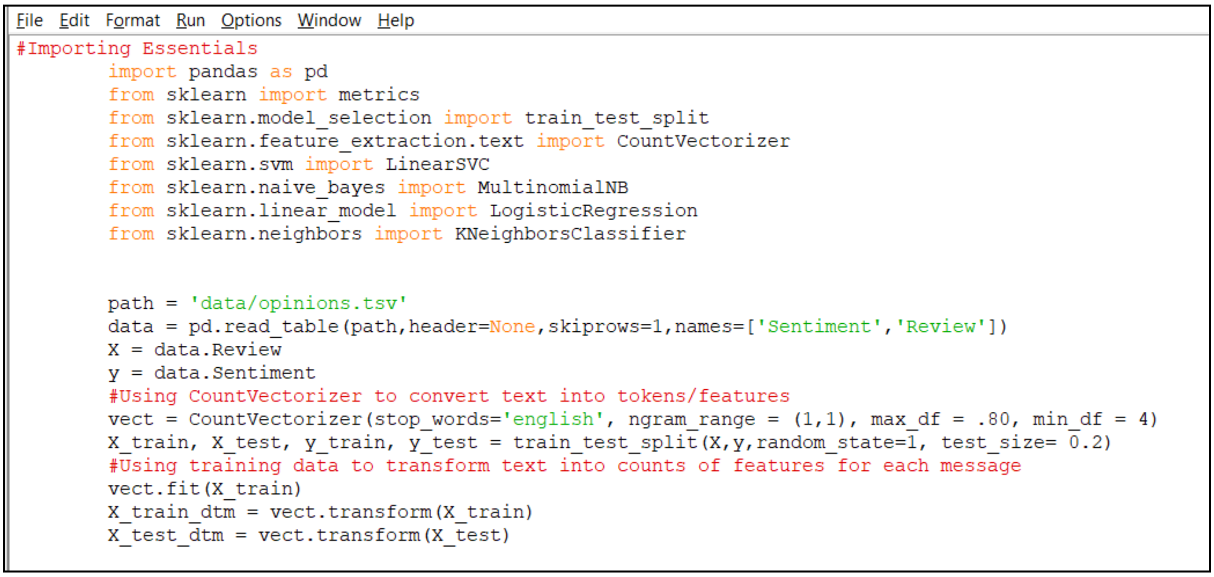
$P(X|Y) = P(Y) P (Y| X) P(X)$

STEP 6: PREDICTING USING NAIVE BAYES

In order to predict the sentiment of a tweet we simply have to sum up the loglikelihood of the words in the tweet along with the log prior. If the value is positive then the tweet shows positive sentiment but if the value is negative then the tweet shows negative sentiment.

VI. EXPERIMENTS RESULTS

Sentimental Analysis can be implemented using various sources like Machine learning and upcoming algorithms like Naive Bayes, linear regression, etc. Here firstly we need to check out the essential packages or files inbuilt in language we using for coding. Sentimental analysis for movie review using machine learning algorithms and python language is described below.



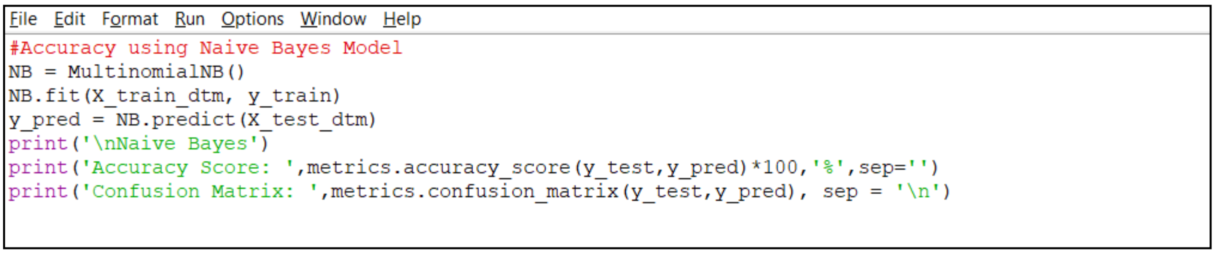
```
#Importing Essentials
        import pandas as pd
        from sklearn import metrics
        from sklearn.model_selection import train_test_split
        from sklearn.feature_extraction.text import CountVectorizer
        from sklearn.svm import LinearSVC
        from sklearn.naive_bayes import MultinomialNB
        from sklearn.linear_model import LogisticRegression
        from sklearn.neighbors import KNeighborsClassifier


        path = 'data/opinions.tsv'
        data = pd.read_table(path,header=None,skiprows=1,names=['Sentiment','Review'])
        X = data.Review
        y = data.Sentiment
        #Using CountVectorizer to convert text into tokens/features
        vect = CountVectorizer(stop_words='english', ngram_range = (1,1), max_df = .80, min_df = 4)
        X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size= 0.2)
        #Using training data to transform text into counts of features for each message
        vect.fit(X_train)
        X_train_dtm = vect.transform(X_train)
        X_test_dtm = vect.transform(X_test)
```

Firstly, we are importing the essential packages from the python editor. Packages like pandas, metrices, count vectors, etc. are implemented at the very first step for coding.

In the very first part we use CountVectorizer for conversion of the texts into tokens

The very next step after this is transforming texts into counts of features for every single message using training data.

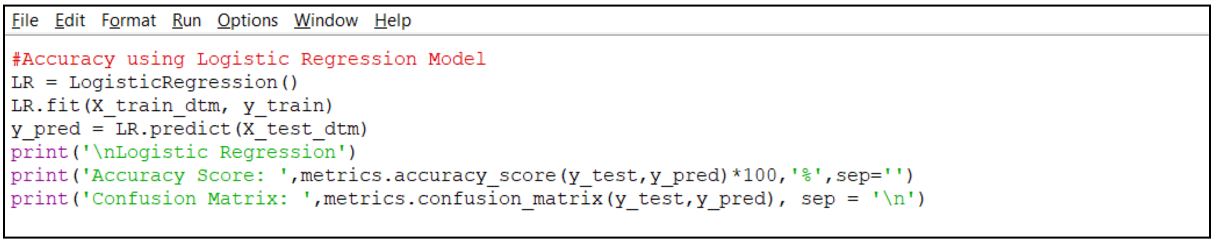
The result we want is the score of positive and negative reviews or tweets. For this we need to firstly look over the accuracy and confusion matrix (A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if we have an unequal number of observations in each class) Thus we can achieve this via various algorithms. Each of them code is stated below.



```
#Accuracy using Naive Bayes Model
NB = MultinomialNB()
NB.fit(X_train_dtm, y_train)
y_pred = NB.predict(X_test_dtm)
print('\nNaive Bayes')
print('Accuracy Score: ',metrics.accuracy_score(y_test,y_pred)*100,'%',sep='')
print('Confusion Matrix: ',metrics.confusion_matrix(y_test,y_pred), sep = '\n')
```

The above code deals with the bringing out the confusion matrix and accuracy score via Naive Bayes.

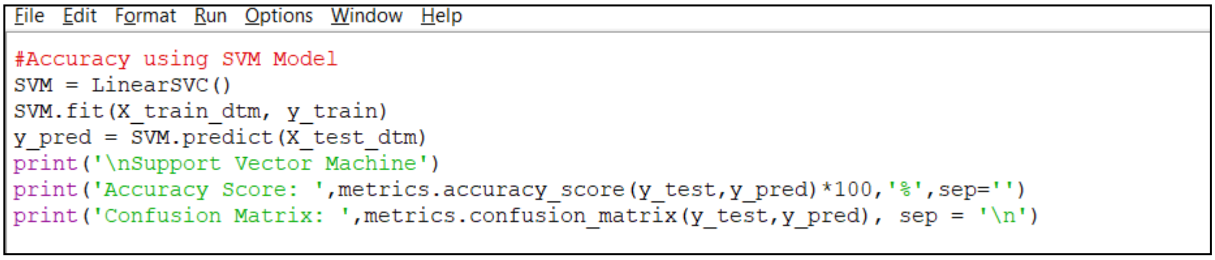
The function we use here is Multinomial Naïve Bayes which in python has its syntax as var=MultinomialNB()



```
#Accuracy using Logistic Regression Model
LR = LogisticRegression()
LR.fit(X_train_dtm, y_train)
y_pred = LR.predict(X_test_dtm)
print('\nLogistic Regression')
print('Accuracy Score: ',metrics.accuracy_score(y_test,y_pred)*100,'%',sep='')
print('Confusion Matrix: ',metrics.confusion_matrix(y_test,y_pred), sep = '\n')
```

The above code deals with the bringing out the confusion matrix and accuracy score via Logistic Regression Model.

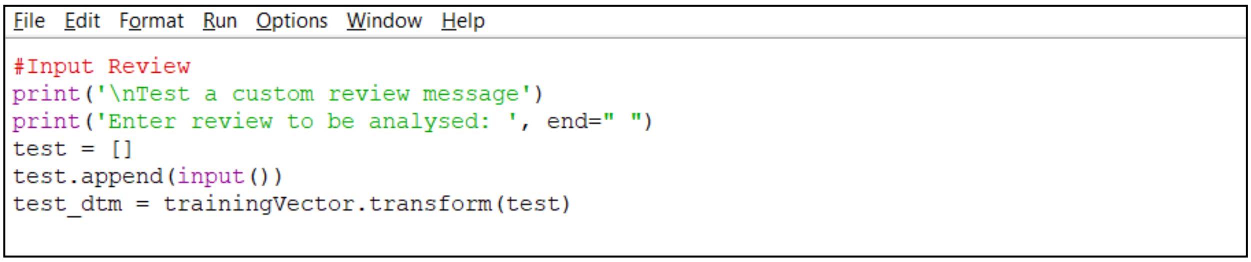
The function we use here is Regression which in python has its syntax as var=LogisticRegression()



```
#Accuracy using SVM Model
SVM = LinearSVC()
SVM.fit(X_train_dtm, y_train)
y_pred = SVM.predict(X_test_dtm)
print('\nSupport Vector Machine')
print('Accuracy Score: ',metrics.accuracy_score(y_test,y_pred)*100,'%',sep='')
print('Confusion Matrix: ',metrics.confusion_matrix(y_test,y_pred), sep = '\n')
```

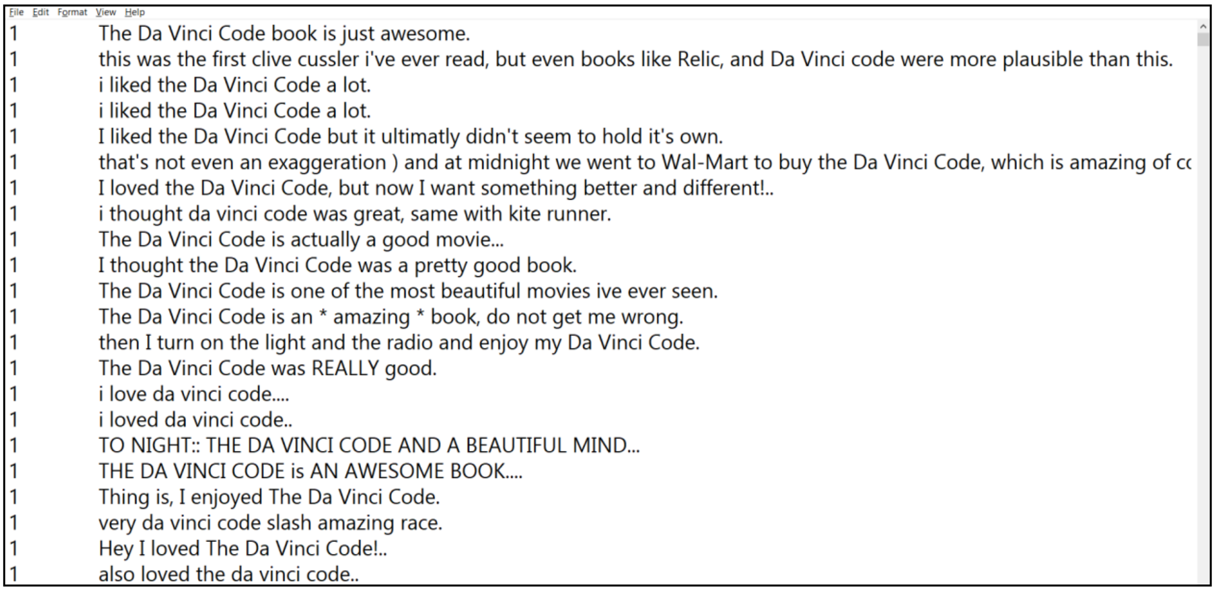
The above code deals with the bringing out the confusion matrix and accuracy score via Logistic Regression Model.

The function we use here is Support Vector Machine which in python has its syntax as var=LinearSVC()



```
#Input Review
print('\nTest a custom review message')
print('Enter review to be analysed: ', end=" ")
test = []
test.append(input())
test_dtm = trainingVector.transform(test)
```

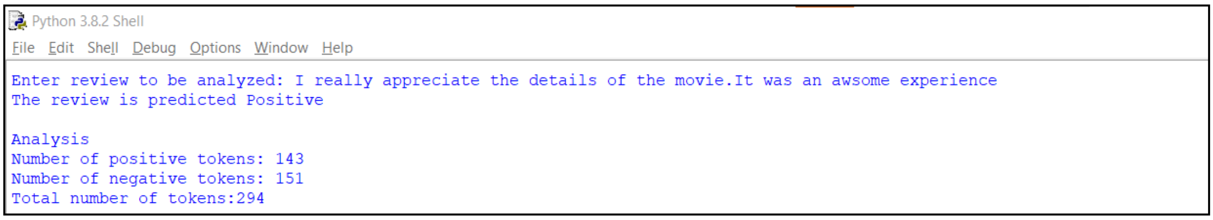
The input is taken as simple way it can be, because we have to store the large amount of raw data from our side before itself. Thus, the input and the data stored with us, it has similarity then it could make the process faster. Thus, the test variable is taken as array of words which is then checked one by one using the analysis code.



```
1	The Da Vinci Code book is just awesome.
1	this was the first clive cussler i've ever read, but even books like Relic, and Da Vinci code were more plausible than this.
1	i liked the Da Vinci Code a lot.
1	i liked the Da Vinci Code a lot.
1	I liked the Da Vinci Code but it ultimatly didn't seem to hold it's own.
1	that's not even an exaggeration ) and at midnight we went to Wal-Mart to buy the Da Vinci Code, which is amazing of c
1	I loved the Da Vinci Code, but now I want something better and different!..
1	i thought da vinci code was great, same with kite runner.
1	The Da Vinci Code is actually a good movie...
1	I thought the Da Vinci Code was a pretty good book.
1	The Da Vinci Code is one of the most beautiful movies ive ever seen.
1	The Da Vinci Code is an * amazing * book, do not get me wrong.
1	then I turn on the light and the radio and enjoy my Da Vinci Code.
1	The Da Vinci Code was REALLY good.
1	i love da vinci code....
1	i loved da vinci code..
1	TO NIGHT:: THE DA VINCI CODE AND A BEAUTIFUL MIND...
1	THE DA VINCI CODE is AN AWESOME BOOK....
1	Thing is, I enjoyed The Da Vinci Code.
1	very da vinci code slash amazing race.
1	Hey I loved The Da Vinci Code!..
1	also loved the da vinci code..
```

The task we perform for our better results is that we collect large amount of raw data for our experimental analysis. Here we collect the data in very common way like storing it in notepad and saved the file with tsv extension in the system where the access to code is easy. Our data has both the negative and positive messages such that the algorithm techniques we using gives out good accuracy.

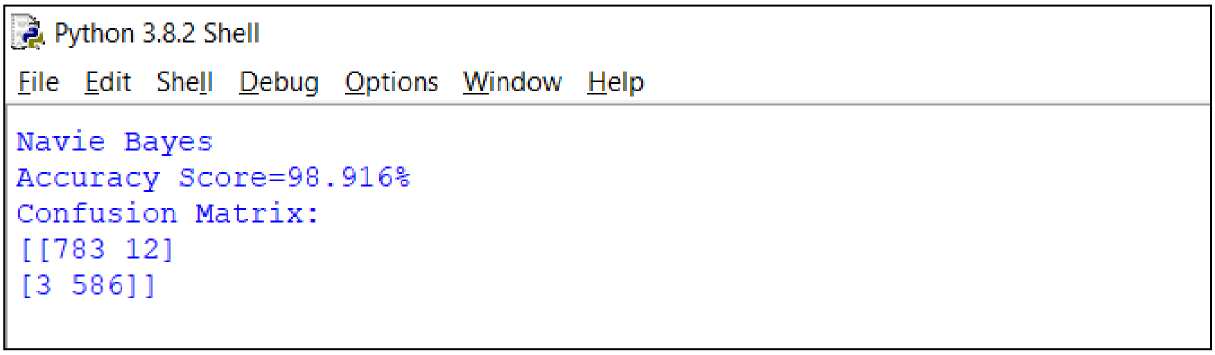
To the custom input review the total number of positive and negative analysis tokens and a predicted phrase comes as output with the accuracy and confusion matrices of each method separately.



```
Enter review to be analyzed: I really appreciate the details of the movie.It was an awsome experience
The review is predicted Positive

Analysis
Number of positive tokens: 143
Number of negative tokens: 151
Total number of tokens:294
```

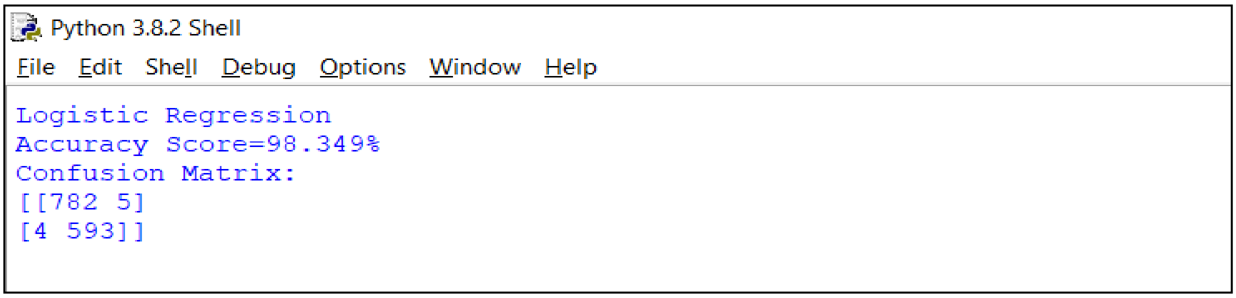
The individual output results to each of the Naive Bayes, Logistic regression and Support Vector Machine methods to give out accuracy and confusion matrices. The values vary from one to other as the procedure and methodology varies. The output for the individual algorithm is screenshotted and well displayed.

NAVIE BAYES:



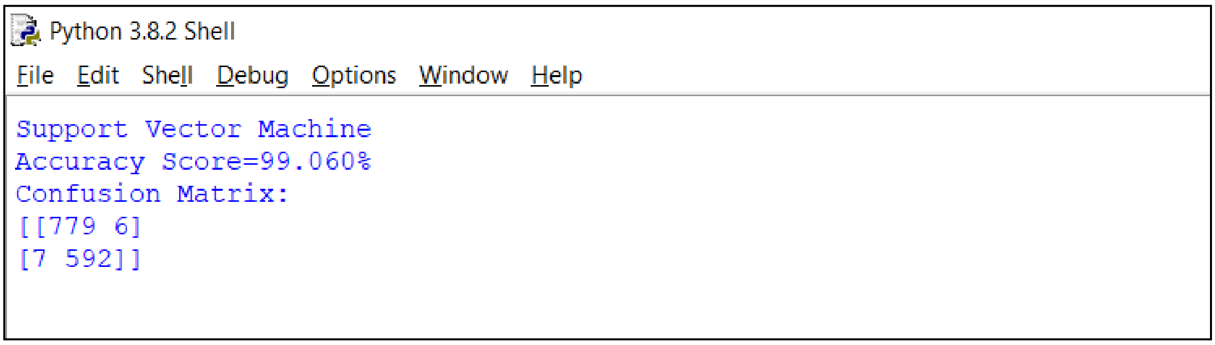
```
Navie Bayes
Accuracy Score=98.916%
Confusion Matrix:
[[783 12]
[3 586]]
```

LOGISTIC REGRESSION:



```
Logistic Regression
Accuracy Score=98.349%
Confusion Matrix:
[[782 5]
[4 593]]
```

SUPPORT VECTOR MACHINE:



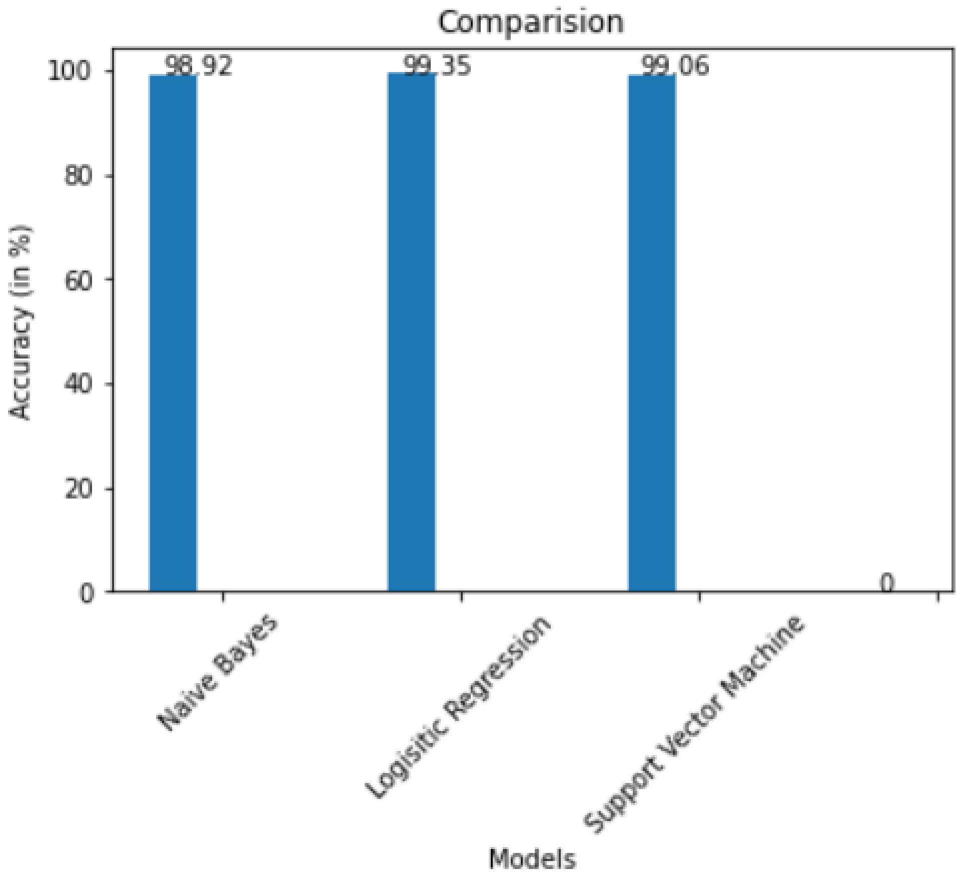
```
Support Vector Machine
Accuracy Score=99.060%
Confusion Matrix:
[[779 6]
[7 592]]
```

VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

Here we compare the whole project in two terms.
First one can be the comparing the accuracy percentage of the models we have chosen. A table is drawn manually to get clarity about the graph that is drawn via code.

MODEL	ACCUARCY IN %
Naïve Bayes	98.916
Logistic Regression	98.349
Support Vector Machine	99.060

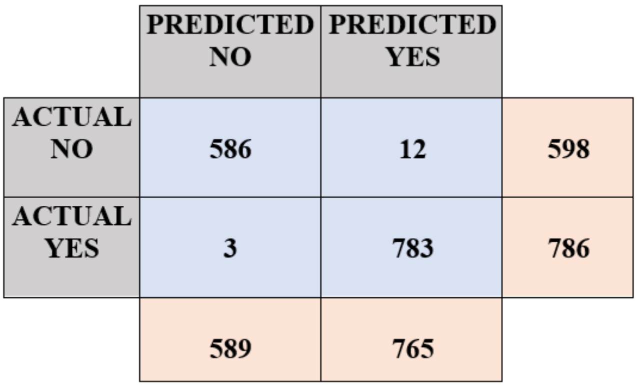
TABLE SHOWING ACCURACCY PERCENTAGE FOR VARIOUS MODELS



GRAPH SHOWING ACCURACCY PERCENTAGE FOR VARIOUS MODELS

The comparison between different model and accuracy level gave out the conclusion that the accuracy level for **Logistic regression** is much better than the other two. So, for training in future we preferred for using Support Vector Machine.

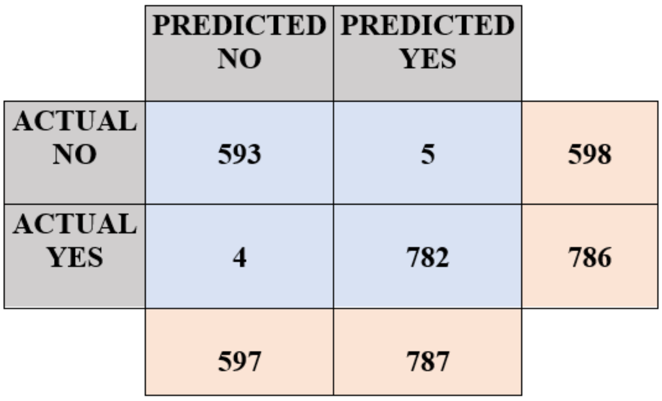
The output for the code gives out accuracy percentage and confusion matrices. The diagrammatic representation of confusion matrices for each model is shown below.



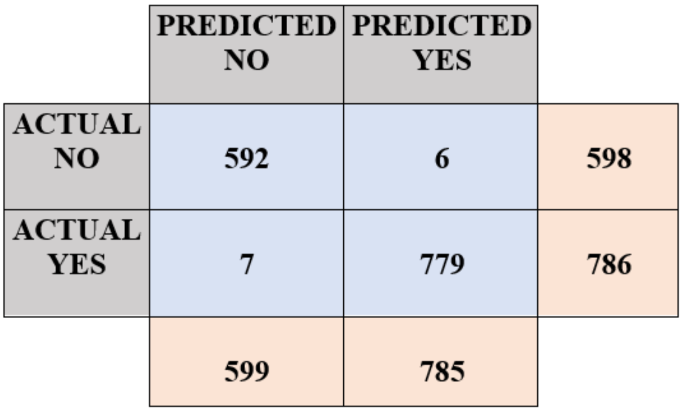
	PREDICTED NO	PREDICTED YES	
ACTUAL NO	586	12	598
ACTUAL YES	3	783	786
	589	765	

DIAGRAMTEIC REPRESANATION OF CONFUSION MATRIX FOR NAÏVE BAYES

DIAGRAMTEIC REPRESANATION OF CONFUSION MATRIX FOR LOGISTIC REGRRSSION



	PREDICTED NO	PREDICTED YES	
ACTUAL NO	593	5	598
ACTUAL YES	4	782	786
	597	787	



	PREDICTED NO	PREDICTED YES	
ACTUAL NO	592	6	598
ACTUAL YES	7	779	786
	599	785	

DIAGRAMTEIC REPRESANATION OF CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE

We choose various different phrases as input for the testing and bring out the output using Support Vector Machine code and get the tabular format as follows with the graphical comparison.

Let the phrases be as

Phrase 1: It was a amazing movie. I liked it a lot

Phrase 2: It was a boring one.

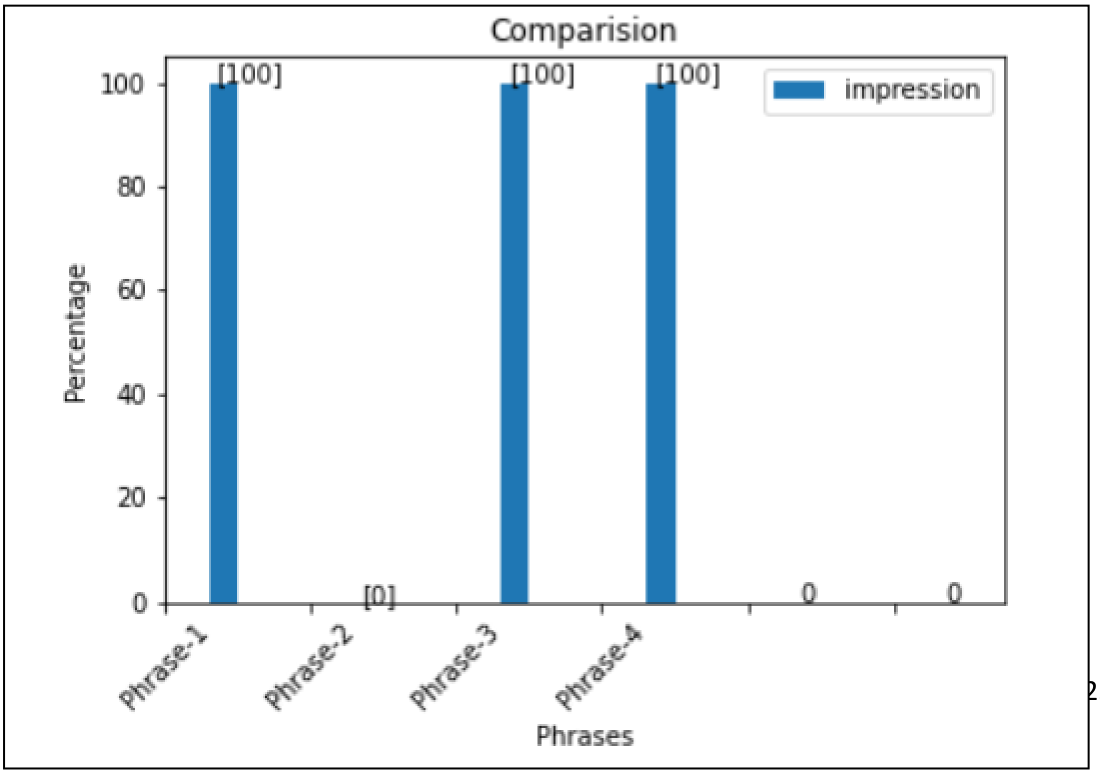
Phrase 3: It was fabulous. Must watch

Phrase 4: It was average one.

REVIEW PHRASES	**PREDICTION (100-GOOD /0-BAD)**
PHRASE 1	100
PHRASE 2	0
PHRASE 3	100
PHRASE 4	0

TABLE SHOWING PREDICTAION PERCENTAGE FOR VARIOUS PHRASES

GRAPH SHOWING PREDICTAION PERCENTAGE FOR VARIOUS PHRASES



VIII. CONCLUSION AND FUTURE WORK

Here first we concluded that Logistic Regression give better accuracy compared to Support Vector Machine and Naïve Bayes and also, it’s very less time taking compared to other two. So, for actual prediction purpose we used Logistic regression algorithm for training our model. We even looked after the confusion matrices and analysed the actual and predicted values. After that we predict the impression of different phrase reviews and checked our output with the online analyser websites like imdb so finally concluded that our model working properly. In future we can use also some advanced techniques like RNN (Recurrent Neural Networks).