VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: M.TECH5 - Software Engineering

Course: SWE1017- Natural Language Processing

Slot: A2

Faculty: Prof. M. PREMALATHA

Component: J – Component

Title: SENTIMENT ANLAYSIS ON MOVIE REVIEWS

Team Member(s):

19MIS1018 – B DEVI PRASAD

19MIS1099 – K CHAITANYA

19MIS1129 – V SANTHOSH

1. ABSTRACT

In this project we planned to do Sentiment analysis on Movie reviews. Sentiment analysis is focused on the extraction of opinions of the people towards a particular topic from a textual data.In this project we try to focus on sentiment analysis on Rotten Tomatoes movie review database. We examine the sentiment expression to classify the polarity of the moviereview on a scale of 0(highly disliked) to 4(highly liked) and perform feature extraction and ranking and use these features to train our multi- label classifier to classify the movie review into its correct label. We are going to implement some of the Machine Learning Algorithms in order tocompare the final Accuracy results In addition, a comparative study on different classification approaches has been performed to determine the most suitable classifier to suit our problem domain.

2. INTRODUCTION

Movie reviews are an important way to gauge the performance of a movie. While providing a numerical/stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews iswhat gives us a deeper qualitative insight on different aspects of the movie.A textual movie review tells us about the strong and weak points of the movie and deeper analysis of a movie review can tell us if the movie in general meets the expectations of the reviewer.

Sentiment Analysis is a major subject in machine learning which aims to extract subjective information from the textual reviews. The field of sentiment of analysis is closely tied to natural language processing and textmining. It can be used to determine the attitude of the reviewer with respect to various topics or the overall polarity of review. Using sentiment analysis, we can find the state of mind of the reviewer while providing the review and understand if the person was “happy”, “sad”, “angry” and so on.In this project we aim to use Sentiment Analysis on a set of movie reviews given by reviewers and try to understand what their overall reaction to themovie was, i.e. if they liked the movie or they hated it. We aim to utilize therelationships of the words in the review to predict the overall polarity of the review. The main objective of this project is to predict the movie is Good or not based on comments from Rotten Tomatoes using Sentiment analysis. This project helps us to predict whether the users liked the movie or not and thewords used in the phrase falls under positive or negative comment.

3. PROBLEM STATEMENT

The main goal is to classify the sentiment of reviews from the Rotten Tomatoes dataset. The Rotten Tomatoes movie review dataset is a corpusof movie reviews used for sentiment analysis. The main task is to label phrases on a scale of five values: negative, somewhat negative, neutral, somewhat positive, positive. There are many obstacles such as sentence negation, sarcasm, language ambiguity, and many others make the sentiment prediction more difficult. In general, this particular Sentiment Analysis is a multiclass classification task to be faced.

4. Literature Review

Paper 1:

Atiqur Rahman, Md. Sharif Hossen(2019) , Sentiment Analysis on Movie Review Data Using Machine Learning Approach,. 019 International Conference on Bangla Speech and Language Processing (ICBSLP) 978-1-7281-5241-7/20/$31.00 ©2019 IEEE. Explanation: Now a day's very person is using internet for very purpose to book their movie tickets, to purchase their things in online (Clothes, Grocery, Books, Electrical products, etc..).To buy all these things first the people will see the before reviews of the same products buyed by other people. And after the buying the products also the people will share their opinion using very social media platform about the movie or product they buyed. Sentiment analysis is the process of getting valuable fact from a big set of data. Sentiment analysis is the most used topic for the purpose to assist one to get a feedback from a large dataset. It mainly centres a systematic search for the truth or facts about the feelings from the text patterns. And it automatically characterizes the expression of feelings, Negative, Positive, or Somewhat about the existence of anything. Various sources like Social media, newspaper, Movie reviews and Surveys can be used in the data analysis. So this paper particularly focused on the movie reviews data to analyse this data five machine learning classifiers are used. They are Naïve Bayes, Decision tree, Support vector machine, Maximum entropy and Multinomial naïve bayes. This analysis many achieve on the better accuracy, precision, F- score , recall and better accuracy than previous experiment over this classifiers. The paper mainly focused on the opinion of Movie reviews data. The sentiment analysis was handled by NLP using several levels. Here the authors mainly focused on the creation of sentiment lexicon by word vector representation. The sentiment analysis can be done on combined techniques(Training and testing data). The 2 methods where supervised and unsupervised methods. It has the several algorithms to conduct the classification techniques based on the trained data. Sentiment analysis classifies the people opinion as a positive or negative view. According to sentiment analysis techniques are sentence, Document, Aspect and user based. Here, the sentiment analysis identify orientation of opinion in a piece of text.[For example, The movie as fabulous, the Movie stars Mr.X, The Movie was horrible] can be generalized to a wider set of emotions. The Authors here described three technique the first technique detects the sentiment of ach sentiment of each sentence as positive or negative. The second technique detects the sentiments of full document as a single unit. The last technique focuses on the properties of an entity. Machine learning and lexicon based techniques are the most common in sentiment analysis where the first uses training and testing data set to categorize data. Here, 2 kinds of lexicons are available, first one is corpus based lexicon and second on is dictionary based lexicon. To get outcome in a context oriented we have to use corpus based lexicon like as SenticNet. Where the small set of counsel words are culled manually with acquainted orientations in dictionary based approach. The Analysis procedure followed is Input data and then pre- process the data after the data pre-processing the method will be followed Test processing which has 6 steps [URL punctuation and removal, Bracket and Number removal, Tokenization, stop words removal, case conversion, and Stemming] after the text processing the Create Features vectors, classification, and Result will be implemented. These are the steps for classification of Sentiment analysis. In each step of process we can see the sentiment classification. In URL removal where url is sharing then first we should remove url from the text. Bracket and number removal brackets and numbers have no meaning in the sentiment classification. So we have to remove the URL from the text. In Tokenization we have to divide our textual data into smaller components this provides turn text into sentences and sentences into words. Removing punctuation like semi –colon, colon, follies-stop and Question mark. Case conversion this helps to remove the distinction between two same words. Omitting stop words such as it, i, you, a, an, the because they won't have any meaning in sentiment classification. And Last one is stemming in stemming we have to reform the inflected words and removes derivational affixes from a word. After the implementation of all ML algorithms the performance statistics of several classifiers the best algorithm that shows the quality of features of selected for movie review data. By seeing results very classifier is sensitive, Result that shows the multinomial Naïve bayes classifier is better than all other 4 algorithms. Multinomial Naïve bayes contains everything shows high in large datasets. Accuracy – 88.50%, Precision – 92.94%, Recall – 83.33%, FScore – 87.87%. Sentiment analysis is very essential to understand the expression of feelings about anything like product, social media, movie reviews. It can implemented by lexicon and machine learning approaches. Machine learning is more efficient where it requires more labelled data. where lexicon can fail to calculate the score of expression where if it can found the word in the dictionary. Polarity classification can be used in ML approach on the movie review data. Where this method divides datasets into two parts one training and other one is testing. Before this process collect dataset on the suitable platform on movie reviews. Then use the Machine learning approach with the classification algorithms to find the best approach with accuracy, precision, recall, and F-score should be high among all algorithms, By this approach we can find the Multinomial Naïve bayes is better than others. Authors mainly discussed on the how this sentiment analysis will work on the movie reviews with the different approaches (ML and Lexicons). And using these classifiers we can find which algorithm will be best for the chosen dataset. By this analysis we can analyse the movie review data. This point we can learn the strong and weak point of the movie. For example some users we will see the IMDB ratting on the particular movie then some people watch movie based on that ratting then if the movie is not upto their expectations then will give feedback and comments on the movie. So, this approach helps us to find the strong and weak point.

Paper 2:

Priya Patel, Devkishan Patel, and Chandani Naik, Sentiment Analysis on Movie Review Using Deep Learning RNN Method (2020) ,. Springer Nature Singapore Pte Ltd. 2021.
Explanation: In this paper we will see about how Sentiment analysis will work on movie reviews using Deep learning RNN method. The usage of internet has growing day by day, Where mainly the social media grows instantly because the it easy to use for every one and using this social media platform we can connect all over the world this social media platforms are becoming more popular now a days, The social media sites are sentiments which we can drive meaningful information. This sentiment analysis is use for the finding the correct feedback and relevant solutions. For movie reviews the sentiment analysis classification is useful to analyse the information in the form of the reviews and opinions in two possible ways positive and negative. The authors in this paper have approached to the deep learning- based classification algorithm RNN. Here authors have used Recurrent neural networks algorithms instead of Machine learning because of Recurrent neural networks works on the multilayer that gives the better output

as comparative to the Machine learning algorithms. In this Paper the authors have used Sentiment analysis in three different levels(Sentence level, Document level, Aspect Level). Sentence level- Two classification types are used objective and subjective. Where each sentence is classified into negative and positive., Document level- Here, the Document will classifies into two classes positive or negative class., Aspect Level- As per the authors it also known as feature level. Where users often express opinions about multiple aspect in the same sentences. The Sentiment analysis is field of study that analyses people's opinions, sentiments, attitudes, and emotions towards entities such as products, services, organization, events. The aim of sentiment analysis is to automatically extract options expressed in the user-generated content. In sentiment analysis we will be handling emotion, feelings of both positive and negative or normal. The authors had used document-level sentiment analysis classification for getting the proposed system results. The authors of enhanced thee RNN language protocol, that is forward LSTM, which effectively covers all past system information and achieves better results than the conventional. The sequential connection among the words that the study of text sentiment analysis. Mainly sentiment classification has been explained by the classification algorithms such as logistic regression ,support vector machines and naïve bayes classifier. Where Deep learning method will also be applied for sentiment analysis. In these approach the authors used the Indian movie dataset. And then they have applied pre-processing on the dataset and also performed the removal of hash, tags, synonyms, acronyms, etc… Here they using long short –term memory to modify version of RNN with word vector features. For Sentiment analysis on movie review data using different sources, like BookMyShow, Netflix, and IMDB. After this preprocessing such as removal of HMTL tags, punctuation, and number and also used removal of stop words. Here RNN with LSTM can be viewed as an enhanced model for the RNN conventional language model from the perception of language. Here, the authors proposed approach is First they will import Movie Review Dataset then it goes to the next phase pre-processing where we should remove tokenization, stemming and stop word, then Feature Extraction to Feature Selection then they have applied the RNN algorithm in, RNN we have Random weight-> result-> Compare with expected result -> Should Repeat step 1-3 until the variable are not defined then after applying the opinion classification then at last phase we will get the labelled Data. To implement the approach authors mentioned that system should have python latest version and 8Gb ram. Data used IMDB movie review dataset. Here the results obtained for TF_IDF is accuracy for 1 Epoch is 50.16 % , and for 3 Epochs 87.64%. Then for word2Vec the accuracy is 94.61 % and for 3Epochs is 94.06. They obtained more accuracy in 1 Epoch because the dataset are over fit with 3 epochs and that is reason it gave us less accuracy with 3 epochs. The Results tell us that using word2vec with RNN we can get better accuracy even on a large amount of training and testing dataset compared to machine learning method such as Naïve bayes and support vector machine. As more number of people the efficient review related to the movie. It can also be useful in the future, we can work on the real-time data with the use of machine learning. To maximize the performance we can also use the deep learning bi-LSTM method and use different model of combinations. The movie review dataset is used for the preparation of pang and lee from IMDB and A famous deep learning framework provided by socher and discussed.

Paper 3:

Dilip Singh Sisodia, Shivangi Bhandari, Nerella Keerthana Reddy and Abinash Pujahari, A Comparative Performance Study of Machine Learning Algorithms for Sentiment Analysis of Movie Viewers Using Open Reviews,. © Springer Nature Singapore Pte Ltd. 2020 Explanation: Here Authors are tend to take others opinions before implementing the even trivial matters like shopping , movie. But these days the internet the is one and only source to the users to gather the important information for their work to implement. So, now a days movie and series are the favourite pass time to the youth now a days. In this modern world the people like to give their feedback and opinion only movies. We can find these reviews mainly in social media platform such as instagram, Facebook, and twitter this provides an feedback to the users to review the feedback before watching the movie. The reviews are not done by the films actors or producers these are mainly done mainly the people who will spend their maximum on the internet. Mainly on the first day of the movie the people watched the movie will post their reviews on the social media. So here the other users will see feedback like it positive or negative feedback and then they will make decision on it. Here the sentiment the analysis will do the analysis, summarize and classify the data for the users. The main goal of sentiment analysis may be different levels like at the sentence level, word level, document level and aspect level. While considering the sentiment analysis it determines the opinion for independent sentence or review. However, both of these analysis fail to find the people opinion for example like if people liked the movie or not. The dataset used in these approach is around 25,000 and 1000 reviews respectively of different movies given on the internet. Here the inputs are used for various machine learning algorithms. These Machine learning algorithms work for the some of the selected features of the data which are extracted using NLP techniques. The authors here compared the performance of different machine learning algorithms for the selected features. Sentiment analysis is the only process of analysing of views of persons toward any service. The perspective of sentiment analysis although subjective analysis system is responsible for the sentiment analysis performance increase and identifying the sentence of any person is a difficult task. Here, the authors used two classifiers the first classifier searches for the subjective sentence in the data set and the other searches for the objective sentence. To implement and validate the machine learning techniques for the sentiment analysis, the data set is obtained and these are described in the detail. Then they found the accuracy of the built model. And the summary of the approach. First to classify the sentiment analysis should the perform the Classification of Sentiment analysis. The Analysis procedure followed is Input data and then pre- process the data after the data pre-processing the method will be followed Test processing which has 6 steps [URL punctuation and removal, Bracket and Number removal, Tokenization, stop words removal, case conversion, and Stemming] after the text processing the Create Features vectors, classification, and Result will be implemented. These are the steps for classification of Sentiment analysis. In each step of process we can see the sentiment classification. In URL removal where url is sharing then first we should remove url from the text. Bracket and number removal brackets and numbers have no meaning in the sentiment classification. So we have to remove the URL from the text. In Tokenization we have to divide our textual data into smaller components this provides turn text into sentences and sentences into words. Removing punctuation like semi –colon, colon, follies-stop and Question mark. Case conversion this helps to remove the distinction between two same words. Omitting stop words such as it, i, you, a, an, the because they won’t have any meaning in sentiment classification. And Last one is stemming in stemming we have to reform the inflected words and removes derivational affixes from a word. The data set extracted from the kaggle and it consists of 25,000 revies. In the dataset the data stored is reviews of different films in different languages in the 2013-2016. Then both are divided for training and testing sets. Then after performing the pre-processing, first data set now it will consist of 20,000 reviews. Here in machine learning the authors have used classification algorithms and with the different feature extraction techniques. By using the data set they have validated both the data set using different classifiers and features using the weka tool. With the ML algorithms they are naïve bayes , Random forest, Support vector machine, Decision tree, MNB , BNB, AdaBoost and bagging. After the performance matrix performed the bagging classifier gives the good results on using the bigrams. In these paper the authors main aim is to find the best classifier to test the movie reviews given by the people so that we would know the overall general opinion of the audience. The classifiers are used to validate using the same data of metrics. The support vector machine classifier outperforms all than the all other classifiers in the set of individual classifiers. While on the ensemble classifiers, the random forest classifier out performers the remaining ones. When compared between individual and ensemble classifiers, the individual classifiers perform the better. Here the authors have implemented the 2nd dataset with same machine learning algorithms with 1000 reviewed data. Here also the Support vector machine performs well then thee other classifiers the random forest classifier out performers the remaining ones. Then at last better features can be extracted which are less in number, yet give acceptable results.

Paper-4

Sentiment Analysis of IMDB Movie Reviews using long short-term Memory November 2020978-1-7281-5467-1/20/$31.00 ©2020 IEEE Conference: 2020 2nd International Conference on Computer and Information Sciences (ICCIS),Author: Saeed Mian Qaisar Effat University Explanation: The sentiment analysis is the process of natural language processing and computational linguistics to extract them, Identify them and categorize by different opinions expressed in the sequence of their analysis. It is mainly a classification problem that merge into both domains of NLP and ML. This has a wide range of opinion mining and business analytics which is potential to implement for governmental purposes which prevents suicide incidents. Lexicon and Machine learning are two fundamental approaches used for this Sentiment analysis. The sentiment analysis is an emerging research area where vast amount of data are being analysed, to generate useful insights in regards to a specific topic. It is a tool which provide service to the governments, corporations and even consumers. Text emotion recognizing lays a key role in this framework. Researchers in the fields of natural language processing (NLP) and machine learning (ML) have explored a variety of methods to implement the process with highest accuracy possible. In this paper the Long Short-Term Memory (LSTM) classifier is used for analysing sentiments of the IMDb movie reviews. It is based on the Recurrent Neural Network (RNN) algorithm. The data is effectively pre-processed and partitioned to enhance the post classification performance. In this paper, datasets are been used for analysing the reviews of movies where there is a python based application used to analyse IMDB dataset which the dataset is divided into two parts as it is trained and tested by the system. For training we use to prepare the LSTM classifier and for the testing used to quantify the classification precision here, the confusion matrix and accuracy results are displayed by the proposed ML models for text based sentiment analysis where the model consists of five main blocks with some minor components integrated in the system. As the dataset contains about 50k movie reviews from IMDB where it is been split into 25k reviews for training purposes while the other half part of 25k is intended for testing the classifiers. By the splitting both sets contain12.5k positive and 12.5k negative ratings which it allows the viewers to rate on a scale from 1-10 and according to the dataset any review which is <-4 stars are labelled as negative review and review which is >-7 stars are labelled as positive. There will be more than 30 reviews for each movie of their average number that dataset contains different words. It will be common to divide the dataset into training and testing vectors where the training is the set of data that trains the particular classifier also the validation vector is a subset of training that does not necessarily train but also is used to give some insights on the performance of the classifier and the test data will be evaluated by the accuracy of the particular models. The recurrent ratio is encountered in various Machine Learning algorithms at 80-20 separating which gives 80% to the training and 20% for testing purpose. while dealing with the IMDb intended database portion, the confusion matrix is generated to assess the performance accuracy of the LSTM classifier. Confusion matrix provides a summary of correct and incorrect predictions This ratio gives by doing the data division of the models provided by datasets. In the preprocessing of the data the cleaning of data is incomplete and organized leads to identify the fake reviews. It is a crucial task in doing data mining process which refers to cleaning up the data from unwanted information which helps in training process which confuses the classification process. By text embedding which is a process of extracting the features from text and pass it as an input in the classifier where each movie reviews are encoded or vectorized into a numeric values which is done by utilizing the genism. Based on Recurrent Neural Network(RNN) machine learning algorithm which has memory but falls short when the data has more dependencies on it. LSTM is used to maintain a level of relevance and permits data from vanishing them elegantly in RNN model using linear memory units and control long range temporal dependencies. The classification accuracy is used to measure that how well, the devised model is able to automatically identify the data. It is the percentage of labels that have been correctly classified. he mathematical formulation for accuracy is given .The LSTM is employed with moment of adaptive learning which is designed for training the model where the result of confusion matrix is given by the accuracy score of 89.9%.By considering the accuracy of the classifier it achieves better precision for the intended dataset with cleaning the data and employing the ensemble classification methods. In this paper the Sentiment analysis is referred as opinion mining which is the process of extracting from text data and classifiers it into both positive and negative or neutral reviews. It is used to adapt automatically and categorize the IMDB movie reviews in total of 10k reviews are considered 5k are positive and the remaining 5k were negative sentiments. Finally the results were concluded that the highest accuracy provided is 89.9% .

Paper-5

Sentiment Analysis of Movie Reviews Using Support Vector Machine Classifier with Linear Kernel Function Sheik Abdullah · K. Akash · J. ShaminThres Thiagarajar College of Engineering, Madurai, Tamil Nadu 625015, India S. Selvakumar GKM College of Engineering and Technology, Chennai 600063, India © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021 V. Bhateja et al. (eds.), Evolution in Computational Intelligence, Advances in Intelligent

Systems and Computing 1176 Explanation: The process of sentiment analysis provides the mechanism of determining the state of attitude and subjective material for the given textual data. It also extracts the opinion or emotions from the textual data which can be further deployed in making decision. In recent days, sentiment analysis is used in social and health care applications for collecting and reviewing customer responses. Sentiment analysis basic task is to classify the given textual data to be as positive or negative. The advanced text analytic process is used to classify the emotions of a person as happy, sad, or angry. In simple terms, sentiment analysis predicts the psychological behaviour of a person. The authors deployed the process of sentiment analysis for data corresponding to social media with the analysis from the tweets collected from the users. The process is then segregated into positive/negative/neutral. The classification process has been carried out using Naïve Bayes, random forest algorithm, and logistic regression analysis. The authors proposed a new classifier approach called ensemble classifier which combines the base classifier into a single classifier, with the intention to improve the accuracy and the performance estimation of the sentiment analysis. By using sentiment lexica are used to check the polarity estimation by matching the words and sentiment polarities in the given text. Sentiment Classification model is used in the proposed methodology and semantic similarity metric is used to measure the performance. Proposed methodology includes two sub-modules document embedding and semantic similarity. For similarity-based feature extraction SIMON algorithm is used and embedding text representation selection over a lexicon is used. An overview of sentiment analysis, Challenges in the evaluation phase of sentiment analysis. Forty-seven papers related to sentiment analysis are used for this survey and classify papers based on domain oriented, Challenge type, Sentiment Analysis Challenge, and Review Type. For Sentiment Analysis Challenge types BOW technique, POS technique, semantic technique, lexicon technique, maximum entropy, and n-gram techniques are used. To improve the accuracy of sentiment analysis various models are used from linear to neural network models. Now a days deep learning methods are used in all fields of research including sentiment analysis. Moreover, some of the Arabic tweets seem to be complex with different dialects. To solve this, the proposed methodology is used for analysis. By using Clustering algorithms of machine learning the movie reviews considered for this research work involve the utilization of k-means clustering algorithm. The feature or aspect with the consideration of given text has to be analysed in accordance with the number of levels in the selected cluster. This algorithm determines the levels of grouping with regard to the similarity among the mean distance measure and the cluster centroids. The Support vector machine classifiers are widely used and most promising algorithms for handling data at all levels including textual data. The algorithm deploys the mechanism of non-linear mapping to transform the original observed training samples to a level of higher order dimensions. Thereby, with the available form of dimensional level it searches for the optimal linearly separable hyperplane which is also considered to be the decision level for segregating the tuples of records from one class to another. After we group the datasets under different clusters as per the number of clusters provided by the user by the usage of k-means clustering, the obtained data is then used for classification sentimentally. By doing the clustering process, we get to divide the data in a better way and utilize the data in a more efficient manner. The final sentiment classification is done using the SVM classifier. In order to perform the mechanism of sentiment analysis for the given set of text corpus, the table corresponding to sentiment and the selected text has to be indicated. The row has the textual part and its segregated sentiments. The column signifies the sentiments accordingly with the classified labels. The data has been taken from movie reviews which signify whether the movie comments represent to positive or negative. The designed model works accordingly with regard to the opinion that has been classified with the corresponding sentiments. Therefore, the opinions considered with the free form of text review are identified from the library executed files. The stop words, noun phrases are identified with the sentiment analysis process and the word cloud is formed as a result. The behaviour of the model developed is observed through the analysis from the applicability of different kernel functions as linear, polynomial, Gaussian, string.Linear Classifier is the first new best hyperplane algorithm. With the observed set of opinion pairs, the nearest observed opinion pairs are analysed and recorded. The ε value of the Support Vector is 50% of the sample ratio. Hence we can easily compute the optimal generalized performance to be smaller than 50%. This paper deals with the analysis of movie review data by using data clustering and data classification process. The classification algorithm used in SVM with a linear kernel process. The classification process produces 87.56% of accuracy level in segregating the positive and negative opinions for the movie review data.

Paper-6

Sentiment Analysis of Movie Reviews using Attention based-LSTM January 2021 Authors: Charu Gupta, Geetansh Chawla, Karan Rawlley, Kritarth Bisht and Mahak SharmaSome of the authors of this publication are also working on these related projects: Neutrosophic Set View projectBhagwan Parshuram Institute of Technology (aff. GGSIPU) © The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021A. Abraham et al. (eds.), Proceedings of 3rd International Conference on Computing Informatics and Networks, Lecture Notes in Networks and Systems 167,PUBLICATION: SPRINGER Explanation: In this paper we will be seeing about the Sentiment analysis refers to the NLP for identifying and extracting the subjective documents which intends to determine by detecting reviews using positive, negative and neutral ways. In the machine learning there are classifier algorithms to test and train the models. Deep learning is a function of AI that copies the operation of the human brain while handling data and forming patterns to make certain decisions. It is a subset of AI, having networks that can learn from unstructured data with no supervision. In this paper, an attention-based long short-term memory (LSTM) model, where it is employed to examine the sentiments in the movie reviews. The aim of this research is to implement deep learning approach in order to analyse the polarity of the movie reviews and compare the performance of various pre-defined machine learning models. LSTM associates both long-term and short-term memory to leverage the artificial neural networks and eliminates the gradient vanishing suffered in RNN. Attention layer further reduces the information loss by allowing the translator to observe the information provided by the original sentence, thus remembering the context . In the work compared numerous ways for the use of semantic data to enhance the sentiment analysis performance. The conventional approaches were not considering the semantic associations among sentences or document contents. In this method knowledge of the sentiment analysis with drug reviews containing numeric data. They processed sentences containing numerical terms for the classification on the basis of opinionated and nonopinionated sentences and further indication of polarity depicted by the use of fuzzy set theory. By integrating the Naïve Bayes and support vector machine having a decent performance in a variety dataset. In the developed method employed a number of semantic attributes and support vector machine classifier to perform sentiment analysis. The method remarked that the previous works done for the mentioned purpose needed to check whether the tweets were predictive or reactive. It was discovered that the tweets were more reactive than predictive. Wang and Zhang proposed bidirectional recurrent neural network which acquire sequences and relations from the unprocessed data. This provides the details of deep neural networks used in designing the Senti_ALSTM. The process of this system is firstly we review the movie dataset then do tokenization, stop word removal and vectorization, test and train the models and finally analyse the input data. In data pre-processing the dataset consists of different anomalies like stop words, punctuations, connectors, special characters, etc., which were not required for the further phases. In data pre-processing stage, the string is the input and this stage processes it by removing the anomalies mentioned, and the processed string is returned at last. First, the tags are modified by replacing the text between the opening and closing tags by replacing the characters by spaces followed by the removal of all the characters except capital and small letters resulting in long meaningless strings. The single characters are then replaced by space. At last, extra spaces are removed from the string. Further, the word encoding is the initial layer of the neural network model used; for this, tokenizer class is used to create a word to index dictionary. In the word to index dictionary, every word present in the collection is used as a key and a correlating distinctive index is used as the value for the key. GloVe is used for creation of embedding layer for word to vector conversion. For the classification of movie reviews into positive and negative, a number of models like vanilla neural network, CNN, LSTM, etc., were tested. SVM was found unable to integrate the feature sets, and therefore, no satisfactory results were achieved. In comparison with the rest of the models, Senti_ALSTM model out-performed the other models applied with an accuracy of nearly 88%. Also, vanilla neural network was found to have least accuracy of nearly 75% out of all the models. The order of performance according to the classification accuracy was found to be Senti_ALSTM > Bidirectional LSTM > LSTM > CNN > Vanilla Neural Network.

Paper – 7

Sentiment analysis is a computational process to identify and classify subjective information such as positive, negative and neutral from the source material. It is able to extract feeling and emotion from a piece of a sentence. This technology has been widely used to extract valuable information from people's views on social media. Hence, this project aims to classify movie reviews into positives, negatives and neutral polarity using lexicon-based method which used R as the language and development framework. Twitter data is used as the source material. Firstly, tweets were extracted using RStudio and Twitter API. Then data pre-processing was done by removing all the stop words and noises. Next was the tokenization process, which separates the words and matches the separated words with positive and negative words vocabulary. Finally, the result of the sentiment analysis is produced into positive, negative and neutral polarities. The results were evaluated using standard evaluation metrics that are the precision, recall, F1 score and accuracy. After all, it is found that the basic lexicon-based method is able to classify sentiment quite well with 52% accuracy. Apparently, the accuracy value achieved in our experiment is not impressive enough, but it is worth corresponding to the simplicity and minimal cost of development for sentiment analysis on Twitter data for movies. Sentiment analysis also known as opinion mining is a field of study that analyses people's sentiment, opinion, attitude and emotion towards entity such as individuals, movies, topics, products, organization and services [1]. This analysis of sentiment had spread across many fields such as marketing, consumer information, hotels, movies, books and socials. In 2004, the author in [2] said that the popularity of the internet had driven people to search for other people's opinions or reviews online before purchasing a product or seeing a movie. Then in 2018, author of [3] reported there were 90% of people, looks for internet reviews before they made a decision especially on buying something or booking hotels. For today, a great source for online reviews is social media. Social media websites like Twitter and Facebook are a major hub for users to publish their opinions online We evaluated the accuracy of the sentiment results to see the performance of the application. In this paper, we present the evaluation of the application. We apply the lexicon-based model and R script as the language of development. We chose movie reviews as our domain because commonly people seek for movie's review when they want to find more information about the movie. It might help them in determining whether they should watch, rent or buy the movie. Fig 1 shows a sample of movie reviResearchers who used Machine Learning (ML) method usually apply several types of classifier such as Naïve Bayes (NB), Support Vector Machine (SVM) and Neural Network (NN) classifier to solve the task. Among the methods, NB is the most popular for research that related to a straightforward classification. Some of them intentionally mixed two types of classifier to increase the accuracy of the result. As an example, [16] has mixed NB and NN which had gain better accuracy up to 80.65%. Then, [17], [18] and [19] had compared between NB and SVM, and all of them reported SVM gave better performance than NB in their experiments. Later author [20] has compared NB, SVM and KNN on 5 different domain of datasets, and they also reported that SVM has outperformed all other techniques. So, in most cases, SVM gave better accuracy even in different domain data.ew from tweets on Twitter. The development of this sentiment analysis project undertakes 4 basic steps. First, extracting the desired tweets using the Twitter application, then clean the tweets for further analysis. After that, a sentiment score is produced for each tweet. Lastly, segregate it into positive, negative and neutral tweets. Tweets that are used for this project is focused only on movie reviews. Hashtag of the tweets is used as the navigator that led to the movie reviews. Only tweets in English are considered for this project. This project used the lexicon-based method, which applied a combination of lexicon and learning-based approaches for concept-level sentiment analysis. It is a basic and straightforward method which count the aggregated value of sentiment from the existing vocabulary (lexicon) to analyse the sentiment of the document. This means the vocabulary of positive and negative words is used to classify the sentiments of tweets word. In this case, it did not apply any machine learning All the tweets went through the data pre-processing part which applies data cleaning process involving removing the unexpected and unwanted (noises) data generated by Twitter such as retweet numbers, stop words, punctuations, whitespaces, changing the case of the letter to lower case and tokenization. These steps led to producing a clean text to be analysed. Once the tweets are preprocessed, the data were ready to be classified and analysed. In the sentiment classification part, the tweet's word is matched with the positive and negative word (the lexicon) which has been preinstalled in the R working directory. This process produced a collection of polarity for negative and positive text (review). Tweet that is not positive or negative is counted as a neutral tweet. Finally, the results of polarity are stored in the RStudio. The sentiment results were presented on a user interface developed using Shiny. Shiny is a package in

Rstudio which offer a great web framework for developing a web application. classifier. The process starts by dividing tweet sentences into words (tokenization) so that it will be able to compare and match the positive and negative movie which has less than 50% accuracy. We consider that if the accuracy is more than 50% than it is a good prediction of sentiment analysis. Based on the results of the experiments in this project, it can be concluded that using lexicon-based model is quite good because out of five movies, only one movie achieve accuracy less than 50%. Based on the literature done, it seems that the accuracy for the lexicon-based approach appears to be lower than the accuracy earned by machine learning approach. According to Chen[9] it is common to have lower accuracy when using lexicon-based model compared to using a machine learning model. Although the lexicon-based model gains lower accuracy, it is much simpler and straightforward to implement. In comparison with other research work on lexicon-based approach, seems that many other researchers get higher accuracy value because they incorporate some additional features to the basic lexicon. The accuracy score for this project is not impressive may be due to its straightforward and plain lexicon used in the algorithm. Besides that, movie review data is also said to be difficult to analyse compared to other domain review We conclude that applying the lexicon-based model using R is succeeds in classifying movie review sentiment and gives the user a piece of valuable information about the polarity of the review for that particular movie by looking at the accuracy percentage that is more than 50%. The average accuracy rate for all five movies is 52%. This result is comparable with other existing experiments. Additionally, the result of this project also proves that lexicon-based model is able to classify sentiment with very minimal computational cost compared to machine learning which is very datadependent (most data need to be trained on a specific corpus). Applying the lexicon-based method using R, make the development of sentiment analysis task become much simpler and efficient. In addition, R script and its tool made the sentiment result presented in a more interesting way.

Paper – 8

Every day a large set of data are collected for various purposes from diferent sources. Analyzing these large data sets is very essential as we do not need to use all the infor mation depending on the application usage. Hence, mining this data set including text and sentiment is gradually becoming very important and useful for application purposes. Market analysis experts make plans for any production by taking into account the users' feedback and buying habits. Using diferent sentiment analysis methods, these tasks can be accomplished successfully. In our research, we discuss here the existing lexicon analysis method and find out the limitations of its methodology, e.g., lower accuracy. Although some researchers have proposed and compared the accuracy of their models with the exist ing lexicon approaches, our proposed and developed customized model shows good accu racy for movie data reviews compared to the existing approaches. Diferent social media including Twitter, Facebook, etc. are being very popular for sharing users' views or comments via text, video, etc. These media are found for various purposes, e.g., selling or buying a product, sharing the issues of politics, seeing movies, etc. Besides, users' feedback can be got from those posts which are very useful for specifc applications. Data from those media are useful for knowing the patterns and trends of users' particular inter ests [1]. Retrieving useful information from those texts using various mathematical and statis tical methods is known as text mining. In text-mining, opinion mining or sentiment analysis is the most attractive research feld. Sentiment analysis is such an approach where appropriate sentiments can be summarized and analyzed from a large text data. It represents the text's opinions as positive, neutral, and negative. It can be implied at various levels- sentence, docu ment, or feature and word level. Document-level represents the sentiments expressed on an object. Sentence level represents the sentiments about an entity excluding the features analysis of sentiments provokes several challenges. Although sentiment analysis is very com plex, it has a very vast application. Businesspersons, scientifc researchers, market analyzers, and other professionals use it. As a result, several techniques are used in sentiment analysis. As languages are evolving day by day, it is making the job harder than before. Researchers are inventing various technologies, methods, and algorithms for sentiment analysis. Those are not out of the box on the basis of accuracy, but their accuracies are progressively improving. We have studied several sentiment analysis methods. We consider their complexity, computational power consumption, and other properties along with their accuracies. We focus on low com putational power consumption methods. These types of methods usually provide lower accu racy due to some limitations. We concentrate on specifc limitations and try to increase the accuracy level at a satisfactory stage [3]. In the remaining paper, Sect. 2 contains the related works of the research. The methodology and the proposed models with the existing lexicon approach are mentioned in Sect. 3. Section 4 analyses the proposed and existing lexicon-based model. Section 5 includes conclustion We can divide sentiment analysis into two categories to fnd out the sentiments from text automatically. The frst category is the lexicon-based sentiment analysis which refers to fnd sentimental classifcation by calculating semantic oriented words or phrases [4]. The second category is known as machine learning-based sentiment analysis. In this approach, classifers are contracted with statistical and complex algorithms. Unlike the frst approach of sentiment analysis, the second approach requires more computational power to calculate sentimental classifcation. Here, we use the frst approach where a dictionary is used which contains a large collection of words with related value to express their polarity. The simplest way to apply lexicon-based sentiment analysis is to present the lexicons as a simple feature [2]. The inten sity-based approach collects the sentiment values by adding sentiment linked in a sentence [3]. Another technique is to take the sentiment related characteristics like maximal, total tokens, and complete scores. There are large collections of the sentimental lexicon. Researchers have used these lexicon resources in their studies. Various studies have been checked their accu racies and efciencies before using them in the main workstream. WordStat sentiment dic tionary is the freely available largest lexicon resource. It has approximately 14,000 words and provides a score for binary classifcation (positive or negative). Bill McDonalds 2014 Master Dictionary provides approximately 85,000 words. Harvard Inquirer has nearly 11,780 words and provides a complex way to score words. Each word is scored into 15+categories, e.g., negative, strong, positive, weak, pain, active, pleasure, passive, virtue and vice, etc. MPQA lexicon provides approximately 8200 words and classifes each word into a binary class. Bing Liu's Opinion Lexicon provides nearly 4782 negative and 2,005 positive words. SentiWordNet provides a score from 0 to 1 for a single word. It contains scored words nearly (either positive or negative) 29,000 Lexicon based technique works by determining the sentimental lexicons for fnding the sentiments for user text. It takes into account how many number of negative and positive words are found in the user text. The text is given a positive score if there exists greater positive words and vice versa for the negative score assignment. To fnd if a word negative or positive, a sentimental lexicon repository is built. Several techniques are used to estab lish a sentimental lexicon. Here, we use a dictionary-based approach s the existing lexicon-based approach. In Fig. 2, we present our proposed model where we design and implement a lexicon-based sentiment analysis approach. We collect 2000 IMDb movie reviews for our study from the IMDb website [5] by a python script. We collect random reviews that can be positive or negative and store those reviews into an excel fle. Here, we apply a dictionary-based technique for utilizing the external dictionary where the synonyms and antonyms of every word can be found apporach follows the lexicon-based dictionary where a collection of lexicons remain, e.g., SentiWordNet [6]. Moroever, it stores the record of semantic relationship between words. In our proposed model, we use micro-WordNet-Opinion 3.0 libraries [7, 8]. In KNIME, we have accessed the IMDb review data in excel format with the XLS Reader node. Then we transform our excel data into KNIMEs' document format using Strings To Document node. After that, we use two Dictionary Tagger nodes and add relevant tags to identify positive and negative words according to the SentiWordNet dictionary [6]. At this stage, we start data preprocessing using several preprocessor nodes provided by KNIME. Here, we use Number Filter, Stop Word Filter, Punctuation Erasure, and Case Conver sion nodes to preprocess our document. Numbers are not important for our analysis, so we flter out all numbers using the Number Filter node. This node is used to flter out all numbers from all reviews. Stop words and punctuations are always unnecessary for natural language processing. Hence we have fltered out all stop words and punctuations using the Stop Word Filter node and Punctuation Erasure node, respectively. We change all cases of our text to lower case with the Case Conversion node of KNIME. To apply lexicon analysis we create a Bag of Words (BoW) from our preprocessed data. Hence Bag of Words creator node is used to do that job. Term to Strings node is used to extract each sentimental term from the bag of words into KNIMEs' data format for further pre processing. At this moment of analysis, we do some essential calculations for lexicon.

Paper – 9

Entertainment is crucial part of human life entertainments like songs, music, drama and movies etc. So for watching good movies most of the peoples generally prefer theater. If movie is not good then we feel nervous and we think wasted our money and time for watching bad movie so people prefer to go to movie by reading reviews and rating for that movie on various apps like IMDb, flixter and voice etc. But by reading one or two reviews we cannot say movie is good or bad because different peoples have different opinions some peoples like action or some like thrill or romance so people gives reviews based on their area of interest. So we cannot predict movie for that we proposed movie reviews based on sentiment analysis and classification algorithms such as Naïve bays and Random forest (RF). Sentiment analysis generally utilized to identify the sentiment of huge amount of text. We compare naïve bayes and RF machine learning techniques for measuring negative, positive and neutral reviews. Index Terms: Sentiment analysis, Naïve Bayes algorithm, Random Forest algorithm. the Machine learning approach makes utilization of a datasets and a test informational collection to build up a classifier. It is preferably more straightforward over Knowledge base methodology. Since the improvement of calculations a few difficulties were looked in the field of Sentiment analysis. The first is that a sentiment word can be sure or negative contingent on the circumstance. The second test is that individuals don't in every case express conclusions similarly. Sentiment mining comprehends the connection between literary audits and the outcomes of those reviews. Sentimental analysis can be utilized to differentiate customers and followers depends on their attitude towards a specific brand or a movie or a product with the help of reviews. One can identify whether the product review is positive or negative or whether the user email is satisfied or not. Feature Extraction categorized into four types Syntactic Feature, Semantic Feature, Link based Feature, Stylistic Feature. The most commonly utilized features are the first two features. Syntactic feature utilizes word tags, patterns, phrases and punctuations. On the other hand, Semantic feature works on the relationship between words, signs and symbols. Linguistic semantics can be utilized to know the human expression through language accurately. The enhancement in the field of web technology has changed the manner by which individuals can express their perspectives. Individuals rely on this user perspective information for analyzing the items for online shopping or while booking film tickets for watching movies in theaters. The users are interfacing together through posts, Facebook, tweets on twitter etc. The measure of information is huge to the point that it is troublesome for a typical human to examine and come to conclusion. Sentiment analysis is extensively arranged in the two kinds initial one is an information based methodology and the other classification techniques. First methodology requires an expansive database of predefined feelings and a proficient information portrayal for recognizing sentiments. Then again Classification is also known as "Supervised learning". Linear Classifiers: Logistic Regression/Naive Bayes Classifier, Support Vector Machines, Decision Trees, Random Forest, Neural Networks are classification algorithms in Machine Learning The section I explains the Introduction of movie review using classification method such as NB and RF. Section II presents the literature review of existing systems and Section III present proposed system implementation details Section IV presents experimental analysis, results and discussion of proposed s SVM and Naïve bayes. Also author utilized lexicon approach to convert structured review into numerical score value. Here Liu, B [2] discussed sentimental analysis applications and its problems also presented types of sentimental analysis, also two relevant and important concepts of subjectivity and emotion were also introduced, which are highly related to but not equivalent to opinion. Based on textual review Mouthami, K., Devi, K.N. and Bhaskaran [5] did classification and sentimental analysis. First they used mining techniques to extract the features from huge data and sentimental analysis utilized to analyze the opinions or emotions of users from text data this analysis is widely adopted in CRM. Analyzing sentiment using Multi-theme document is very difficult and the accuracy in the classification is less. So proposed a new algorithm called Sentiment Fuzzy Classification algorithm with parts of speech tags is utilized to improve the classification accuracy on of Movies reviews dataset. Here Kanakaraj, M. and Guddeti [6] analyzing the mood, emotions of the society on a particular news from Twitter posts. The key point of this is to enhance the accuracy of classification by including Natural Language Processing Techniques (NLP) especially semantics and Word Sense Disambiguation. Author Chaovalit, P. and Zhou, L [7] examines movie review mining utilizing two methodologies that is machine leaning and semantic analysis. The methodologies are adjusted to movie review area for correlation. The outcomes demonstrate that our outcomes are equivalent to or surprisingly better than past discoveries. We likewise find that film audit mining is a more difficult application than numerous different sorts of review mining. The difficulties of film review mining lie in that authentic data is constantly blended with genuine audit information and amusing words are utilized recorded as a hard copy film review. Text analysis important type are Sentiment analysis or opinion mining that aims to support decision making by extracting and analyzing opinion oriented text, finding positive and negative opinions, and estimating how positively or negatively an entity regarded. Most of the users express their political or religious views on Twitter so tweets become valuable sources of individual express. Tweets data can be efficiently utilized to infer people's opinions for social studies. Author proposes a Tweets Sentiment Analysis Model (TSAM) [8] that can spot the social interest and general people's opinions in a social event. Gautam, G. and Yadav, D [9] analyze opinions or sentiments of the twitter data using machine learning approaches and semantic analysis. machine learning approaches such as Naive Bayes, Maximum entropy and SVM utilized to analyze the twits on twitter and lastly measured the performance of classifier in terms of recall, precision and accuracy. By using unigram feature extraction technique twitter dataset is analyzed and utilized. Then after that, different machine learning techniques [10] trains the dataset with feature and then the semantic analysis gives a large set of similarity and synonyms which gives the polarity of the content. WordNet enhances the accuracy. As a part of preprocessing they have cleared ambiguous information and not required blank spaces. After preprocessing, this preprocessed data is converted into numerical vector using TF-IDF and Count Vectorizer. SVM and NB classifiers are utilized to classify numerical vector [11]. Zhang et al. [15] utilized rule based semantic analysis to In proposed system NB and RF classification techniques and sentimental analysis are utilized that will provide interested movie reviews to web user. Generally sentimental analysis is termed as an opinion mining. It is utilized to identify user’s emotions, mood, interest and behavior by using text pattern data. Classification techniques such as NB and RF utilized for feature selection and extraction, NB will not work properly in terms of execution time and it requires more memory. RF takes less time and less memory to execute than NB. Then we compare NB and RF on the basis of time and memory, results show that Random Forest Algorithm is better than Naive Bayes Algorithm in terms of time and memory to recommend the good movie to usersAll the experimental cases are implemented in Java in congestion with Netbeans tools and MySql as backend, algorithms and strategies, and the competing classification approach along with various feature extraction technique, and run in environment with System having configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM , this project aims to classify movie reviews into positives, negatives and neutral polarity using lexicon-based method which used R as the language and development framework. Twitter data is used as the source material. Firstly, tweets were extracted using RStudio and Twitter API. Then data pre-processing was done by removing all the stop words and noises. Next was the tokenization process, which separates the words and matches the separated words with positive and negative words vocabulary. Finally, the result of the sentiment analysis is produced into positive, negative and neutral polarities. The results were evaluated using standard evaluation metrics that are the precision, recall, F1 score and accuracy. After all, it is found that the basic lexicon-based method is able to classify sentiment quite well with 52% accuracy. Apparently, the accuracy value achieved in our experiment is not impressive enough, but it is worth corresponding to the simplicity and minimal cost of development for sentiment analysis on Twitter data for movies.

5. DATASET

Our Dataset is taken from Kaggle and it is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. The train/test split has been preserved **for** the purposes of benchmarking, but the sentences have been shuffled from their original order. Each Sentence has been parsed intomany phrases by the Stanford parser. Each phrase has a Phrase Id. Each sentence has a Sentence Id. Phrases that are repeated (such as short/common words) are only included once in the data.

ATTRIBUTES:

We have 4 Attributes in our dataset

- Phrase ID
- Sentence ID
- Phrase (comments)
- Sentiment labels

The sentiment labels are:

- 0 - negative
  1 - somewhat negative
  2 - neutral
  3 - somewhat positive
  4 - positive

6. METHODOLOGY

In this project,we are using tf-idf transformation in order to develop feature vectors for each phrase. Tf means term-frequency while tf-idf means term-frequency into inverse document-frequency. The main purpose of using tf-idf instead of using raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occurvery frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus. Scikit python library provides several transformers for tf-idf transformation such as count vectorizer, tf-idf transformer and tfidf vectorizer. There are 3 approaches for transforming text data to tf-idf data.

- First use count vectorizer to build feature vectors with frequencyvalues.
- Then convert them to tf-idf values using tf-idf transformer.
- Tf-idf vectorizer combines the functionality of both count vectorizer and tf-idf transformer. It directly converts text data to feature vectors with tf-idf values.

All tokens are converted to lowercase to avoid duplicates when transferring text data to tf-idf format. In machine learning and data mining usually stop words are removed when processing text data to reduce the complexity of data. In this project, stop word removal has not performed as each phrase consists only few words and some phrases only consist of stop words.

7. ANALYTICAL APPROACH

- Data Extraction
- Text Pre-processing
- Removing Stop words and Punctuation
- Study of the variables by exploring the data
- Polarity
- Stemming and Lemmatization
- Division of data into train and test
- Count Vectorizer and Tf-Idf Vectorizer
- Model Development
- Evaluation

8. MACHINE LEARNING MODELS

We are using 3 machine learning models for prediction;

1. Logistic regression
2. Support vector machine
3. Naïve Bayes

LOGISTIC REGRESSION

It is a linear model for classification rather than regression. It is also knownin the literature as logit regression, log-linear classifier and maximum- entropy classification. In this model, the probabilities describing the possible outcomes of a single trial are model using a logistic function. A logistic function is used to find the accuracy and precision score for the prediction. Performance have been evaluated using holdout approach and validation approaches and categorization accuracy values

SUPPORT VECTOR MACHINE

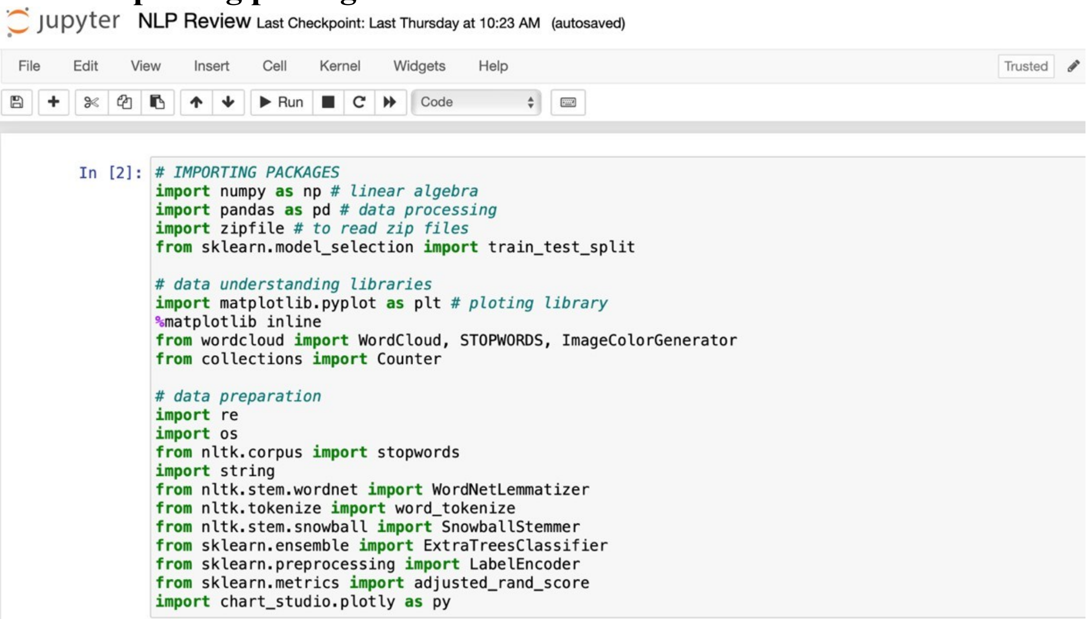
Support vector machines are set of supervised learning methods used for classification, regression and outlier detection. It tries to find a hyperplanethat can effectively divide the given training data into two parts. The major advantage of support vector machines is effectiveness in high dimensional spaces. Also it uses a subset of training points in the decision function called support vectors, so it is also memory efficient. SVM model isused to find the accuracy and precision score for the prediction. The one drawback in SVM is when training data is highly unbalanced, resulting model tends to perform well on majority data but perform bad on minority data. In scikit library, different types of kernels such as linear are provided. It implements the multi class classification using one against one approach.

NAÏVE BAYES

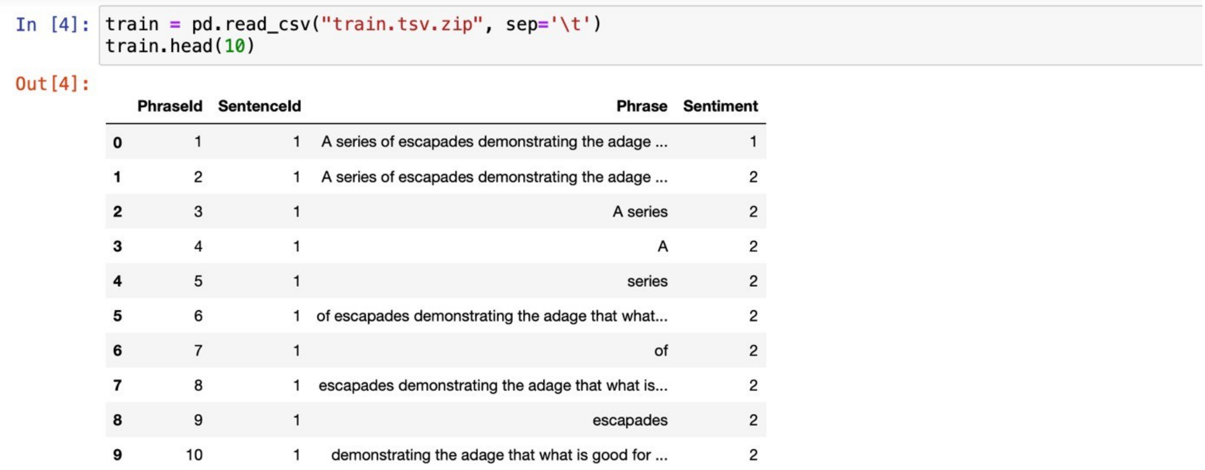
Naive Bayes methods are a set of supervised learning algorithms basedon applying Bayes theorem with the naive assumption of independencebetween every pair of features. Performance have been evaluated using holdout approach and validation approaches and categorization accuracyvalues.

9. IMPLEMENTATION

9.1 Importing packages and dataset


jupyter NLP Review Last Checkpoint: Last Thursday at 10:23 AM (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Run
Code

In [2]:
```
# IMPORTING PACKAGES
import numpy as np # linear algebra
import pandas as pd # data processing
import zipfile # to read zip files
from sklearn.model_selection import train_test_split

# data understanding libraries
import matplotlib.pyplot as plt # ploting library
%matplotlib inline
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from collections import Counter

# data preparation
import re
import os
from nltk.corpus import stopwords
import string
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import adjusted_rand_score
import chart_studio.plotly as py
```



In [4]:
```
train = pd.read_csv("train.tsv.zip", sep='\t')
train.head(10)
```

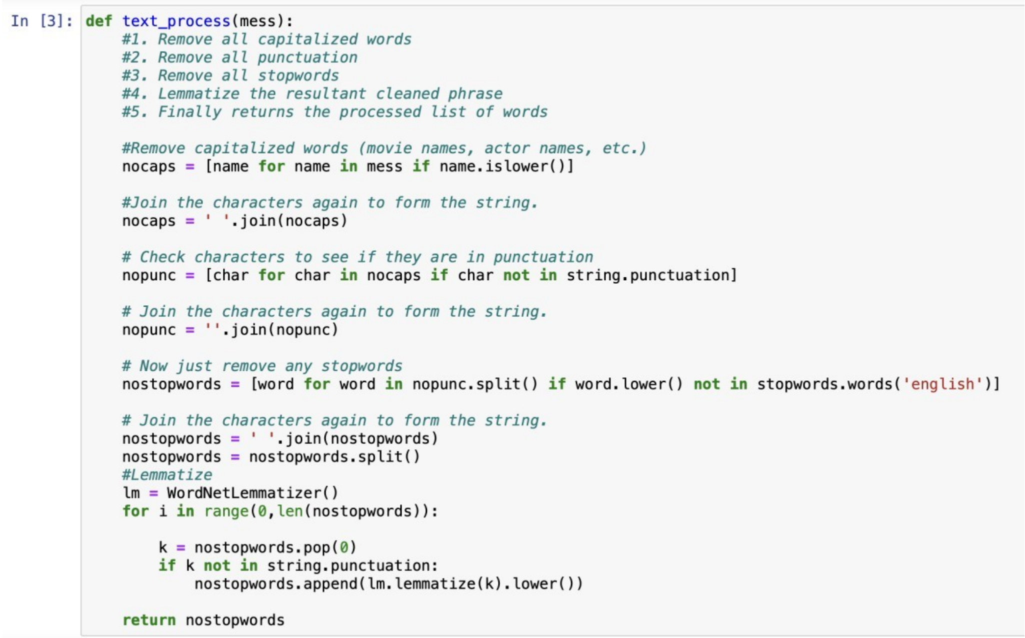
Out[4]:

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
1	2	1	A series of escapades demonstrating the adage ...	2
2	3	1	A series	2
3	4	1	A	2
4	5	1	series	2
5	6	1	of escapades demonstrating the adage that what...	2
6	7	1	of	2
7	8	1	escapades demonstrating the adage that what is...	2
8	9	1	escapades	2
9	10	1	demonstrating the adage that what is good for ...	2

9.2TEXT PRE-PROCESSING

9.2.1 Remove all capitalized words
9.2.2 Remove all punctuation
9.2.3 Remove all stop words
9.2.4 Lemmatize the resultant cleaned phrase
9.2.5 Finally returns the processed lists of words

Lemmatization
We are using Word-Net-Lemmatizer for this text pre-processing. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary formof a word, which is known as the lemma.

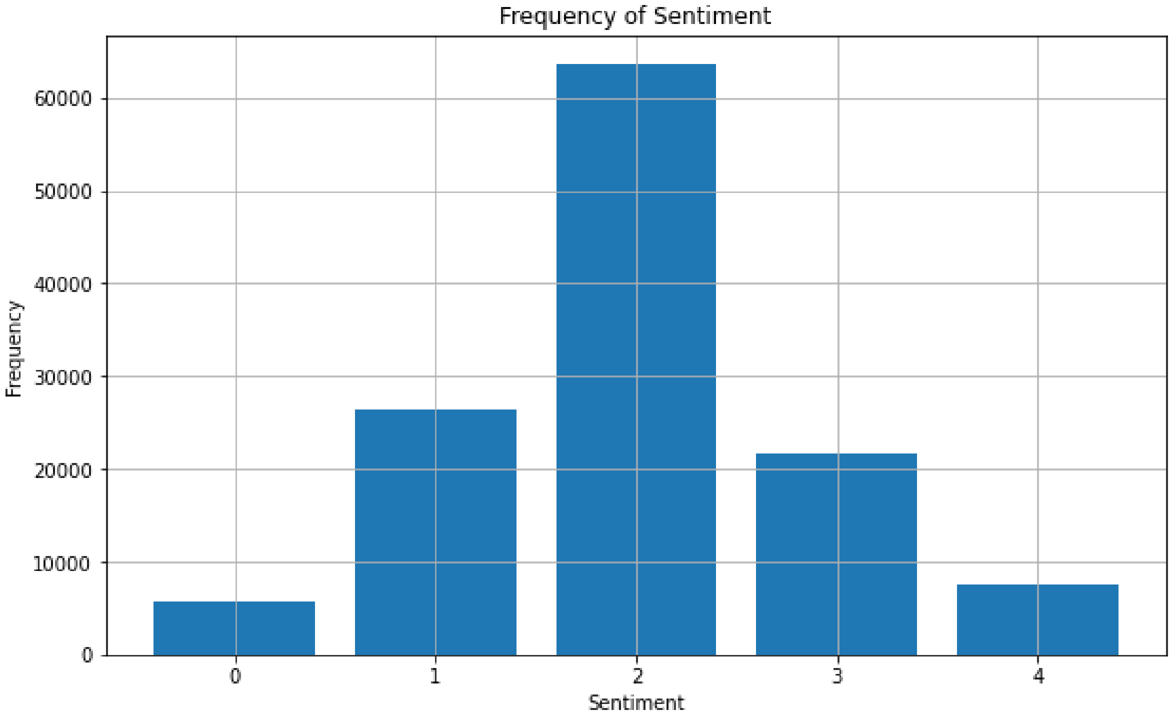


```
In [3]: def text_process(mess):
            #1. Remove all capitalized words
            #2. Remove all punctuation
            #3. Remove all stopwords
            #4. Lemmatize the resultant cleaned phrase
            #5. Finally returns the processed list of words

            #Remove capitalized words (movie names, actor names, etc.)
            nocaps = [name for name in mess if name.islower()]

            #Join the characters again to form the string.
            nocaps = ' '.join(nocaps)

            # Check characters to see if they are in punctuation
            nopunc = [char for char in nocaps if char not in string.punctuation]

            # Join the characters again to form the string.
            nopunc = ''.join(nopunc)

            # Now just remove any stopwords
            nostopwords = [word for word in nopunc.split() if word.lower() not in stopwords.words('english')]

            # Join the characters again to form the string.
            nostopwords = ' '.join(nostopwords)
            nostopwords = nostopwords.split()
            #Lemmatize
            lm = WordNetLemmatizer()
            for i in range(0,len(nostopwords)):

                k = nostopwords.pop(0)
                if k not in string.punctuation:
                    nostopwords.append(lm.lemmatize(k).lower())

            return nostopwords
```

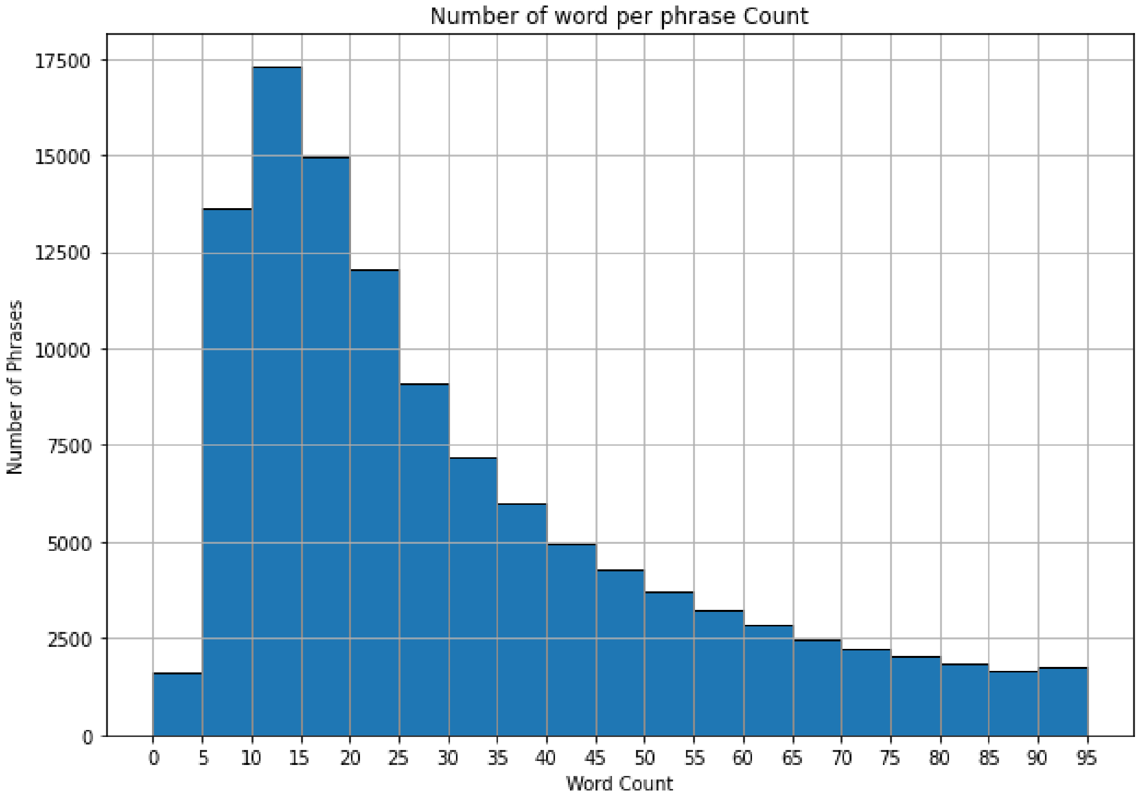
10. Results and Discussion:

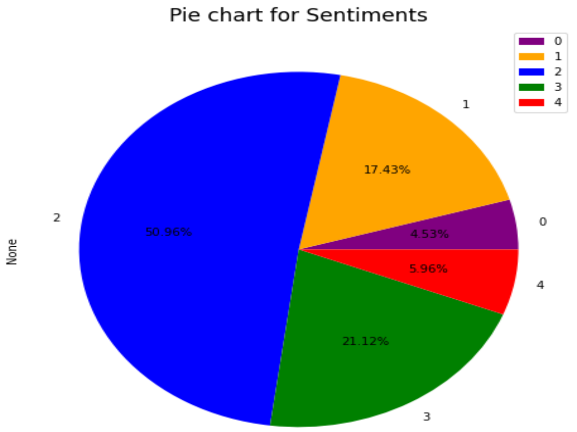
10.1 Splitting Train And Test Data:

The dataset provided in this competition is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. The dataset has been divided to training and test set for the purpose of benchmarking, but the sentences have been shuffled from their original order. Each Sentence in the dataset has been parsed into many phrases by the Stanford parser. Each phrase has a phrase Id. Each sentence has a sentence Id. Phrases that are repeated (such as short/common words) are only included once in the data.



10.2 DATA EXPOLARION VISUALIZATION:



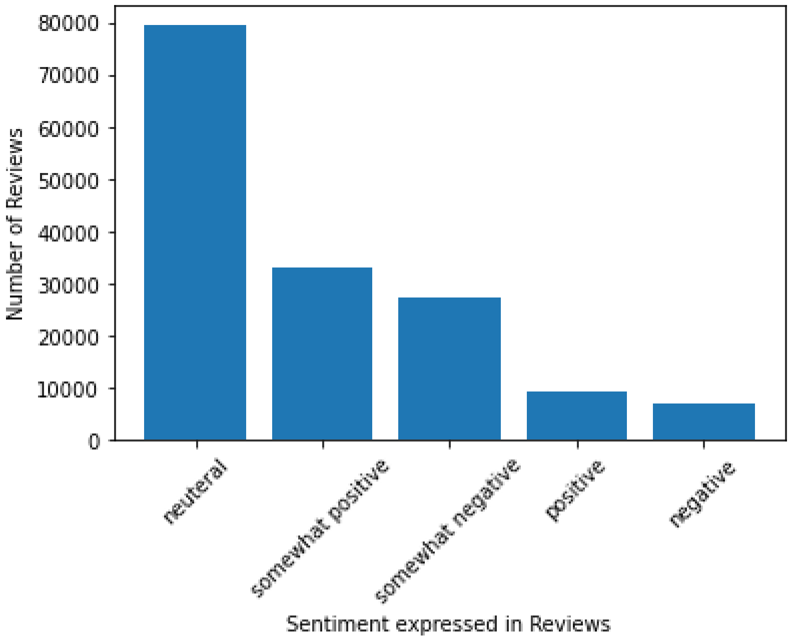


10.3 POLARITY:

There are 5 sentiment used to calculate the polarity.

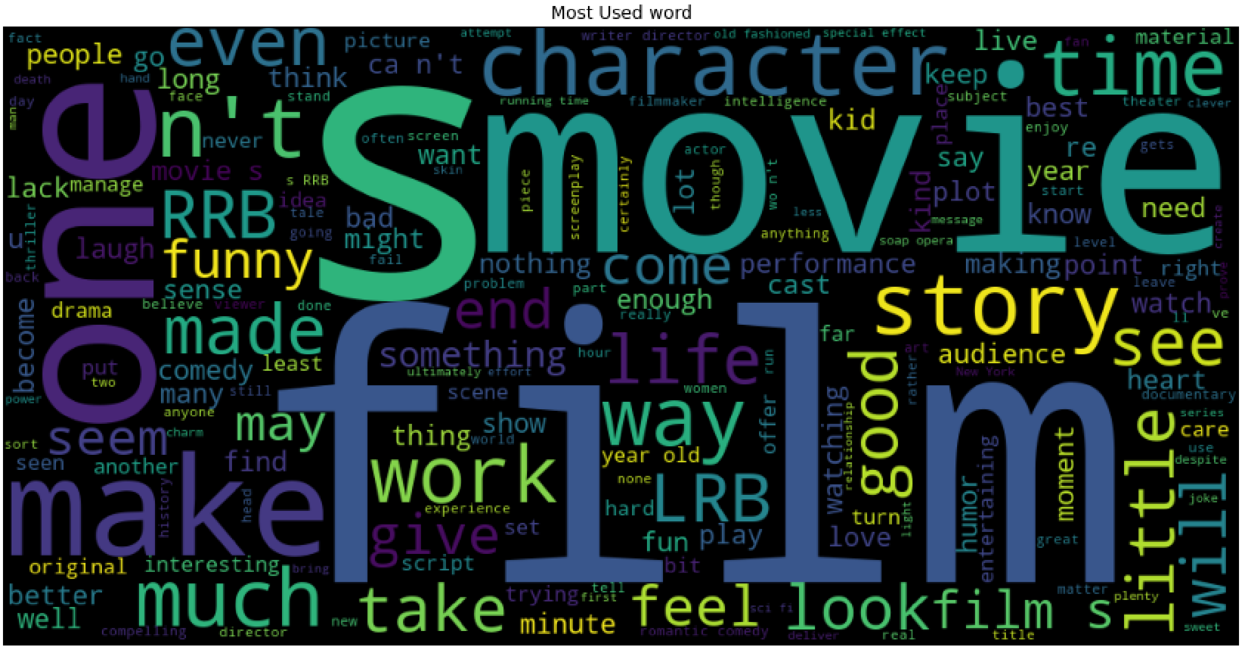
- Positive
- Somewhat positive
- Neutral
- Somewhat Negative
- Negative

```
neuteral              79582
somewhat positive     32927
somewhat negative     27273
positive               9206
negative               7072
Name: Polarity, dtype: int64
Text(0.5, 0, 'Sentiment expressed in Reviews')
```



10.4 WORD CLOUD:

Most words used in Phrase



Here, we can also see the other word counts,

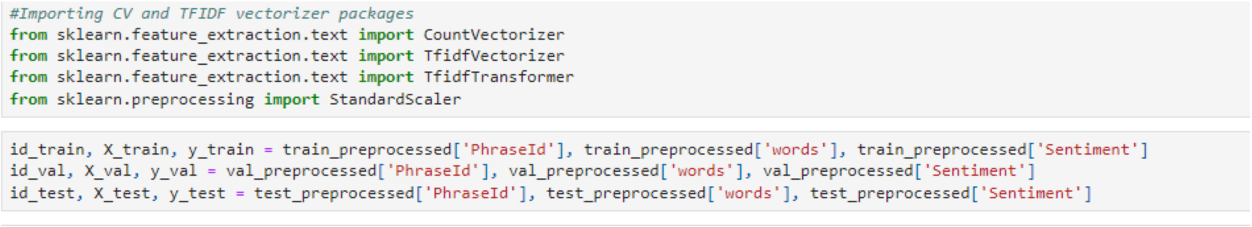
- Most Positive

- Most Negative

10.5 COUNT AND TF-IDF VECTORIZER:

Tf-idf Vectorizer is the base building block of many NLP pipelines. It is a simple technique to vectorize text documents. It transform sentences intoarrays of numbers and use them in subsequent tasks. Transforms text to feature vectors that can be used as input to estimator. vocabulary_ Is a dictionary that converts each token to feature index in the matrix, each unique token gets a feature index.The only difference is that the Tf-idf Vectorizer() returns floats while the Count Vectorizer() returns int. And that's to be expected – as explained in the documentation quoted above, Tfidf Vectorizer() assigns a score while Count Vectorizer() counts.

N-grams: N-grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighbouring sequences of items in a document. They come into play when we deal with text data in NLP(Natural Language Processing) tasks.

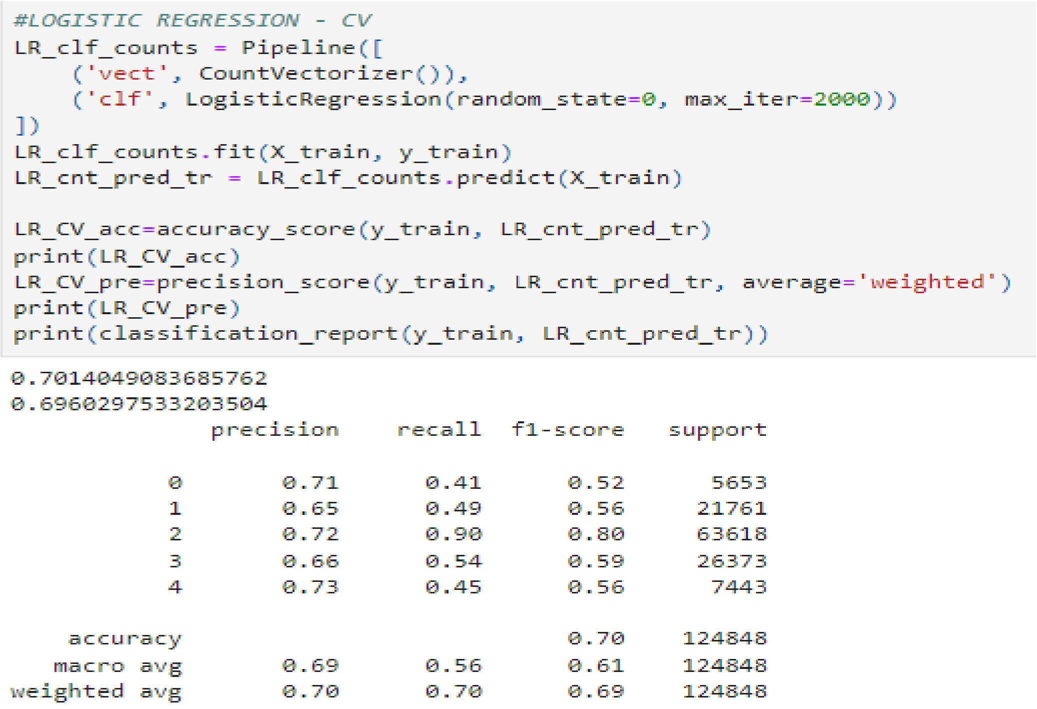


```
#Importing CV and TFIDF vectorizer packages
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.preprocessing import StandardScaler
```

```
id_train, X_train, y_train = train_preprocessed['PhraseId'], train_preprocessed['words'], train_preprocessed['Sentiment']
id_val, X_val, y_val = val_preprocessed['PhraseId'], val_preprocessed['words'], val_preprocessed['Sentiment']
id_test, X_test, y_test = test_preprocessed['PhraseId'], test_preprocessed['words'], test_preprocessed['Sentiment']
```

10.6 Result who Model Creation's has implemented:

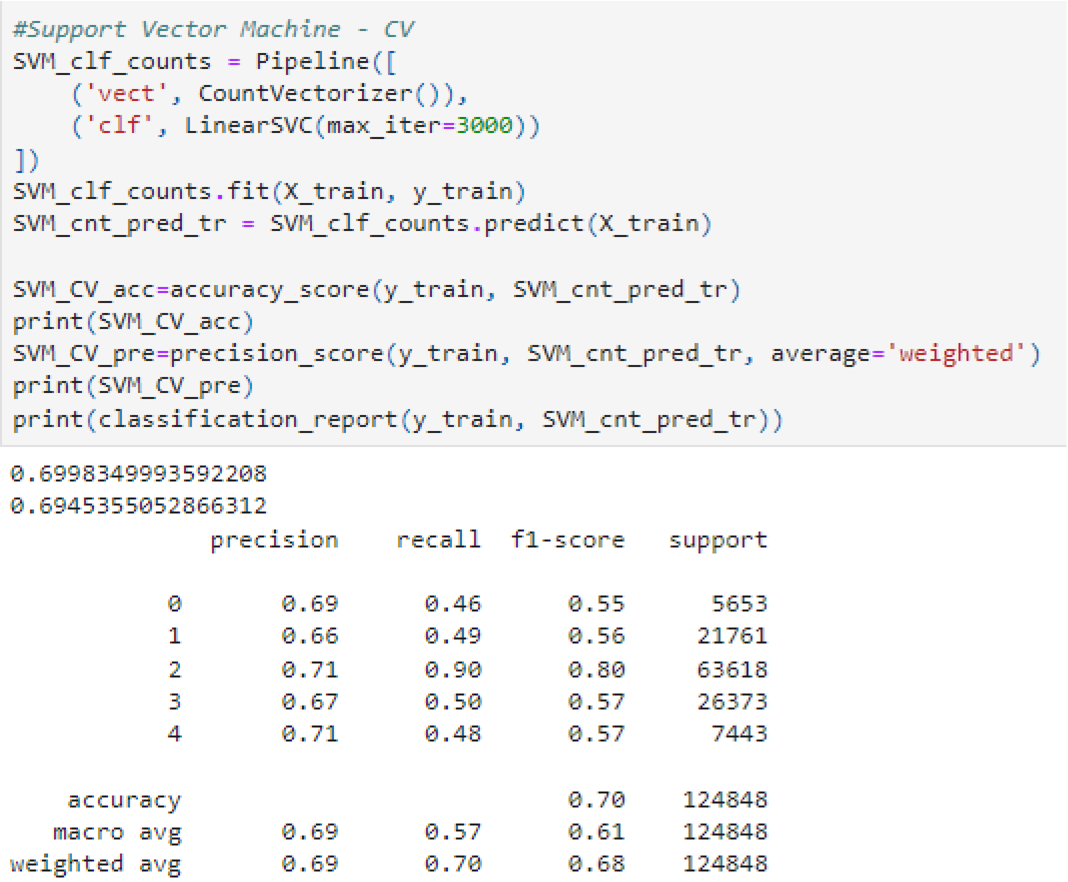
#LOGISTIC REGRESSION



```
#LOGISTIC REGRESSION - CV
LR_clf_counts = Pipeline([
    ('vect', CountVectorizer()),
    ('clf', LogisticRegression(random_state=0, max_iter=2000))
])
LR_clf_counts.fit(X_train, y_train)
LR_cnt_pred_tr = LR_clf_counts.predict(X_train)

LR_CV_acc=accuracy_score(y_train, LR_cnt_pred_tr)
print(LR_CV_acc)
LR_CV_pre=precision_score(y_train, LR_cnt_pred_tr, average='weighted')
print(LR_CV_pre)
print(classification_report(y_train, LR_cnt_pred_tr))
```

```
0.7014049083685762
0.6960297533203504
              precision    recall  f1-score   support

           0       0.71      0.41      0.52      5653
           1       0.65      0.49      0.56     21761
           2       0.72      0.90      0.80     63618
           3       0.66      0.54      0.59     26373
           4       0.73      0.45      0.56      7443

    accuracy                           0.70    124848
   macro avg       0.69      0.56      0.61    124848
weighted avg       0.70      0.70      0.69    124848
```

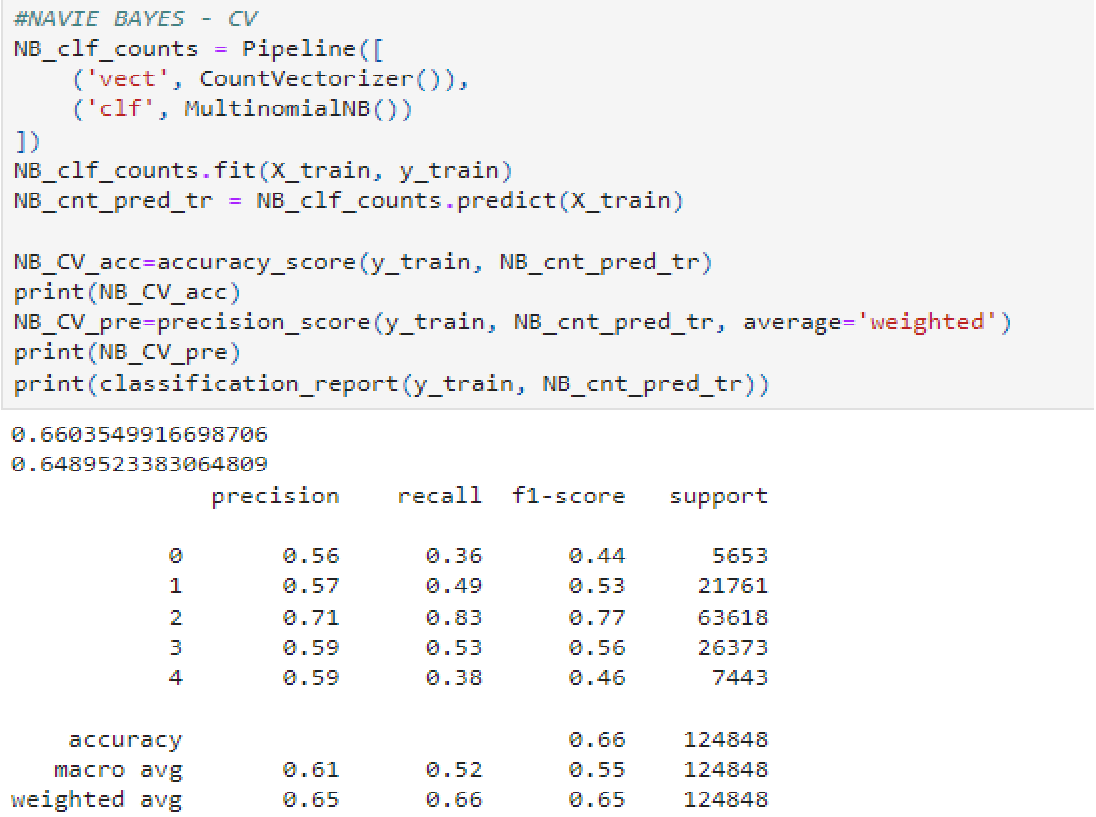
#SUPPORT VECTOR MACHINE



```
#Support Vector Machine - CV
SVM_clf_counts = Pipeline([
    ('vect', CountVectorizer()),
    ('clf', LinearSVC(max_iter=3000))
])
SVM_clf_counts.fit(X_train, y_train)
SVM_cnt_pred_tr = SVM_clf_counts.predict(X_train)

SVM_CV_acc=accuracy_score(y_train, SVM_cnt_pred_tr)
print(SVM_CV_acc)
SVM_CV_pre=precision_score(y_train, SVM_cnt_pred_tr, average='weighted')
print(SVM_CV_pre)
print(classification_report(y_train, SVM_cnt_pred_tr))
```

```
0.6998349993592208
0.6945355052866312
              precision    recall  f1-score   support

           0       0.69      0.46      0.55      5653
           1       0.66      0.49      0.56     21761
           2       0.71      0.90      0.80     63618
           3       0.67      0.50      0.57     26373
           4       0.71      0.48      0.57      7443

    accuracy                           0.70    124848
   macro avg       0.69      0.57      0.61    124848
weighted avg       0.69      0.70      0.68    124848
```

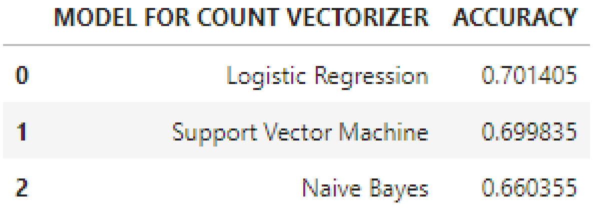
#NAVIE BAYES



```
#NAVIE BAYES - CV
NB_clf_counts = Pipeline([
    ('vect', CountVectorizer()),
    ('clf', MultinomialNB())
])
NB_clf_counts.fit(X_train, y_train)
NB_cnt_pred_tr = NB_clf_counts.predict(X_train)

NB_CV_acc=accuracy_score(y_train, NB_cnt_pred_tr)
print(NB_CV_acc)
NB_CV_pre=precision_score(y_train, NB_cnt_pred_tr, average='weighted')
print(NB_CV_pre)
print(classification_report(y_train, NB_cnt_pred_tr))
```

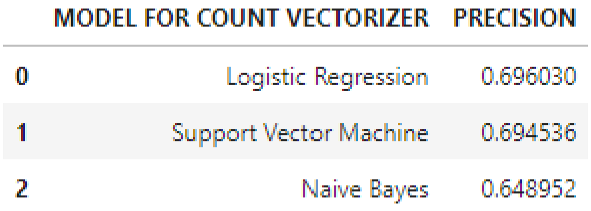
```
0.6603549916698706
0.6489523383064809
              precision    recall  f1-score   support

           0       0.56      0.36      0.44      5653
           1       0.57      0.49      0.53     21761
           2       0.71      0.83      0.77     63618
           3       0.59      0.53      0.56     26373
           4       0.59      0.38      0.46      7443

    accuracy                           0.66    124848
   macro avg       0.61      0.52      0.55    124848
weighted avg       0.65      0.66      0.65    124848
```

Result: ACCURACY:



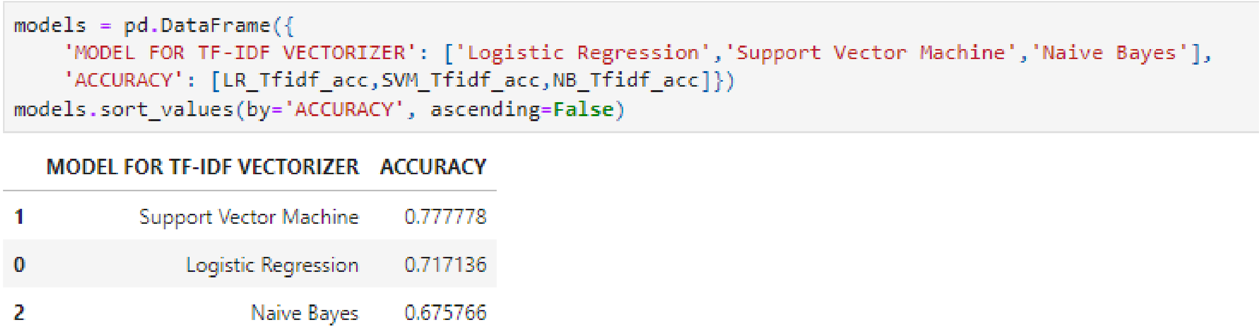
	MODEL FOR COUNT VECTORIZER	ACCURACY
0	Logistic Regression	0.701405
1	Support Vector Machine	0.699835
2	Naive Bayes	0.660355

Result: PRECISION:



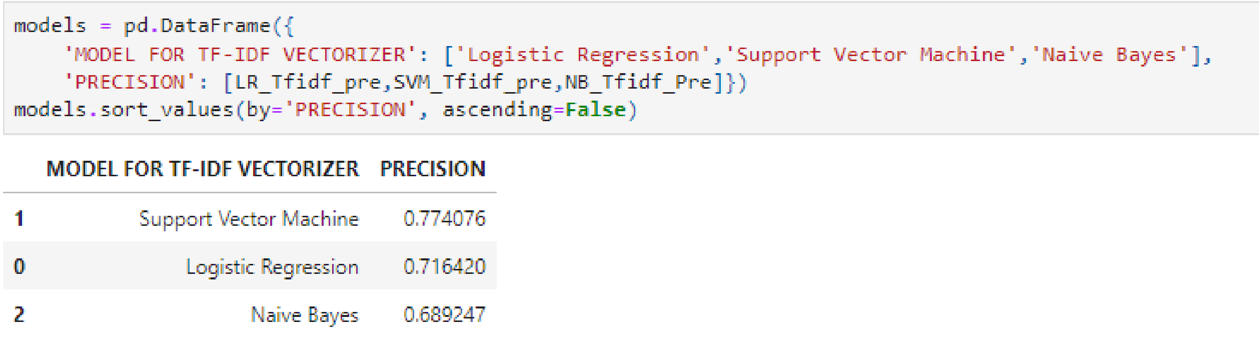
	MODEL FOR COUNT VECTORIZER	PRECISION
0	Logistic Regression	0.696030
1	Support Vector Machine	0.694536
2	Naive Bayes	0.648952

Result: TF-IDF VECTORIZER:



```
models = pd.DataFrame({
    'MODEL FOR TF-IDF VECTORIZER': ['Logistic Regression','Support Vector Machine','Naive Bayes'],
    'ACCURACY': [LR_Tfidf_acc,SVM_Tfidf_acc,NB_Tfidf_acc]})
models.sort_values(by='ACCURACY', ascending=False)
```

	MODEL FOR TF-IDF VECTORIZER	ACCURACY
1	Support Vector Machine	0.777778
0	Logistic Regression	0.717136
2	Naive Bayes	0.675766



```
models = pd.DataFrame({
    'MODEL FOR TF-IDF VECTORIZER': ['Logistic Regression','Support Vector Machine','Naive Bayes'],
    'PRECISION': [LR_Tfidf_pre,SVM_Tfidf_pre,NB_Tfidf_Pre]})
models.sort_values(by='PRECISION', ascending=False)
```

	MODEL FOR TF-IDF VECTORIZER	PRECISION
1	Support Vector Machine	0.774076
0	Logistic Regression	0.716420
2	Naive Bayes	0.689247

11. Summary:

According to above results it has become clear that the model implemented with the support of the NLP approach. Basically this method other than just carrying out normal classification process some important data pre- processing steps on the movie review data were carried out using the available tools within the NLP library, such as Stemming and Lemmatizing, in order to obtain further improvement of the performance. Moreover the Count and TF-IDF vectorizer module supported with the model building process of this approach which considerably increased the accuracy of the model which was built. Therefore with the discussion which has been carried out it has been clear that proper pre-processing of data based on natural language processing approaches as well as incorporating already existing models in the domain of sentiment analysis altogether with appropriate classification process can improve the performance of the model for multiclass classification of movie reviews.

12. Conclusion:

This project helps us to classify the dataset of movie reviews by utilizing some of the supervised Machine learning approaches like Logistic Regression, Support vector machine and Naïve bayes. For the Count vectorizer logistic regression has higher accuracy and precision score of 70% . For the Tf-idf vectorizer SVM has higher accuracy and precision score of 77% . and hence we used these techniques for better result of predicting best model to classify movie reviews from Rotten Tomatoes.

13. Future Work:

Future of Brand Monitoring, Sentiment analysis can help companies understand how customers feel about a brand: positive, negative, or neutral. Brand monitoring, including sentiment analysis, is one of the most important ways to keep customers engaged and interested. Branding can help a company improve its recognition, trust, and loyalty among customers as well as the effects of advertising.

14. References:

[1] Sisodia DS, Bhandari S, Reddy NK, Pujahari A. A comparative performance study of machine learning algorithms for sentiment analysis of movie viewers using open reviews. Performance Management of Integrated Systems and its Applications in Software Engineering. 2020:107-17.

[2] Patel, Priya, Devkishan Patel, and Chandani Naik. "Sentiment analysis on movie review using deep learning RNN method." Intelligent Data Engineering and Analytics: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 2 (2020): 155-163.

[3] Sentiment Analysis on Movie Review Using Deep Learning RNN Method Priya Patel, Devkishan Patel & Chandani Naik @© Springer Nature Singapore Pte Ltd. 2020.

[4] Atiqur Rahman, Md. Sharif Hossen(2019) , Sentiment Analysis on Movie Review Data Using Machine Learning Approach,. 019 International Conference on Bangla Speech and Language Processing (ICBSLP) 978-1-7281-5241-7/20/$31.00 ©2019 IEEE.

[5] Rahman, A., & Hossen, M. S. (2019, September). Sentiment analysis on movie review data using machine learning approach. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-4). IEEE.

[6] Qaisar, S. M. (2020, October). Sentiment analysis of IMDb movie reviews using long shortterm memory. In 2020 2nd International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE.

[7] Gupta, C., Chawla, G., Rawlley, K., Bisht, K., & Sharma, M. (2021). Senti_ALSTM: sentiment analysis of movie reviews using attention-based-LSTM. In Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020 (pp. 211-219). Springer Singapore.

[8] Gupta, C., Chawla, G., Rawlley, K., Bisht, K., & Sharma, M. (2021). Senti_ALSTM: sentiment analysis of movie reviews using attention-based-LSTM. In Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020 (pp. 211-219). Springer Singapore.

[9]Azizan, Azilawati, et al. "Lexicon-based sentiment analysis for movie review tweets." 2019 1st International conference on artificial intelligence and data sciences (AiDAS). IEEE, 2019.

[10]Rahman A, Hossen MS. Sentiment analysis on movie review data using machine learning approach. In2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 2019 Sep 27 (pp. 1-4). IEEE.

[11] Azizan, Azilawati, Nurul Najwa SK Abdul Jamal, Mohammad Nasir Abdullah, Masurah Mohamad, and Nurkhairizan Khairudin. "Lexicon-based sentiment analysis for movie review tweets." In 2019 1st International conference on artificial intelligence and data sciences (AiDAS), pp. 132-136. IEEE, 2019.

[12] Sisodia, Dilip Singh, et al. "A comparative performance study of machine learning algorithms for sentiment analysis of movie viewers using open reviews." Performance Management of Integrated Systems and its Applications in Software Engineering (2020): 107-117.

[13] Sisodia, D.S., Bhandari, S., Reddy, N.K. and Pujahari, A., 2020. A comparative performance study of machine learning algorithms for sentiment analysis of movie viewers using open reviews. Performance Management of Integrated Systems and its Applications in Software Engineering, pp.107-117.

[12] Adam, Noor Latiffah, Nor Hanani Rosli, and Shaharuddin Cik Soh. "Sentiment analysis on movie review using Naïve Bayes." In 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), pp. 1-6. IEEE, 2021.
