# **ACKNOWLEDGEMENT**

First of all, we would like to express our sincere thanks to the department of Computer Engineering for providing us the opportunity to work on the project named “**A Comparative Study of algorithms used in Sentiment Analysis”** to explore our abilities and knowledge in the field of Artificial Intelligence and Machine Learning.

Finally, we would like to thank our supervisor **Er. Milan Chikanbanjar** who inspired us to gather different information and conduct proper research on our study of sentiment analysis.

**Project Team Members:**

1. Arun Prajapati (740305)

2. Neetu Phaiju (740324)

3. Rabin Phaiju (740329)

4. Rodip Duwal (740334)

# **ABSTRACT**

Sentiment analysis is the computational study of people’s opinions, sentiments, attitudes and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years. It has a wide range of applications because opinions are key influencers of our behavior. In decision making, the opinions of others have a significant effect. The approaches of text sentiment analysis typically work at a particular level like phrase, sentence or document level. This system aims at analyzing a solution for the sentiment classification.

Keywords: Opinions, Sentiment Analysis, Sentiment Classification, Binary Classification, Inverse Document Frequency

# **LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Fig. No | Title | Page no. |
| 1.1 | Sentiment Classification | 1 |
| 1.2 | Sentiment analysis procedure | 2 |
| 4.1 | The proposed methodology for sentiment analysis | 8 |
| 4.2 | Porter Stemmer examples | 9 |
| 4.3 | Feature Extraction Procedure | 10 |
| 4.4 | Accuracy Comparison between SVM with TF-IDF and SVM with Doc2Vec | 10 |
| 4.5 | Datasets used for our research | 14 |
| 4.6 | Flowchart for Sentiment Analysis | 15 |
| 5.1 | Comparison of different performance measure of 6 different Algorithms for Movie dataset | 17 |
| 5.2 | Comparison of different performance measure of 4 different Algorithms for Conversation dataset | 17 |
| 5.3 | Comparison of different performance measure of Multinomial Naïve Bayes with respect to amount of data | 18 |
| 5.4 | Comparison of different performance measure of Decision Tree with respect to amount of data | 19 |
| 5.5 | Comparison of different performance measure of Logistic Regression with respect to amount of data | 19 |
| 5.6 | Comparison of different performance measure of KNN with respect to amount of data | 20 |
| 5.7 | Comparison of different performance measure of Random Forest with respect to amount of data | 20 |
| 5.8 | Comparison of different performance measure of SVC with respect to amount of data | 21 |
| 5.9 | Screenshot of Sentiment Analysis Application | 23 |
| 5.10 | Sentiment Tab | 23 |
| 5.11 | Sentence Tab | 24 |
| 5.12 | Result Tab | 24 |
| 5.13 | REST API | 24 |

# **LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| Table. No | Title | Page no. |
| 3.1 | Gantt chart for the completion of Program | 7 |
| 5.1 | Article Summary for Sentiment analysis using Naïve Bayes | 21 |
| 5.2 | Article Summary for Sentiment analysis using Decision Tree | 21 |
| 5.3 | Article Summary for Sentiment analysis using Logistic Regression | 22 |
| 5.4 | Article Summary for Sentiment analysis using KNN | 22 |
| 5.5 | Article Summary for Sentiment analysis using Random Forest | 22 |
| 5.6 | Article Summary for Sentiment analysis using SVC | 22 |

# **LIST OF ABBREVIATION**

1. IDF Inverse Document Frequency
2. KNN K-Nearest Neighbor
3. NLP Natural Language Processing
4. POS Part of Speech
5. SA Sentiment Analysis
6. SC Sentiment Classification
7. SVC Support Vector Classifier
8. SVM Support Vector Machine
9. TF Term Frequency

**TABLE OF CONTENTS**

[ACKNOWLEDGEMENT ii](#_Toc96938618)

[ABSTRACT iii](#_Toc96938619)

[LIST OF FIGURES iv](#_Toc96938620)

[LIST OF TABLES v](#_Toc96938621)

[LIST OF ABBREVIATION vi](#_Toc96938622)

[CHAPTER 1 1](#_Toc96938623)

[INTRODUCTION 1](#_Toc96938624)

[1.1 Background 1](#_Toc96938625)

[1.2 Motivation 3](#_Toc96938626)

[1.3 Statement of Problems 3](#_Toc96938627)

[1.4 Objectives 3](#_Toc96938628)

[1.5 Scope & Limitation 4](#_Toc96938629)

[CHAPTER 2 5](#_Toc96938630)

[LITERATURE REVIEW 5](#_Toc96938631)

[CHAPTER 3 7](#_Toc96938632)

[PROJECT MANAGEMENT 7](#_Toc96938633)

[3.1 Team Members 7](#_Toc96938634)

[3.2 Feasibility Study 7](#_Toc96938635)

[3.3 Work break down structure 7](#_Toc96938636)

[CHAPTER 4 8](#_Toc96938637)

[METHODOLOGY 8](#_Toc96938638)

[4.1 Background 8](#_Toc96938639)

[4.2 Datasets 14](#_Toc96938640)

[4.3 Comparison Measure 14](#_Toc96938641)

[4.4 Flowchart 15](#_Toc96938642)

[4.5 Tools and Platform 16](#_Toc96938643)

[CHAPTER 5 17](#_Toc96938644)

[RESULT AND DISCUSSION 17](#_Toc96938645)

[5.1 RESULT ANALYSIS 17](#_Toc96938646)

[5.2 DATASET EVALUATION 18](#_Toc96938647)

[5.3 RESEARCH ANALYSIS 21](#_Toc96938648)

[5.4 THE SENTIMENT ANALYZER 23](#_Toc96938649)

[5.5 REST API 24](#_Toc96938650)

[CHAPTER 6 25](#_Toc96938651)

[CONCLUSION AND FUTURE RECOMMENDATION 25](#_Toc96938652)

[6.1 Conclusion 25](#_Toc96938653)

[6.2 Future Recommendation 25](#_Toc96938654)

[REFERENCES 26](#_Toc96938655)

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 Background**

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs [1].

There are many researches that focus on sentiment classiﬁcation, with different goals and different supervised and unsupervised methods. These are usually based on machine learning models (e.g., support vector machines, maximum entropy, time series, Bayes, Hoeffding trees, artiﬁcial neural networks, etc.). Some of them use lexical resources and different features (e.g., unigrams, bigrams, etc.). The accuracy of such methods is incrementally better, being their main goal, the extraction of the sentiment based on the subjectivity and linguistic features of the words used in an unstructured text. Thus, it is quite common the division into two well differentiated types of methods for SA:

• Machine learning based methods - These methods are also divided in groups: supervised and unsupervised techniques, together with semi-supervised learning, a hybrid between them.

• Lexicon based methods - These techniques are based on dictionaries of words annotated with their semantic polarity.

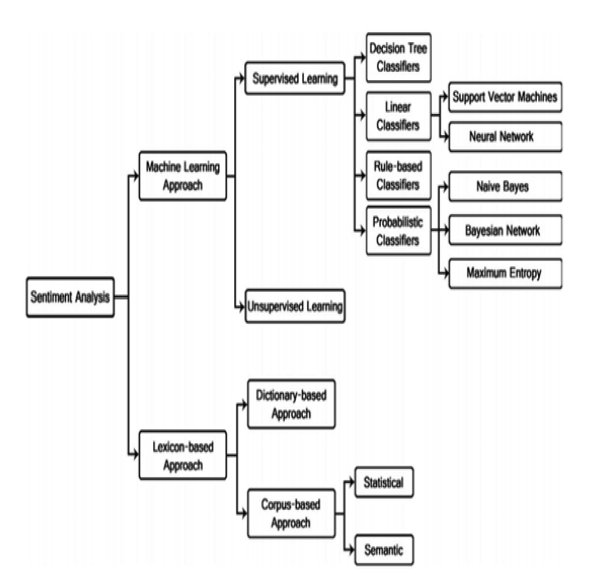


Fig 1.1: Sentiment Classification [Source: www.researchgate.net].

Sentiment Analysis can be considered a classification process as illustrated in Fig. 1.2. There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions. Wilson et al. have pointed out that sentiment expressions are not necessarily subjective in nature. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents [2]. Classifying text at the document level or at the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity like this sentence “The voice quality of this phone is not good, but the battery life is long”. This survey tackles the first two kinds of SA.

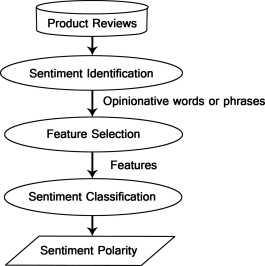


Fig 1.2: Sentiment analysis procedure

In this paper we have presented a comparative study of most commonly used algorithms for sentimental analysis. The task of classification is a very vital task in any system. We present a study of algorithms viz.

1. Naïve Bayes
2. Decision tree
3. Logistic regression
4. KNN
5. Random Forest
6. SVC

We showcase the basic theory behind the algorithms, when they are generally used and their pros and cons including their accuracy, F-score, Recall, Precision etc.

The data sets used in SA are an important issue in this field. The main sources of data are from the product reviews. These reviews are important to the business holders as they can take business decisions according to the analysis results of users’ opinions about their products. The reviews sources are mainly review sites. SA is not only applied on product reviews but can also be applied on stock markets, news articles, or political debates. In political debates for example, we could figure out people’s opinions on a certain election candidates or political parties. The election results can also be predicted from political posts. The social network sites and micro-blogging sites are considered a very good source of information because people share and discuss their opinions about a certain topic freely. They are also used as data sources in the SA process that performs sentiment analysis. In this paper proposes a model to analyses sentiment of different dataset of different source such as political statements, reviews in comments, posts from twitter, Facebook and/or other social media. One of the main tasks to do is the preparation and the process of the corpus as well as the selection of the features that finally will support the process of classification of the comments in positive, neutral, and negative.

### **1.2 Motivation**

Due to the exponential growth of the social network, the sentiment analysis has been applied to analyze the user’s opinions. The majority of the data that is constantly generated in the social networks could contain valuable information like perceptions and tendencies from the users to the objects, personalities, or services.

Due to its popularity nowadays, SA has been expanded to other fields like education, medicine, politics, and others. But reading all reviews is both time and money consuming therefore instead of spending times in reading and figuring out the positivity and negativity of text we can use automated techniques for sentiment analysis.

### **1.3 Statement of Problems**

A major benefit of social media is that we can see the good and bad things people say about the particular brand or personality.

The bigger your company gets difficult it becomes to keep a handle on how everyone feels about your brand. For large companies or a public figure with thousands of daily mentions on social media, news sites and blogs, it’s extremely difficult to do this manually.

To combat this problem, sentimental analysis is necessary. It can be used to evaluate the people's sentiment about particular brand or personality.

### **1.4 Objectives**

* To evaluate persons opinion in certain cases and the performance parameter of different algorithms used for sentiment analysis.

### **1.5 Scope & Limitation**

Sentiment analysis is the area which deals with judgments, responses as well as feelings, which is generated from texts, being extensively used in fields like data mining, web mining, and social media analytics because sentiments are the most essential characteristics to judge the human behavior [2].

In context of existing techniques, there are inadequate accuracy, incapability to deal with complex sentences and inability to perform well in different domains.

# **CHAPTER 2**

## **LITERATURE REVIEW**

In recent years a lot of work has been done in the field of “Sentiment Analysis” by number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative.

Sentiment analysis is the process of computationally identifying and categorizing opinions from piece of text, and determine whether the writer’s attitude towards a particular topic or the product, is positive, negative, or neutral. Finding the sentiments of a person (also called the polarization of a text) is a classical NLP problem. This problem brings a relatively new perspective by considering twitter tweets and Facebook posts, which are effectively a dialect of the English language. This problem has been considered already by a few authors. Since it is a necessary step but is not central to our questions, we plan to reuse some filters already available. If these detectors do not work as well as expected, we plan to use a robust binary classifier.

In 2019, Saad and Yang have aimed for giving a complete tweet sentiment analysis on the basis of ordinal regression with machine learning algorithms. The suggested model included pre-processing tweets as first step and with the feature extraction model, an effective feature was generated [3].

Pablo et. al. presented variations of Naïve Bayes classifiers for detecting polarity of English tweets. Two different variants of Naïve Bayes classifiers were built namely Baseline(trained to classify tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Neutral tweets neglected). The features considered by classifiers were Lemmas (nouns, verbs, adjectives and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters [4].

Po-Wei Liang et. al. (2013) used Twitter API to collect twitter data. Their training data falls in three different categories (camera, movie, mobile). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were filtered. Unigram Naïve Bayes model was implemented and the Naïve Bayes simplifying independence assumption was employed. They also eliminated useless features by using the Mutual Information and Chi square feature extraction method. Finally the orientation of an tweet is predicted i.e. positive or negative [5].

Pak and Paroubek (2010) proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons [6].

Agarwal et al. (2011) developed a 3-way model for classifying sentiment into positive, negative and neutral classes. They experimented with models such as: unigram model, a feature based model and a tree kernel based model. For tree kernel-based model uses 100 features and the unigram model uses over 10,000 features. They arrived on a conclusion that features which combine prior polarity of words with their parts-of-speech(pos) tags are most important and plays a major role in the classification task. The tree kernel-based model outperformed the other two models [7].

In 2011, Han-Xiao Shi, Xiao-Jun Li created a sentiment analysis model for hotel reviews based on supervised learning approach using unigram feature with two types of information (frequency and TF-IDF) to realize polarity classification of documents [8].

Barbosa et al. (2010) designed a two-phase automatic sentiment analysis method for classifying tweets. They classified tweets as objective or subjective and then in second phase, the subjective tweets were classified as positive or negative. The feature space used included retweets, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS [9].

Xia et al. used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naïve Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy [10].

We will compare most popular algorithms such as Naïve Bayes, Decision Tree, Logistic Regression, KNN, Random Forest and SVC on sentiment analysis. We will use 2 different datasets to find the polarity as positive or negative.

# **CHAPTER 3**

## **PROJECT MANAGEMENT**

In order to design our system on **Sentiment Analysis,** first we will collect the related information. Then we will plan the work for the success of the concerning concept.

### **3.1 Team Members**

For this project, we have a group of four members:

1. Arun Prajapati (740305)
2. Neetu Phaiju (740324)
3. Rabin Phaiju (740329)
4. Rodip Duwal (740334)

### **3.2 Feasibility Study**

The aim of feasibility study is to understand thoroughly all aspects of a concept, or plan. During the study, problems in the system are determined. So, it’s always good to have a contingency plan that test to make sure it’s a viable alternative in case the first plan fails. This study will be reliable for forecasting the detail of the system.

### **3.3 Work break down structure**

All the team members will work on different modules. During the course of work, team members will be in touch so that no problems will arise in future. After the completion of individual work, whole work will be combined to develop a proposed system.

Table 3.1 Gantt Chart for Completion of Program

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S.  N | Week Job  Description | 1st  Week | 2nd  Week | 3rd  Week | 4th  Week | 5th  Week | 6th  Week | 7th  Week | 8th  Week |
| 1. | Problem  Identification |  |  |  |  |  |  |  |  |
| 2. | Analysis |  |  |  |  |  |  |  |  |
| 3. | Design |  |  |  | |  |  |  |  |
| 4. | Coding |  |  |  |  | | |  |  |
| 5. | Implementation  and testing |  |  |  |  |  |  | | |
| 6 | Documentation |  |  |  |  |  |  | | |

# **CHAPTER 4**

## **METHODOLOGY**

### **4.1 Background**

The step involved for feature extraction are described as follows:

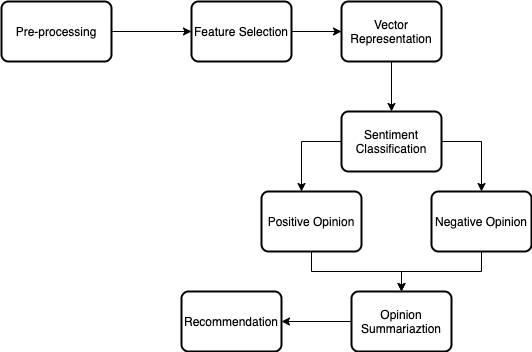


Fig 4.1: The proposed methodology for sentiment analysis.

**4.1.1.** **Preprocessing**: In Preprocessing the unanalyzed data is handled for feature extraction. In this phase, the several techniques like POS tagging, Stemming and Stop word removal are applied to data set for noise reduction and facilitating feature extraction [24].

It is further divided into below steps:

* **Tokenization**: White spaces, symbols and special characters are removed and a sentence is divided into words. Tokenization helps to filter out unnecessary tokens. For example, a document into paragraphs or sentences into words. For our paper we tokenized each sentence for making stemming and stop word removal easier.
* **Cleaning**: For this we created a cleaner class that:
  + Converts uppercase letters to lowercase
  + Normalization: Convert number to its equivalent word or completely removes numbers.
  + Removes punctuations, accent marks and whitespaces etc.
  + Removes contractions.
* **Stemming and Lemmatization:** Stemming and Lemmatization are two essential morphological processes of preprocessing module during feature extraction. It has been seen that most of the times the morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. Since the meaning is same but the word form is different it is necessary to identify each word form with its base form. To do this a variety of stemming algorithms have been developed [8].

In our paper we used Porter stemmer, is as of now one of the most popular stemming methods proposed in 1980. This reduces all words with the same root to a single form, the stem, by stripping the root of its derivational and inflectional affixes [27].

There are mainly total of 5 rules: The rule looks like the following:

<condition> <suffix> → <new suffix>

For example, a rule (m>0) EED → EE means “if the word has at least one vowel and consonant plus EED ending, change the ending to EE”. So “agreed” becomes “agree” while “feed” remains unchanged [4].

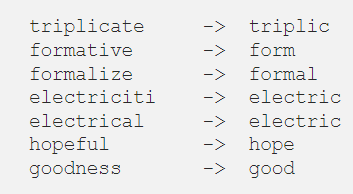


Fig 4.2: Porter Stemmer examples

* **Stop Word Removal:** Stop word concept was first introduced by Hans Luhn, H.P. Stop words are common and high frequency words like “a”, “the”, “of”, “and”, “an”.]; ultimately enhance performance of feature extraction algorithm.

The stop words removal reduces dimensionality of the data sets and thus key words left in the review corpus can be identified more easily by the automatic feature extraction techniques. Words to be removed are taken from a commonly available list of stop words. Savoy had given huge collection stop word list. At simplest level stop words are iterated in chosen word list and removed from text. We made use of list that contains stop words and filter out the words for sentence that matches those stop words.

**4.1.2.**  **Feature extraction:** Feature extraction handles the following task:

* Feature Type: In this step features are identified like the term frequencies, term cooccurrences, Opinion word, OS information, Negation Syntactic Dependencies.
* Selection of Feature: Good features are selected for classification using the following ways like Information gain, Document frequency, Odd ratio and Mutual Information.
* Feature Weighting Mechanism: The features are ranked by computing the weight using term presence, term frequency and Inverse document frequencies.
* Reduction of Feature: To optimize the classifiers performance the vector size is reduced.

Two techniques are used for feature extraction -TF-IDF and Doc2vec. Figure 4.3 shows the steps involved in extracting features using both the techniques.

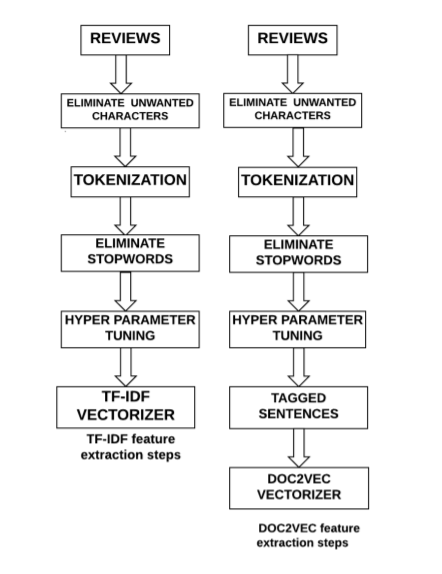


Fig. 4.3: Feature Extraction Procedure [Source: sciendo.com].

But we only used TF-IDF for our project because the performance of Doc2vec is poor for short documents and according to study of performance comparison between these two models TF-IDF with SVM performed better than Doc2vec with SVM [12].

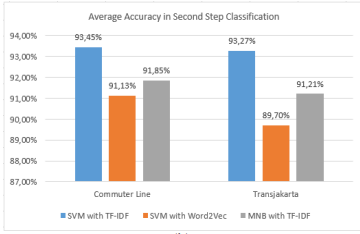


Fig. 4.4: Accuracy Comparison Between SVM with TF-IDF and SVM with Doc2Vec [Source: sciendo.com]

**TF-IDF**

TF-IDF is short form for term frequency-inverse document frequency.TF-IDF is one of the largely used methods in information retrieval and text mining. TF-IDF is a weight metric which determines the importance of word for that document [11].

**TF**

Term Frequency measures number of times a particular term occurred in a document d. Frequency increases when the term has occurred multiple times. TF is calculated by taking ratio of frequency of term t in document d to number of terms in that particular document d.

**IDF**

TF measures only the frequency of a term t. Some terms like stop words occur multiple times but may not be useful. Hence Inverse Document Frequency (IDF) is used to measure term’s importance. IDF gives more importance to the rarely occurring terms in the document d. IDF is calculated as:

The final weight for a term t in a document d is calculated as:

**4.1.3.** **Sentiment Classification:** Sentiment classification uses two approaches to classify the nature of documents/sentence. They are Machine Learning Approaches and Lexicon Based Approaches.

* The Machine Learning belongs to supervised learning and classification of text in particular. So, it is called as ―Supervised Learning.
* Lexicon Based approach consist of Dictionary based and corpus based.

Classifiers are trained on training data sets with the features obtained from the feature extraction techniques TF-IDF and Doc2vec and with the corresponding output labels from the dataset. The test data is tested with the trained classifiers to predict the sentiments and the accuracy is measured for test data. For our purpose we used 6 popular classifiers, they are: Naïve Bayes, Decision tree, Logistic regression, KNN, Random Forest, SVC.

**4.1.3.1. Naïve Bayes**

Bayesian classifiers are based around the Bayes rule, a way of looking at conditional probabilities that allows you to flip the condition around in a convenient way. A conditional probably is a probably that event X will occur, given the evidence Y. That is normally written P(X | Y). The Bayes rule allows us to determine this probability when all we have is the probability of the opposite result and of the two components individually:

This restatement can be very helpful when we are trying to estimate the probability of something based on examples of it occurring.

In our case, we are trying to estimate the probability that a document is positive or negative, given its contents. We can estimate the probability of a word occurring, given a positive or negative sentiment by looking through a series of examples of positive and negative sentiments and counting how often it occurs in each class. This is what makes this supervised learning, the requirement for pre-classified examples to train on.

**4.1.3.2. Decision Tree**

This classifier worked on the basis of dividing the working area repetitively into multiple sub parts. Decision Tree can possibly produce important insight into interactions between variables. Because, there are two different regions of same class. Class is either divided on the basis of purity in classification or impurity in classification. Clear separation of work classes is done easily. But when there is impurity in classes. It is to be solved by the concept of Entropy. Entropy works on the basis degree of how particular element is occurred randomly [13]. Mathematically it is calculated using probability of element as:

P(x) is a probability of element x.

**4.1.3.3. Logistic Regression**

Logistic Regression belongs to the family of classifiers known as the exponential or log-linear classifiers. It works by extracting some set of weighted features from the input, taking logs, and combining them linearly (meaning that each feature is multiplied by a weight and then added up). Technically, logistic regression refers to a classifier that classifies an observation into one of two classes, and multinomial logistic regression is used when classifying into more than two classes [22]. Logistic Regression is a discriminative classifier. Recall that the job of a probabilistic classifier is to choose which output label y to assign an input x, choosing the y that maximizes P(y|x).

**4.1.3.4. KNN**

KNN uses feature similarity where it assigns a data point based on how close it is to its neighbor. The algorithm for the KNN shown below is used for the classification of the data.

Algorithm:

Step 1: Load dataset

Step 2: Select the value of k

Step 3: Calculate the distance between each data point using Euclidean distance

Step 4: Sort data point according to the distance calculated

Step 5: Select the top k row

Step 6: Assign data point on the most frequent class

Step 7: End

To calculate the distance of the data point in the KNN algorithm Euclidean distance is calculated.

**4.1.3.5. Random Forest**

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Random forest takes the prediction from each tree and based on the majority votes of predictions and it predicts the final output. The greater number of trees in the forest leads to higher accuracy.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

Algorithm:

Step 1: Select random k data points from the training set.

Step 2: Build the decision trees associated with the selected data points.

Step 3: Choose the number N for decision trees that we want to build.

Step 4: Repeat Step 1 & 2.

Step 5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

**4.1.3.6. SVC**

The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. Support Vector Classification (SVC) can be extended to solve regression problem. That extended method is called Support Vector Regression (SVR).

Linear SVC is based on liblinear and only supports a linear kernel. So SVC (kernel = ‘linear’) is in theory “equivalent” to: LinearSVC( ). The objective of a Linear SVC is to fit to the data provided returning a best fit hyperplane that divides, or categorizes, the data.

**4.1.4.** **Visualization**: An application software that takes text or voice input and shows the sentiment behind each sentence, Also the software will compare the different plots to show the frequency of words in the customer tweets and the sentiment scores.

### **4.2 Datasets**

We employ two datasets in this paper.

**4.2.1 Movie reviews**

This is a dataset for binary sentiment classification containing 40K movie reviews for natural language processing or Text analytics. This dataset consists of 20,000 highly polar movie reviews for training and 20,000 for testing. This dataset was collected by Stanford research [21].

**4.2.2 Conversation Dataset**

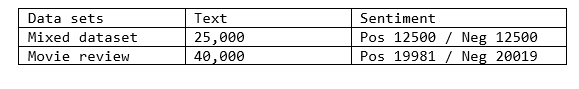
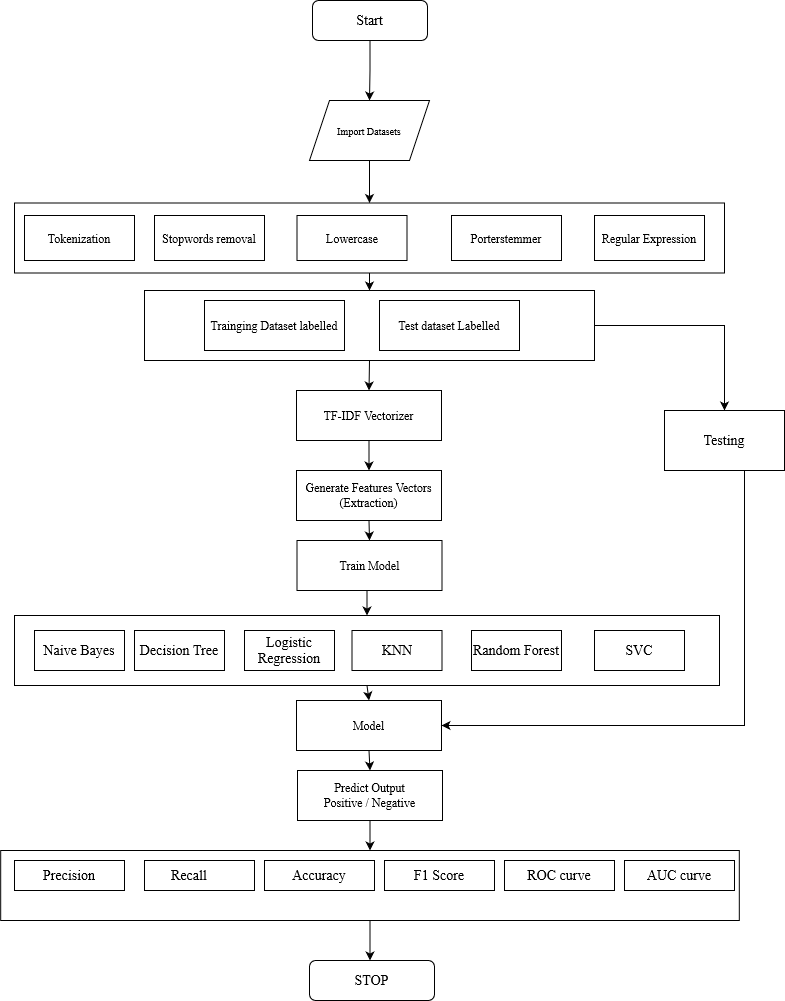
This dataset consists data of casual conversations to train model for sentiment analysis. This dataset consists of 25000 conversations.

Fig 4.5: Datasets used for our research.

### **4.3 Comparison Measure**

Let a represent the number of messages correctly classiﬁed as positive (i.e., true positive), b the number of negative messages classiﬁed as positive (i.e., false positive), c the number of positive messages classiﬁed as negative (i.e., false negative), and d the number of messages correctly classiﬁed as negative (i.e., true negative). In order to compare and evaluate the methods, we consider the following metrics, commonly used in information retrieval: true positive rate or recall: R=a/(a+c), false positive rate or precision: P=a/(a+b), accuracy: A= (a+d)/(a+b+c+d), and F-measure: F= 2 ·(P·R)/(P+R). We will in many cases simply use the F-measure, as it is a measure of a test’s accuracy and relies on both precision and recall. We report all the metrics listed above since they have direct interpretation in practice. The true positive rate or re-call can be understood as the rate at which positive messages are predicted to be positive (R), whereas the true negative rate is the rate at which negative messages are predicted to be negative. The accuracy represents the rate at which the method predicts results correctly (A). The precision rate, also called the positive predictive rate, calculates how close the measured values are to each other (P). We also use the F-measure to compare results, since it is a standard way of summarizing precision and recall (F). Ideally, a polarity identiﬁcation method reaches the maximum value of the F-measure, which is 1, meaning that its polarity classiﬁcation is perfect.

### **4.4 Flowchart**

****Fig 4.6: Flowchart for process of sentiment analysis

### **4.5 Tools and Platform**

1. Python
2. Google Colab
3. Jupyter Notebook
4. Django
5. VS code
6. Manjaro Linux and Windows 10
7. Nvidia 940mx
8. Intel 6200U

# **CHAPTER 5**

## **RESULT AND DISCUSSION**

### **5.1 RESULT ANALYSIS**

In this section we analyzed different comparison measures for 6 popular algorithms of Sentiment Analysis. Fig 5.1 and Fig 5.2 shows the measure i.e., Accuracy, F1 Score, Recall and Precision of different algorithms for SA. The analysis of Fig.5.1 was performed using conversation dataset and Fig 5.2 using Movie Review Dataset. We trained our own model for this analysis. The figure shows that among all these 6 algorithms Logistic regression gave higher value for all comparison measures.

Fig 5.1: Comparison of different performance measure of 6 different Algorithms for Movie dataset.

Fig 5.2: Comparison of different performance measure of 6 different Algorithms for Conversation Dataset.

Some people have this preconceived notion that logistic regression shouldn’t perform well because it is too simple. While that might be true for certain types of complex problems, but in our study, we only used datasets with binary classifier, true and false. So, that might be the reason as to why our model with logistic regression gave higher measures.

Another reason as to why Logistic regression algorithm compared better may be due to the fact that the Logistic Regression doesn’t make as many assumptions as of other algorithm such as naïve bayes.

But in the world of fine-grained Sentiment Analysis, Naïve Bayes and SVC are the widely used and popular ones. Our comparison study focused on detecting the polarity of content (i.e., positive and negative aﬀects) and does not yet consider other types of sentiments (e.g., psychological processes such as anger or calmness)). For this case these two algorithms might be the best option [11 18].

More important, we try to show that the prediction performance of method vary largely across datasets. It can be inferred from the results of this study that, using sufficient data set for predictive model construction can lead to better accuracy and the model’s ability to generalized. Although, due to vagueness that surrounds the size of the dataset, it is difficult to say precisely when a dataset can be considered to be big [8].

We also found that among these four no single methods is always best across diﬀerent text sources. As mentioned, for only detecting the polarity of a text, Logistic regression gave higher accuracy. Furthermore, for diverse categories of sentiment beyond positive and negative polarity logistic might not be the best option.

### **5.2 DATASET EVALUATION**

We evaluated the performance of our classification models used for our sentiment analysis using different amount of data. These are shown in the figure below:

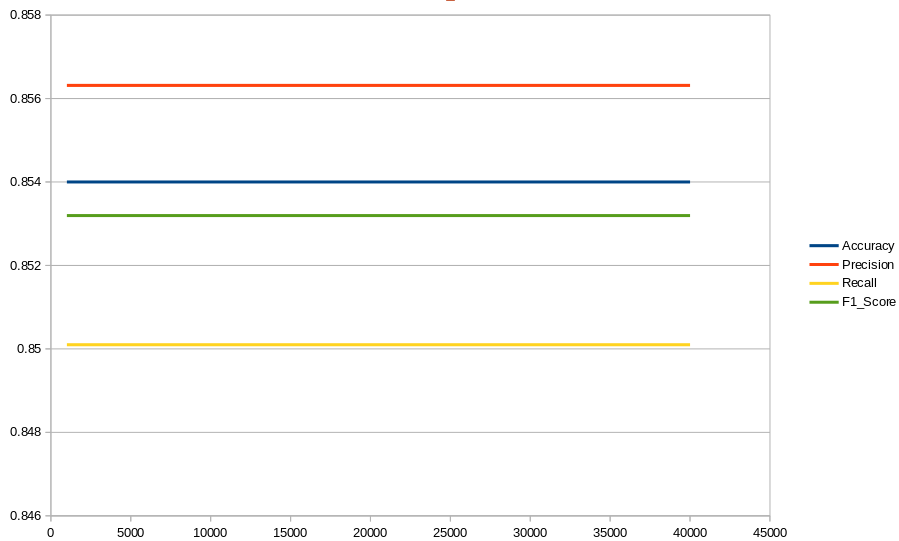


Fig 5.3: Comparison of different performance measure of Multinomial Naïve Bayes with respect to amount of data.

In figure 5.3, We can see that in Multinomial Naïve Bayes number of data does not affect the performance parameters because it needs enough data to understand the probabilistic relationship of each attribute in isolation with output variable. In other word Naïve bayes doesn’t need a lot of data to perform well.

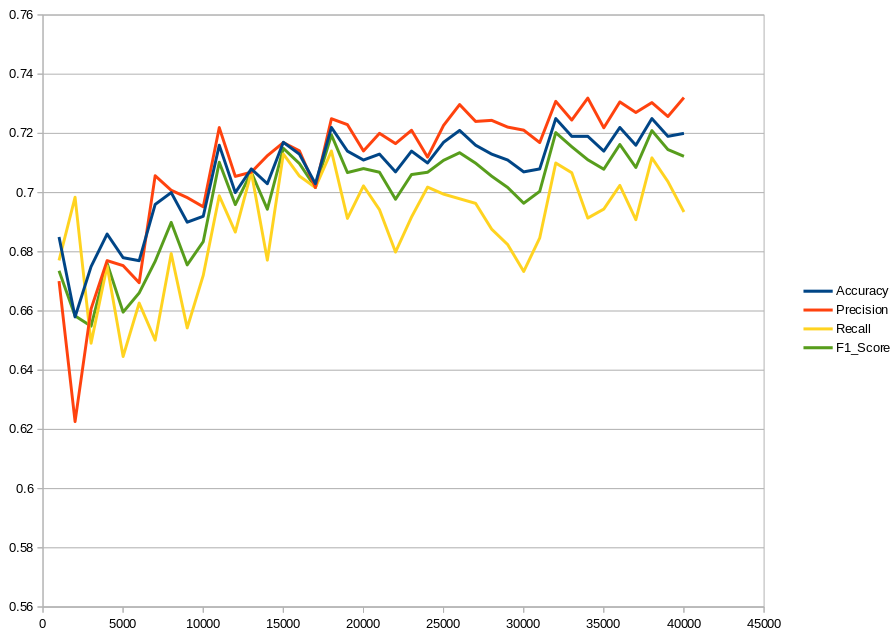


Fig 5.4: Comparison of different performance measure of Decision Tree with respect to amount of data.

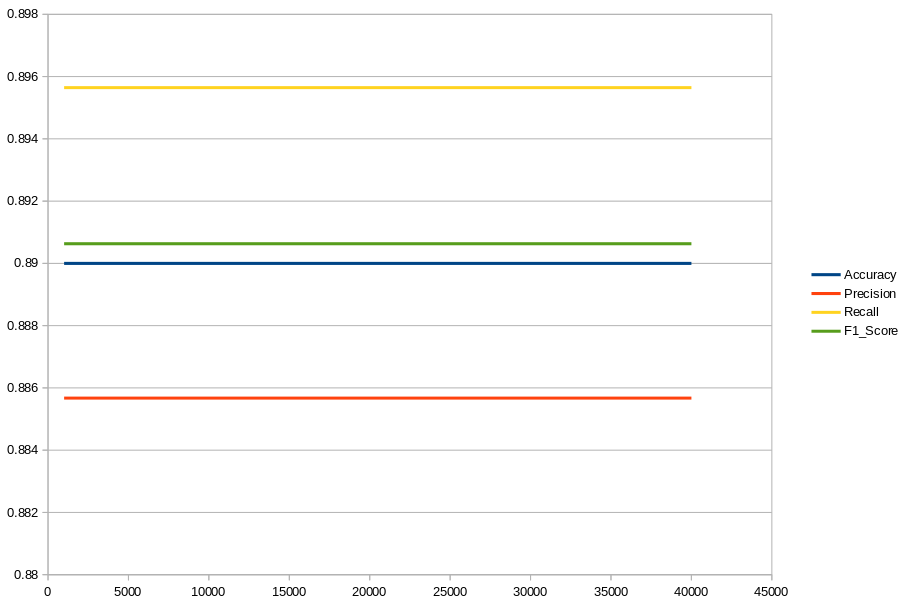
In figure 5.4, the performance parameter increases with increasing number of data.

Fig 5.5: Comparison of different performance measure of Logistic Regression with respect to amount of data

In figure 5.5, we can clearly see that whatever the interaction order, increasing the dataset size didn’t significantly affect model performance.

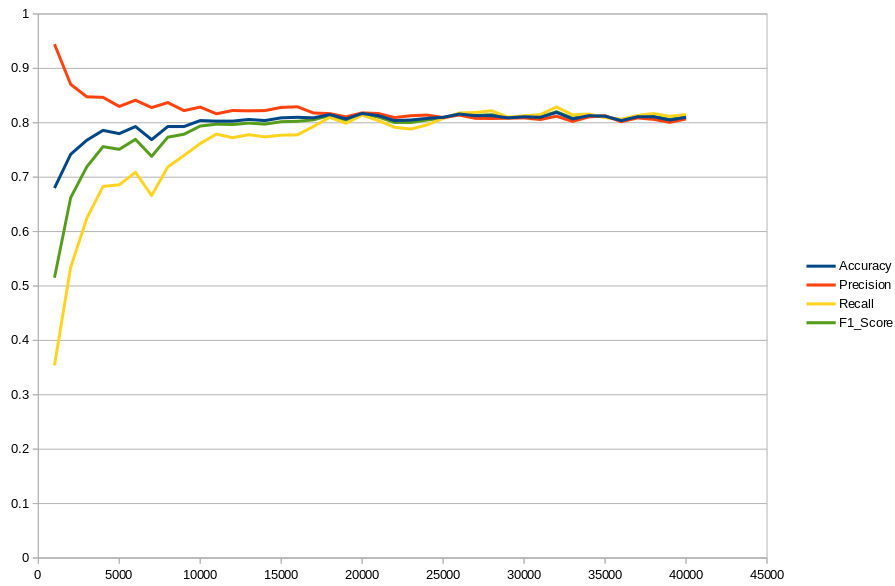


Fig 5.6: Comparison of different performance measure of KNN with respect to amount of data.

In figure 5.6, This chart shows that NN is near the optimal classifier at infinity but data is never infinity, so the larger the data better the performance parameters in KNN.

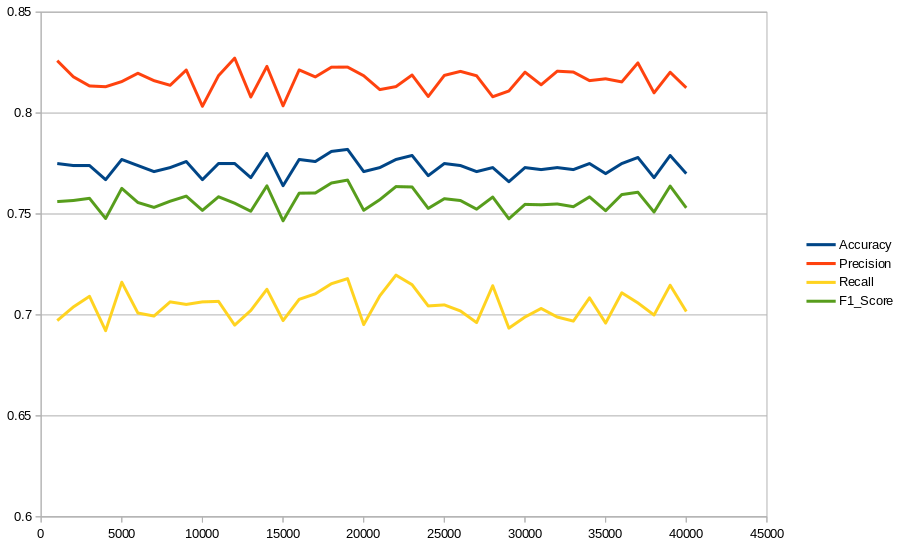


Fig 5.7: Comparison of different performance measure of Random Forest with respect to amount of data.

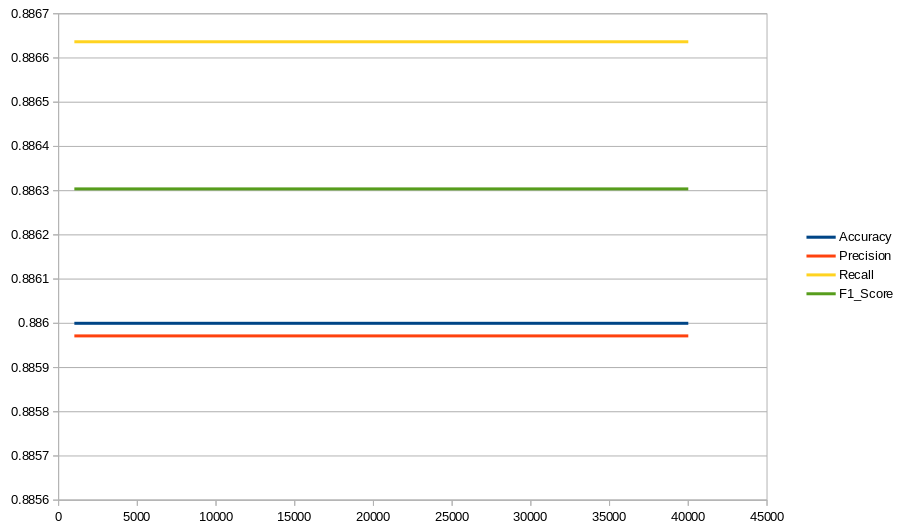


Fig 5.8: Comparison of different performance measure of Support Vector Classifier with respect to amount of data.

### **5.3 RESEARCH ANALYSIS**

The articles of sentiment analysis using naïve bayes presented in the survey are summarized in table 5.1

Table 5.1: Article Summary for Sentiment analysis using Naïve Bayes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| References | Year | Data Scope | Accuracy | Precision | Recall | F-score |
| [11] | 2016 | Twitter Feeds | 0.89 | 0.88 | 0.885 | 0.88 |
| [12] | 2013 | Facebook status | 0.78 | 0.77 | 0.68 | 0.72 |
| [13] | 2019 | Twitter Comments | 0.8343 | 0.5629 | 0.6513 | 0.5768 |

The articles of sentiment analysis using decision tree presented in the survey are summarized in table 5.2

Table 5.2: Article Summary for Sentiment analysis using Decision Tree

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| References | Year | Data Scope | Accuracy | Precision | Recall | F-score |
| [13] | 2019 | Twitter comments | 0.8291 | 0.8291 | 1 | 0.9066 |
| [14] | 2019 | Tweets | 0.63 |  |  |  |
| [15] | 2017 | Tweets | 0.8466 | 0.9596 |  |  |

The articles of sentiment analysis using logistic regression presented in the survey are summarized in table 5.3

Table 5.3: Article Summary for Sentiment analysis using Logistic Regression

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| References | Year | Data Scope | Accuracy | Precision | Recall | F-score |
| [1] | 2017 | Food Review | 0.934 | 0.93 | 0.93 | 0.93 |
| [16] | 2020 | Yelp Review | 0.949 |  |  | 0.902 |
| [17] | 2021 | Food Review | 0.8235 | 0.8878 | 0.8764 |  |

The articles of sentiment analysis using KNN presented in the survey are summarized in table 5.4

Table 5.4: Article Summary for Sentiment analysis using KNN

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| References | Year | Data Scope | Accuracy | Precision | Recall | F-score |
| [13] | 2019 | Tweets | 0.67 |  |  |  |
| [19] | 2017 | Tweeter data | 0.84 | 0.85 | 0.81 | 0.83 |

The articles of sentiment analysis using Random Forest presented in the survey are summarized in table 5.5

Table 5.5: Article Summary for Sentiment analysis using Random Forest

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| References | Year | Data Scope | Accuracy | Precision | Recall | F-score |
| [11] | 2016 | Tweets | 0.8291 | 0.56 | 0.333 | 0.30 |
| [20] | 2020 | Hindi Tweets | 0.902 |  |  | 0.6635 |

The articles of sentiment analysis using SVC presented in the survey are summarized in table 5.6

Table 5.6: Article Summary for Sentiment analysis using SVC

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| References | Year | Data Scope | Accuracy | Precision | Recall | F-score |
| [21] | 2020 | Social Media | 0.798 | 0.508 | 0.866 |  |
| [22] | 2021 | User Review | 0.875 | 0.88 | 0.88 | 0.88 |

The research data also proves our result analysis that the Logistic regression is better in term of accuracy, but again all these papers used different datasets and preprocessing techniques that hugely impact the accuracy but in general, Logistic Regression shows higher accuracy. The reason as to why might be similar to as explained above in this section.

### **5.4 THE SENTIMENT ANALYZER**

Finally, having compared the different sentiment methods and tested different performance parameter, we have created a python-based desktop application. This application takes text or voice (voice-to-text) as an input. A paragraph is divided as multiple sentences of which sentiment is calculated using 6 different algorithms as shown in Fig. 5.3. Also, the result tab will estimate the performance parameters of each algorithm used.

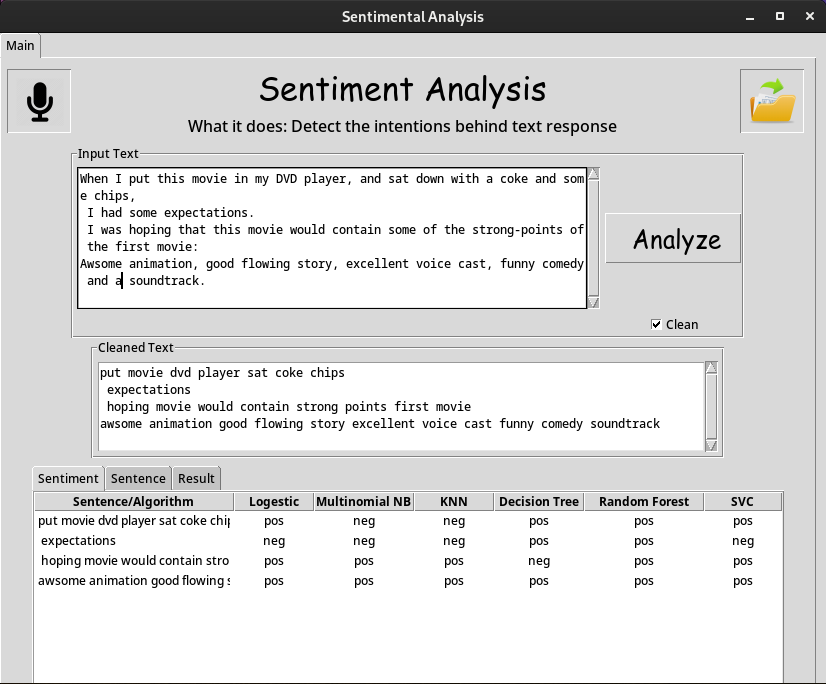


Fig 5.9: Screenshot of Sentiment Analysis Application

There are three tabs in our application.

1. Sentiment tab

In this tab, comparison of sentiment of each sentence is shown using different algorithm.

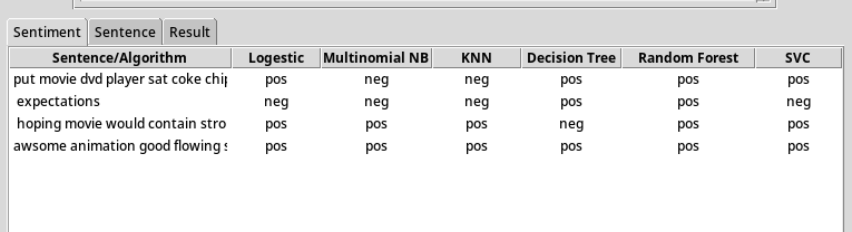


Fig 5.10: Sentiment Tab

1. Sentence tab

In this part, each sentence is shown on a separate line. And the sentiment of a given sentence is presented.

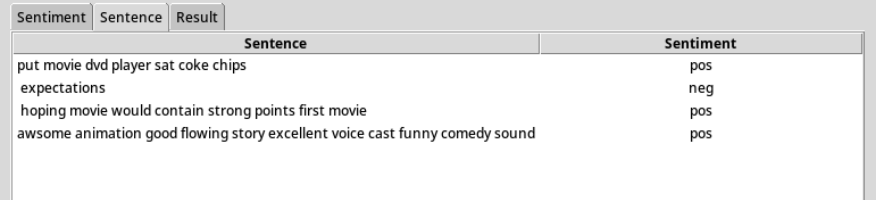


Fig 5.11: Sentence Tab

1. Result tab

In this part, performance measure on the basis of accuracy, precision, recall and f1-score of different algorithm are shown.

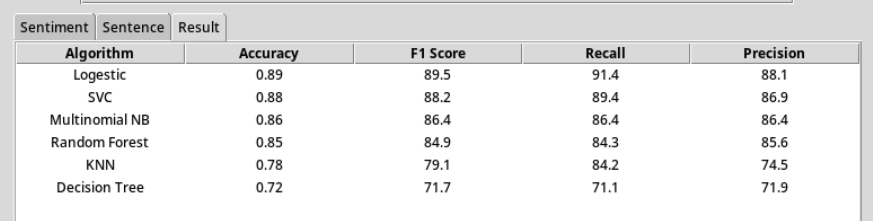


Fig 5.12: Result Tab

### **5.5 REST API**

We have also developed a REST API using Django that is similar to our Sentiment analyzer so that we can easily allow for interaction with RESTful web services.

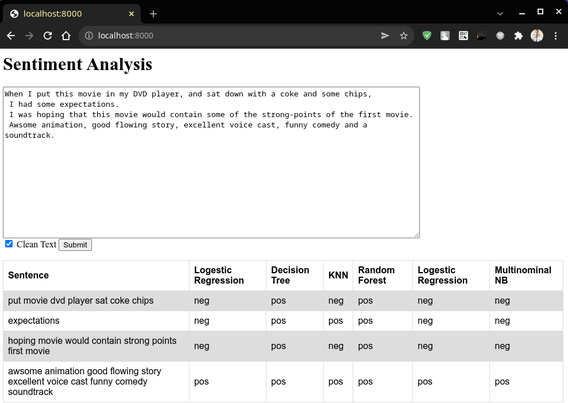


Fig 5.13: REST API

# **CHAPTER 6**

## **CONCLUSION AND FUTURE RECOMMENDATION**

### **6.1 Conclusion**

In this project, we have done the comparative study of different algorithms used in sentiment analysis. We have used Naïve Bayes, Decision Tree, Logistic Regression, KNN, Random Forest and SVC. We have measured performance on the basis of accuracy, precision, recall and F1 score. In our study we only used datasets with binary classifier. Through analysis done throughout the project we concluded that Logistic Regression performed best amount other five algorithms with the accuracy of 89.2% for conversation dataset and 89.1% for movie dataset.

Furthermore, we have concluded that, for diverse categories of sentiment beyond positive and negative polarity logistic might not be the best option. For this case Naïve Bayes and SVC outperformed are widely accepted algorithms.

### **6.2 Future Recommendation**

We can further enhance the performance of this project. And newer models of machine learning (such as Bidirectional Encoder Representations from Transformers (BERT)) can be used.

## **REFERENCES**

|  |  |
| --- | --- |
| [1] | Sentiment Analysis: A definitive Guide, Retrieved from: -Monkeylearn.com/sentiment-analysis/ |
| [2] | Avinash M and Sivasankar E, A Study of Feature Extraction techniques for Sentiment Analysis, 2019 |
| [3] | S. E. Saad and J. Yang, “Twitter Sentiment Analysis Based on Ordinal Regression”, 2019 |
| [4] | Pablo Gamallo, Marcos Garcia, “Citius: A Naïve-Bayes Strategy for Sentiment Analysis on English Tweets”, 2014 |
| [5] | Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social Media Data”, 2013 |
| [6] | A. Pak and P. Paroubek., Twitter as a corpus for Sentiment Analysis and Opinion Mining”, 2010 |
| [7] | Agarwal, B. Xie, I. Voysha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data”, 2011 |
| [8] | Han-Xiao Shi, Xiao-Jun Li, A Sentiment analysis model for hotel reviews based on supervised learning, 2011 |
| [9] | L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitter from Biased and Noisy Data”, 2013 |
| [10] | R. Xia, C.Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification”, 2011 |
| [11] | Anmol Nayak, Dr. S Natarajan, Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds, 2016 |
| [12] | Jaime Caro, Sentiment analysis of Facebook statuses using Naïve Bayes classifier for language learning, 2013 |
| [13] | Veny Amilia Fitri Rachmadita Andreswari, Muhammad Azani Hasibuan, Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm, 2019 |
| [14] | Dimple Tiwari, NAnhay Singh, Ensemble Approach for Twitter Sentiment Analysis, 2019 |
| [15] | Md Shoeb, Jaweb Ahmed, Sentiment Analysis and Classification of Tweets Using Data Mining, 2017 |
| [16] | Dehao Xhang, Sentiment Classification with Logistic Regression – Analyzing Yelp Reviews, 2020 |
| [17] | Mayur Wankhade, A Chandra Sekhara Rao, Suresh Dara, Baijnath Kaushik, A Sentiment Analysis of Food Review using Logistic Regression, 2017 |
| [18] | Dimple Tiwari, NAnhay Singh, Ensemble Approach for Twitter Sentiment Analysis, 2019 |
| [19] | Md Shoeb, Jaweb Ahmed, Sentiment Analysis and Classification of Tweets Using Data Mining, 2017 |
| [20] | Arnav Munshi, Sanchit Sapra, M.Arvindhan, “A Novel Random Forest Implementation of Sentiment Analysis”, 2020 |
| [21] | Shoffan Saifullah, Yuli Fauziah, Agus Sasmito Aribowo, “Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety based on social media data”, 2020 |
| [22] | Asad Khattak, Muhammad Zubair Asghar, Zain Ishaq, Waqas Haider Bangyal, Ibrahim A Hameed, “Enhanced concept-level sentiment analysis system with expanded ontological relations for efficient classification of user reviews”, 2021 |