Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766162

AN ANALYSIS OF MOVIE REVIEWS IN SOCIAL MEDIA DATA USING DATA MINING TECHNIQUES

S.V. Harshini[1], M. Archana[2], U.Latha[3] ,T.Velmurugan [4]

[1]Student, [2,3]Assistant Professor, [4]Associate Professor.

[1]Department of Computer Science and Design, [2,3]PG Department of Computer Applications (M. Sc IT & BCA), [4]PG and Research Department of Computer Science, [1]RMK Engineering College, Kavarapettai, Chennai,

[2,3,4]Dwaraka Doss Goverdhan Doss Vaishnav College, Arumbakkam, Chennai, India

[1]hars21120.cd@rmkec.ac.in,[2]archanadgvc@gmail.com, [3]dgvclatha@gmail.com,
[4]velmurugan_dgvc@yahoo.co.in,

Abstract: The entertainment for the real-world peoples is very huge and analyzing such kind of entertainments are not easy. One of the areas of this topic is movies. Different types of movie reviews are produced by various reviewers. Some of the movies create a high impact in the society as well as it changes some viewers mind to do better thinks and also behave in an improper way. The movie reviews are based on their understanding and usefulness of the information and the main theme adopted in the movies. The personal feelings are produced in the form of movie reviews. To analyze all such useful and un useful information is a tedious task. This research work utilizes the datamining techniques like Naïve Bayes, Support Vector Machine, Random Forest and Logistic Regression to find the pros and cons of the movie reviews. To visualize the feelings of the reviewers. This work identifies the best movie released and accepted based on the reviewer comments. Finally, the best algorithm is suggested by means of its accuracy and performance.

Keywords: Naïve Bayes Algorithm, Decision Tree Methods, Support Vector Machine Method, Random Forest Algorithm

1. Introduction

Internet facilitates interpersonal connections. People use the internet to voice their opinions through social media, blogs, movie reviews, product reviews, etc. Every day, users generate enormous amounts of data. The best kind of entertainment known to man is undoubtedly movies, and it is usual for individuals to watch movies and share their impressions on social media. By examining movie review data, we may discover a film's strong and weak points and determine whether it lived up to audience expectations [1]. A person always reads the review and rating of a film before deciding to watch it. Finding the movie's review is made easier with the aid of sentiment analysis (SA). SA is the process of extracting important information from a large body of data. It categorizes people's opinions as either positive or negative automatically.

Social Media Listening
Focus Groups & Interviews
Surveys
Email, Chat & Call Data
NPS
Customer Reviews
Blogs & Forums
Voice of the Customer

Figure 1: Vision of Social media users

The movie review dataset used in this study was taken from a social media website called Twitter [2]. In particular, classification algorithms are used to collect and analyze reviews of Tamil movies. To determine the dataset with the highest accuracy, machine learning algorithms such Naive Bayes, Support Vector Machine, Random Forest, and Decision Tree are utilized [3]. It aids in recommending the top film in the field and also strongly urges viewers to watch the best films among the rest.

The rest of the paper is organized as follows. Section 2 describes about the related work of the relevant information. The materials and methods used for this research work is explained in section 3. Section 4 explores the preprocessing methods and its results. The experimental results are explained in section 5. Finally, section 6 concludes the research work via its findings.

2. Review of Literature

The emotive movie reviews, which are gathered from various social media sites like Facebook, Instagram, etc., aid social media users in developing an understanding of a certain film. The methodologies and techniques of sentiment analysis are applied in the movie review dataset to find the best ones, according to the many research articles listed below. A research work done by Rahman et al. in [7], In which that the tweets that are gathered from social media are categorised using machine learning algorithms. The Bernoulli Nave Bayes (BNB), Decision Tree (DE), Support Vector Machine (SVM), Maximum Entropy (ME), and Multinomial Nave Bayes algorithms are among the five types of algorithms employed (MNB). The Multinomial Naive Bayes (MNB) algorithm has the greatest accuracy among these methods, at 88.5%.

The research paper carried out by Başarslana et al. in [8]. This research study examines how customers express their emotive reviews of movies on social media and how those reviews are extracted to determine whether they are good, negative, or neutral using classification algorithms. Naive Bayes, Support Vector Machine, Artificial Neural Network, and TF-DF and W2V modelling approaches are used with the datasets that have been selected (Word2Vec). This study found that, when compared to other algorithms, Artificial Neural Network methods had the highest accuracy.

Another research work titled as “Sentiment analysis of movie reviews using machine learning techniques”, carried out by baid et al. in [9], in which that the researcher collected the tweets about movies from the various social media websites like facebook, blogs and twitter to analyzed the data and the three classification techniques like Naïve Bayes, K-Nearest

Neighbour and Random Forest to find the best algorithm. Among the three algorithms the Naïve Bayes Algorithm provides the best accuracy of 81.45%.

A research work done by Daeli et al. in [10] The k-Nearest Neighbor, Naive Bayes, Support Vector Machine, and Random Forest machine learning techniques are used to collect and analyzed movie reviews. K-Nearest Neighbor achieved the greatest accuracy of 96.8% out of these techniques. Another paper titled as "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques", carried out by Tran et al.in [11], In this study, the researchers employ a variety of machine learning methods, such as Decision Trees, Naive Bayes, Support Vector Machines, Blending, Voting, and Recurrent Neural Networks, to perform sentiment analysis on the two distinct movie review datasets (RNN). The experimental findings have demonstrated that our suggestions can perform better, particularly the voting and RNN-based classification models, which can produce more accurate predictions.

The research paper carried out by Bandana and Rachana [12], According to this research study's findings, machine learning algorithms are employed to identify customers emotional reactions to films. Naive Bayes, Linear Support Vector Machine (SVM), and suggested heterogeneous features are just a few of the supervised techniques that are used. Finally, it is concluded that the suggested approach produces the most accuracy. A research work titled as "Detecting fake reviews through sentiment analysis using machine learning techniques", done by Elmurngi et al. in[13], In this study, supervised algorithms are used to identify false movie-related reviews left by customers. To identify the false reviews, Naive Bayes, Support Vector Machine, k-Nearest Neighbor IBK, Kstar, and Decision tree are employed. These methods are applied to assess the fictitious data, and it is discovered that Support Vector Machine performs significantly better than other algorithms.

Another research paper done by Kalaivani, P., and K. L. Shunmuganathan [14]. In this study of research work, the movie review dataset is compared using three supervised algorithms: SVM, Naive Bayes, and KNN. The performance of Support Vector Machine is superior to other algorithms, and it also offers 80% accuracy. The article comes to the conclusion that the consumer reviews of movies are analyzed using seven classification methods. The text-based accuracy is compared to the algorithms Naive Bayes, SVM, Maximum Entropy, Decision tree, KNN, Winnow, and Adaboost. The SVM performs best and offers the maximum accuracy, in the end.The research work titled as "Sentiment Analysis of Movie Review Using Machine Learning Techniques", done by Ramya et al. [15]. In this research paper the researchers analyzed the movie review data by using the machine learning algorithms of Support Vector Machine and Multinomial Naïve Bayes and Logistic Regression which are applied to compared the techniques to find the best method. Finally Multinomial Naïve Bayes yield the best result. A research paper carried out by Singh et al. in[16]. In which that the Modern machine learning classifiers for optimising sentiment analysis include Naive Bayes, J48, BFTree, and OneR. Three manually compiled datasets are used in the tests; two of them were obtained from Amazon and one from IMDB movie reviews. Examining and contrasting the effectiveness of these four classification strategies and OneR technique outperforms the others.

Table 1: A comparison of Various Methods

Paper Ref. No.	Researcher	Methods Used	Results & Accuracy
7	Rahman, Atiqur, and Md Sharif Hossen	Bernoulli Nave Bayes (BNB), Decision Tree (DE), Support Vector Machine (SVM), Maximum Entropy (ME), and Multinomial Nave Bayes algorithms	Multinomial Naïve Bayes yields 88.5% of accuracy.
8	Başarslana, Muhammet Sinan, and Fatih Kayaalpb	Naive Bayes, Support Vector Machine, Artificial Neural Network	Artificial Neural Network provides the best result
9	Baid, Palak, Apoorva Gupta, and Neelam Chaplot	Naïve Bayes, K-Nearest Neighbor and Random Forest	Naïve Bayes provides the best accuracy of 81.45%
10	Daeli, Novelty Octaviani Faomasi, and Adiwijaya Adiwijaya	k-Nearest Neighbor, Naive Bayes, Support Vector Machine, and Random Forest	k-Nearest Neighbor yields the greatest accuracy of 96.8%.
11	Tran, Duc Duy, Thi Thanh Sang Nguyen, and Tran Hoang Chau Dao	Decision Trees, Naive Bayes, Support Vector Machines, Blending, Voting, and Recurrent Neural Networks	Recurrent Neural Networks acquired the greatest accuracy
12	Bandana, Rachana	Naive Bayes, Linear Support Vector Machine (SVM) and proposed method	Proposed method provides the best method
13	Elmurngi, Elshrif, and Abdelouahed Gherbi	Naive Bayes, Support Vector Machine, k-Nearest Neighbor IBK, Kstar, and Decision tree	Support Vector Machine yields the highest accuracy
14	Kalaivani, P., and K. L. Shunmuganathan	Support Vector Machine, Naive Bayes, and KNN	Support Vector Machine provides the highest accuracy of 80%.
15	Ramya, V. Uma, and K. Thirupathi Rao	Support Vector Machine and Multinomial Naïve Bayes and Logistic Regression	Multinomial Naïve Bayes outperforms the others
16	Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh	Naive Bayes, J48, BFTree, and OneR	OneR acquires the best result.

3. Materials and Methods

Natural language processing tasks like classification rely on machine learning techniques. The most common classification task is sentiment analysis, although there are many more types as well. Because each algorithm is utilized to tackle a particular problem, each task frequently demands a unique algorithm. In this study, the classification methods I Bayes, Support Vector Machine, Random Forest, and Decision Tree are used to determine which of these algorithms performs the best overall.

Naïve Baye: One of the well-known classification machine learning methods, the I Bayes Algorithm helps to categorize the data based on the computation of conditional probability values. It uses class levels represented as feature values or vectors of predictors for classification and applies the Bayes theorem to the computation [4]. A quick algorithm for categorization issues is the I Bayes algorithm. Real-time prediction, multi-class prediction, recommendation systems, text categorization, and sentiment analysis use cases can all benefit from this technique.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

P(B|A) stands for Likelihood Probability, which quantifies the likelihood that a given hypothesis is true based on the available data.

Support Vector Machine: A supervised machine learning approach called Support Vector Machine (SVM) is used for both classification and regression [5]. Although we also refer to regression concerns, categorization is the most appropriate term. Finding a hyperplane in an N-dimensional space that clearly classifies the data points is the goal of the SVM method.

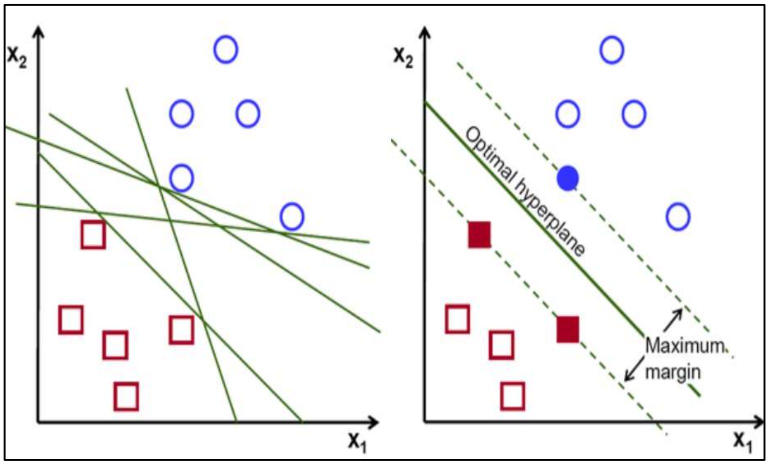

Figure 2: Workflow of Support Vector Machine

Figure 2 explains that the comparison of linear and logistic regression models, Support Vector Machines (SVM) attain a substantial degree of accuracy with less computational power. The SVM looks for a hyperplane that clearly classifies the data with the greatest amount of margin. Support vectors, which are utilized to maximize the margin, are datapoints that are close to the hyperplane. Various data points are disregarded.

Random Forest: Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression [6]. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance.

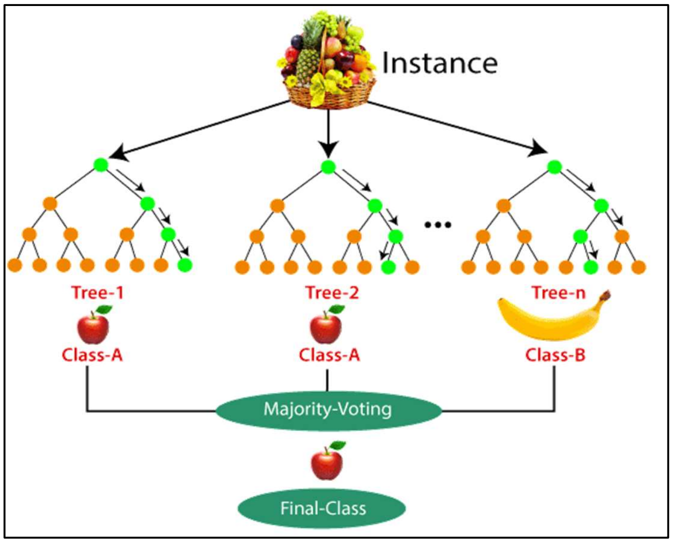

Figure 3: Random Forest classification example

Figure 3 shows that there is a dataset with several fruit photos. Therefore, the Random Forest classifier receives this dataset. Each decision tree is given a portion of the overall dataset. Each decision tree generates a prediction result during the training phase, and the Random Forest classifier predicts the outcome based on the majority of results when a new data point is encountered.

Decision Tree Algorithms: The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label.

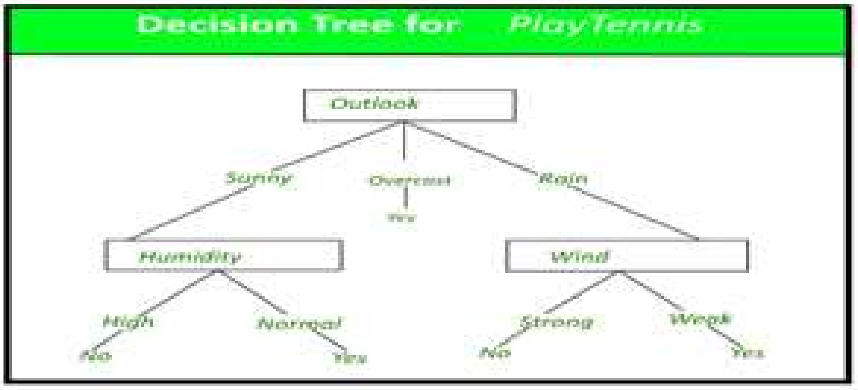

Figure 4: Example for Decision Tree Algorithm

Figure 4 explains that by dividing the source set into subgroups based on an attribute value test, a tree can be “trained”. It is known as recursive partitioning to repeat this operation on each derived subset. When the split no longer improves the predictions or when the subset at a node has the same value for the target variable, the recursion is finished. Decision tree classifier building is ideal for exploratory knowledge discovery because it doesn’t require parameter configuration or domain understanding. High-dimensional data can be handled via decision trees. Decision tree classifiers are often accurate. A popular inductive method for learning classification information is decision tree induction.

System Flow: Figure 5 elaborates that the analysis of the movie review training data that have been gathered from social media such as Twitter, Facebook, Instagram, websites, blogs etc. The term "overall" refers to how viewers rated the film (1 being the lowest evaluation and 5 being the highest evaluation used in this survey). Referring to this study's findings regarding user opinions of the review's use and value is beneficial.

The workflow explains how the preprocessing techniques of stop word removal, stemming, parts of speech tagging, tokenization, and named entity recognition are used to process the movie review dataset that is gathered from the various websites. I Bayes, Support Vector Machine, Decision Tree, and Random Forest classification algorithms are applied to the processed data to determine the accuracy.

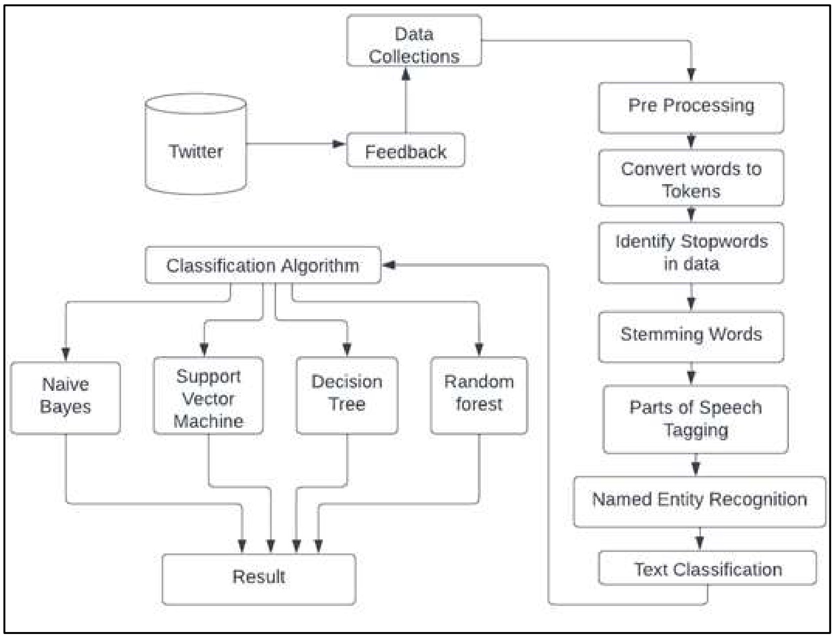

Figure 5: Architecture of the Research work

4. Preprocessing Methods

Preparing text data for machines to use in activities like analysis, prediction, etc. is known as text pre-processing. Text pre-processing involves a variety of phases, and several libraries can be used to get rid of things like stop words, stemming, and tokenization.

Convert Words to Tokens: Tokenization is the division of text into a collection of meaningful fragments. These objects are known as tokens. For instance, in a selected dataset, the text data can be broken down into chunks of text, words, and sentences. The researcher can specify their own criteria to split the input text into relevant tokens depending on the work at hand.

Identify Stop words into Data: Any human language has an abundance of stop words. By eliminating these terms and the basic information from our text, we can draw attention to the crucial details. Because there are fewer tokens involved in training, the removal of stop words obviously reduces the size of the dataset and, consequently, the training time.

Stemming Words: Stemming is a technique used to get rid of any kind of suffix from a word and bring it back to its root form, although occasionally the root word produced by stemming is meaningless or does not belong in the English lexicon.

For Example, the words “helpful”,”helped”, ”helping” , after the stemming process the words will be changed in to “help” .

Parts of Speech Tagging: It involves breaking down a sentence into its component parts, such as a list of words or a list of tuples, each of which has a form (word, tag). The part-of-speech tag “in case of” indicates if a word is a noun, adjective, verb, etc.

Named Entity Recognition: The most common data preprocessing activity is named entity recognition (NER). It entails locating important information in the text and classifying it into a number of predetermined categories. A constant subject of discussion or reference in a book is referred to as an entity.

Results of Pre-processing: Twitter movie reviews are used in this study’s analysis. As was previously said, the input data is first preprocessed before the Nave Bayes, Decision Tree, Random Forest, and SVM models are applied. There are four sections in this paragraph. It first describes the input data, then describes the dataset, evaluates the outcomes, and then presents the comparison outcomes.

Input Data: The methodologies employed in this research project include I Bayes, decision trees, Random Forest, and SVM algorithm. Various tamil movie reviews from Twitter are the dataset utilised, along with the training dataset downloaded from the website, to determine which is more accurate. The use of pre-processing techniques is previously covered in the section above.

Both linguistic and non-linguistic data are present in the movie reviews. Where only linguistic information is considered when varying machine learning algorithms classify the provided text. Based on the people reviews posted on twitter platform, the comments are examined and processed to determine good, negative, and neutral reviews using text data. Table 1 displays the sample dataset.

Table 1: Sample Dataset of Movie Reviews

Month	**Movie Name**	**Review Text**
September	Ponniyin Selvan	Movie is awesome, very good screenplay all the actors done their role very mass.
August	Thiruchitrambalam	Good performance by everyone. Had few feel-good moments
October	Sardar	It would have outgrossed
June	Vikram	The unexpected action sequence made me speechless
September	Venthu Thaninthathu Kaadu	STR has put in a lot of hard work for this film and can celebrate the 50th day of the festival very happily.
July	Yaanai	Stunning performance as always! Such a feel good movie

August	Viruman	Worst writing, worst comedy scenes, worstest cringe elements, worst debut for Adithi, worst interval, worst BGM.
April	Beast	The filming technique is very unassuming
August	Cobra	This was the worst movie and it also received the least amount of applause after few weeks
March	Etharkum thuninthavan	Neither boring nor interesting.

Twitter _latest_tamil_movie_reviews_2022 is the name of the dataset relation, which contains 4721 instances, 212 characteristics, and a total weight of 4721. The dataset is divided into parts and categorised using 11 cross-validations based on detailed accuracy with class. In the dataset, each attribute has two or more different values. The table below displays the same dataset with regard to Movie and Review Text.

Table 2: Weightage of Reviews

No.	**Class**	
	Movie Name	***Count***
1	Ponniyin Selvan	996
2	Thiruchitrambalam	320
3	Sardar	500
4	Vikram	890
5	Venthu Thaninthathu Kaadu	410
6	Yaanai	439
7	Viruman	290
8	Beast	410
9	Cobra	129
10	Etharkum thuninthavan	337

In table 2 displays the name of the movie with the count which specifies the total number of tweets which are given by the reviewers. These are classified and analyzed for suggesting the best movie to the social media users.

A bar graph is a particular style of graphical display of the data in which bars of uniform width are created with equal spacing between them on one axis (often the x-axis), displaying the variable. The height of the bars serves as a representation of the variables' values. In Figure 6 elaborates that the graphical representation shows the various tamil movie name and the total count of tweets which are given by the social media users or reviewers.

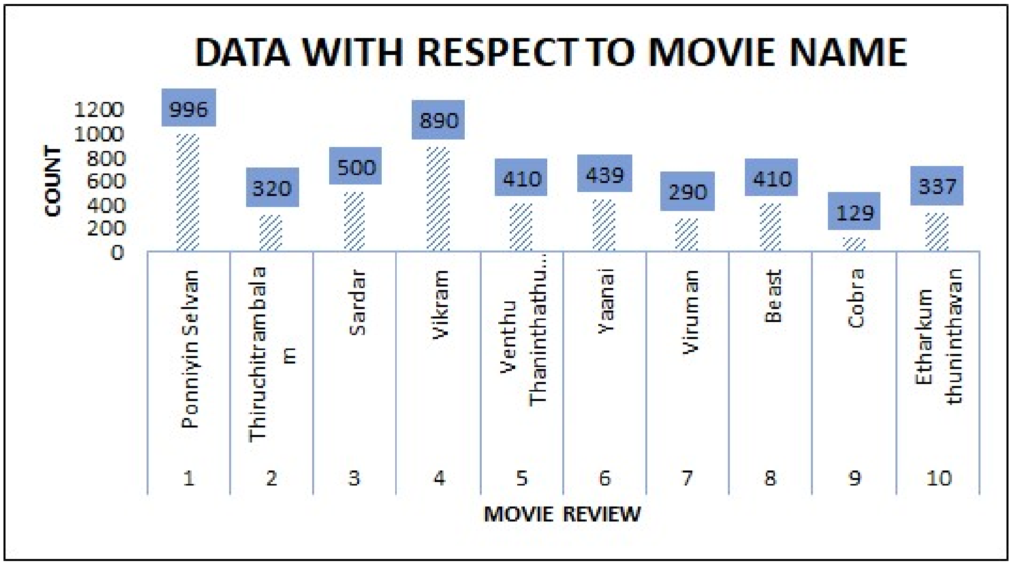

Figure 6: Graphical representation of Movie Reviews

Table 3: Review of Text with respect to Data

No.	**Class**	
	Review text	***Count***
1	Linguistic	3256
2	Non-lingistics	1465

In table 3, it shows that the linguistic and non-linguistic texts are identifies from the dataset. In Linguistic, specific variables, such as worst, best and so forth, have values composed of linguistic notions (sometimes referred to as linguistic words) rather than numbers. For example, Let's outline TWEETS as a linguistic parameter

TWEETS= { “Worst”, “Wonderful”, “Excellent”}

Each linguistic phrase used in a tweet has a membership function for a particular range. Each function maps the same value to several membership values between 0 and 1. The comment's status can then be determined using these membership values and identified the positive and negative tweets.

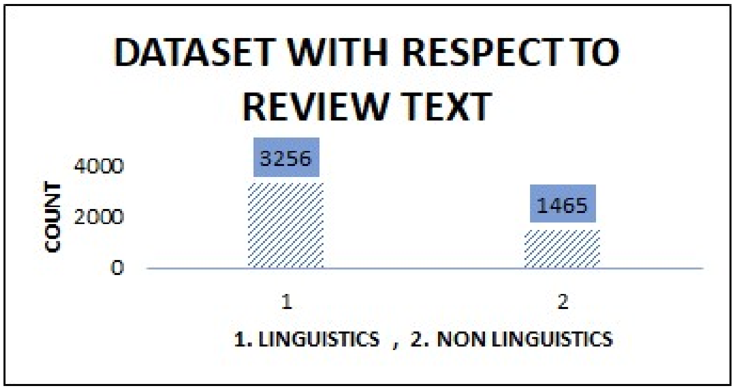

Figure 7: Dataset with respect to Review Text

The graphical representation of the dataset with respect to review text is shown in this figure 7. The linguistics method which identifies the text-based reviews and the non-linguistics method is found count of emojs in the text reviews. The review text dataset for the classes that were provided. The classes are separated into linguistic and non-linguistic metrics. The classification of evaluations based on emojis differs from that of text reviews, according to linguistics.

5. Experimental Results

The twitter_latest_tamil_movie_review is used in this section's implementations of the I Bayes, decision tree, random forest, and SVM algorithms. This table includes the experimental findings from each individual algorithm. This dataset's accuracy in I Bayes, Decision Tree, Random Forest, and Support Vector Machine algorithms is 93.17%, 92.20%, 96.14%, and 99.9%, respectively. Compared to the I Bayes method, decision tree algorithm, and random forest algorithm, the weighted average of the support vector machine algorithm produces better results. SVM outperforms the other three algorithms in terms of accuracy. As a result, the accuracy of the support vector machine method is better than the accuracy of the I bayes algorithm, the decision tree algorithm, and the random forest algorithm.

Result Comparison: The table and graphical representation demonstrate the comparative outcomes of all four algorithms.

Table 4: Performance Measure

	Naïve Bayes	*Decision Tree*	*Random Forest*	*SVM*
TP Rate	0.627	0.765	0.627	0.827
FP Rate	0.627	0.299	0.627	0.827
Precision	0.399	0.699	0.403	0.786
Recall	0.627	0.77	0.627	0.827
F-Measure	0.486	0.754	0.486	0.691
ROC Area	0.3	0.815	0.4	0.9
PRC Area	0.528	0.821	0.528	0.928

Table 4 contains the various performance metrics for all four methods of classification techniques are Naïve Bayes, Decision tree, Random forest and Support Vector Machine in relation to the chosen dataset. The various performance metrics are given below.
True Positive (TP) values are those that are both real and anticipated positive values.
False Positive (FP) values are those that are projected to be positive but are actually negative.
False Negative (FN) values are ones that are projected to be negative but are actually positive.
Values that are both genuinely negative and expected to be negative are referred to as True Negatives (TN).

Precision: The number of positive class forecasts that actually fall into the positive class is measured by precision. Precision is determined by dividing the total number of true positives and false positives by the imbalanced classification problem's two classes. The outcome is a number that ranges from 0.0 (zero precision) to 1.0 (full or perfect precision).

Precision = True Positive / (True Positive + False Positive)
This model's accuracy is calculated as follows:
Precision = 80/ (80 + 40)
Precision = 80 / 120
Precision = 0.6

Recall: Recall measures how many accurate class predictions were made using all the accurate examples in the dataset. Recall measures how many accurate class predictions were made using all the accurate examples in the dataset. Recall is determined by dividing the total number of true positives by the sum of true positives and false negatives in a two-class unbalanced classification issue.

Recall = True Positive / (True Positive + False Negative)

The outcome is a number that ranges from 0.0 for no memory to 1.0 for complete or perfect recall. A model provides predictions, 90 of which are accurate for the positive class and 10 which are not. For this model can compute the recall using the formula below.

Recall = True Positive / (True Positive + False Negative)
Recall = 80 / (80 + 20)
Recall = 80 / 100
Recall = 0.8

F Measures: Precision and memory issues are balanced in a single number by F-single Measure's score.

The formula for the conventional F measure is as follows:

F Measure = (2 × Precision × Recall) / (Precision + Recall)

A perfect F-Measure score, for instance, would be produced by a perfect precision and recall score.

F Measure = (2 × Precision × Recall) / (Precision + Recall)
F Measure = (2 × 1.0 × 1.0) / (1.0 + 1.0)
F Measure = (2 × 1.0) / 2.0
F Measure = 1.0

ROC: The performance of a classifier for each potential threshold is shown on a graph called the ROC. The real positive rate (on the y axis) and the false positive rate are shown on a graph (on the x axis).

PRC: A simple graph with Precision values on the y-axis and Recall values on the x-axis is what makes up a PR curve. In other words, the TP/(TP+FN) on the y-axis and the TP/(TP+FP) on the x-axis are present in the PR curve.

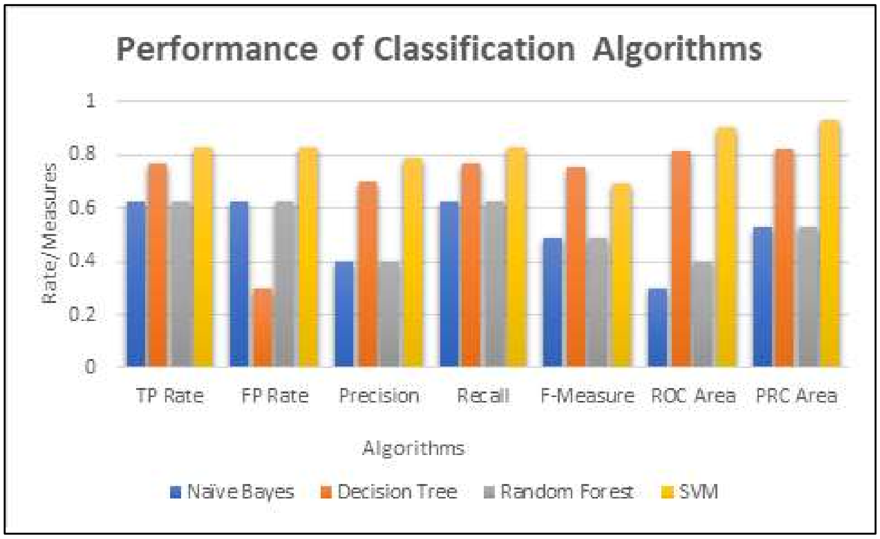

Figure 8: Performance of Classification Algorithm

Figure 8 displays the classification algorithms of Naïve Bayes, Support Vector Machine, Decision Tree and Random Forest performance are analyzed and classified into a graphical format. Among the other methods Support Vector Machine which indicates the highest range in chart.

Table 5: Accuracy of Classification Algorithm

Classification Algorithm	Accuracy (%)
Naïve Bayes	93.17
Decision Tree	92.61
Random Forest	96.24
Support Vector Machine	99.99

According to Table 5, the Support Vector Machine method outperforms the Naïve Bayes, Decision Tree, and Random Forest algorithms for text analysis.

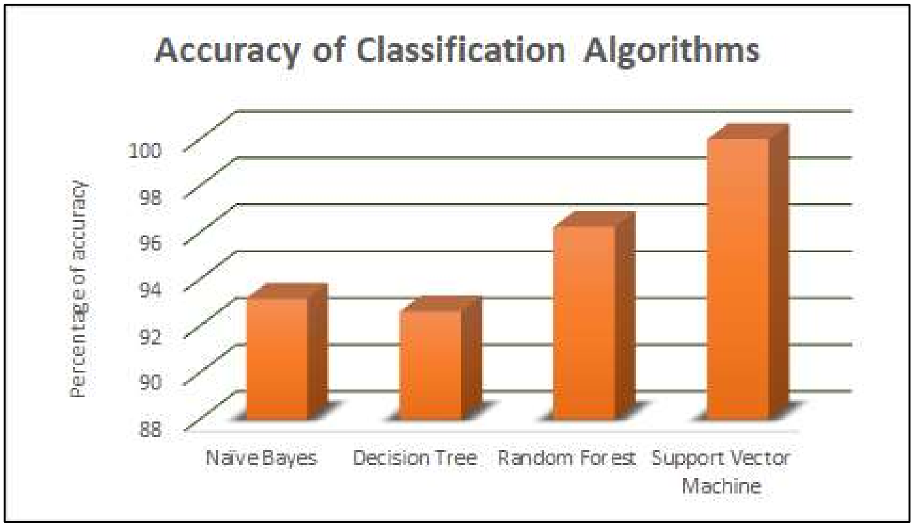

Figure 9: Graphical representation of Classification Results Accuracy

Figure 9 shows the performance analysis of each of the four algorithms is displayed graphically. Compared to the naive bayes algorithm, decision tree algorithm, and random forest algorithm, the support vector machine algorithm is more accurate.

6. Conclusion

Currently, many types of reviews are carried out for the better understanding of the information provided in the Social Medias like twitter, Facebook, Instagram etc. One such information is taken for the analysis in this work. Particularly, this work analyzed the Tamil language movie reviews data which is taken from the different repositories. Data mining algorithms such as Naïve Bayes, Support Vector Machine, Random Forest and Decision tree algorithms are applied to find the accuracy of the algorithms for reviewing the text based movies reviews. Taken data set was preprocessed and then the modified data to be analyzed. The analysis was carried out by considering the Positive, Negative and Neutral commands of the reviewers. The words are categorized based on the text provided by the movie reviewers. The performance of the algorithms for the text based information on accuracy was resulted. This research work identify that the Support Vector Machine method yields the better results compared with the other algorithms in terms its accuracy.

References

[1] Ramanathan, Vallikannu, T. Meyyappan, and S. M. Thamarai, “Sentiment Analysis: An Approach for Analysing Tamil Movie Reviews Using Tamil Tweets”, *Recent Advances in Mathematical Research and Computer Science*, Vol.no.3 pp. 28-39, 2021.

[2] Sheik Abdullah, A., K. Akash, J. ShaminThres, and S. Selvakumar, “Sentiment Analysis of Movie Reviews Using Support Vector Machine Classifier with Linear Kernel Function”, *In Evolution in Computational Intelligence*, pp. 345-354, 2021.

[3] Başarslan, Muhammet Sinan, and Fatih Kayaalp, “Sentiment analysis on social media reviews datasets with deep learning approach”, *Sakarya University Journal of Computer and Information Sciences 4*, Vol. no. 1, pp. 35-49, 2021.

[4] Gandhi, Usha Devi, Priyan Malarvizhi Kumar, Gokulnath Chandra Babu, and Gayathri Karthick, "Sentiment analysis on twitter data by using convolutional neural network (CNN) and long short term memory (LSTM)", *Wireless Personal Communications*, pp. 1-10, 2021.

[5] He, Lu, Tingjue Yin, and Kai Zheng, "They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets", *Journal of Biomedical Informatics*, pp. 132, 2022.

[6] Kumar, AV Mohan, M. Suhas, and Noah Fedrich. "Sentiment Analysis on Twitter Data", *International Conference on Cognitive and Intelligent Computing*, Springer Nature, Vol. 1, p. 441,2021.

[7] Rahman, Atiqur, and Md Sharif Hossen, "Sentiment analysis on movie review data using machine learning approach", IEEE *International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1-4, 2019.

[8] Başarslana, Muhammet Sinan, and Fatih Kayaalpb, "Sentiment Analysis with Machine Learning Methods on Social Media", Vol. 5, 2015.

[9] Baid, Palak, Apoorva Gupta, and Neelam Chaplot, "Sentiment analysis of movie reviews using machine learning techniques", *International Journal of Computer Applications 179*, no. 7, pp. 45-49, 2017.

[10] Daeli, Novelty Octaviani Faomasi, and Adiwijaya Adiwijaya, "Sentiment analysis on movie reviews using Information gain and K-nearest neighbor", *Journal of Data Science* and Its Applications, Vol. 3, no. 1, pp. 1-7, 2020.

[11] Tran, Duc Duy, Thi Thanh Sang Nguyen, and Tran Hoang Chau Dao. "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques." In Proceedings of Sixth International Congress on Information and Communication Technology, 2022, pp. 361-369.

[12] Bandana, Rachana, "Sentiment analysis of movie reviews using heterogeneous features", IEEE 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), pp. 1-4, 2018.

[13] Elmurngi, Elshrif, and Abdelouahed Gherbi, "*Detecting fake reviews through sentiment analysis using machine learning techniques*", IARIA/data analytics, pp. 65-72, 2018,

[14] Kalaivani, P., and K. L. Shunmuganathan, "Sentiment classification of movie reviews by supervised machine learning approaches", *Indian Journal of Computer Science and Engineering, Vol. 4*, no.4, pp. 285-292, 2013.

[15] Ramya, V. Uma, and K. Thirupathi Rao, "Sentiment analysis of movie review using machine learning techniques." *International Journal of Engineering & Technology,* Vol. 2, no.7 pp. 676-681, 2017

[16] Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh, "Optimization of sentiment analysis using machine learning classifiers", Human-centric Computing and information Sciences, Vol. 7, no. 1, pp. 1-12, 2017.

[17] Sharma, Anuj, and Shubhamoy Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis", *Proceedings of ACM research in applied computation symposium*, pp. 1-7, 2012.

[18] Singla, Zeenia, Sukhchandan Randhawa, and Sushma Jain, “Sentiment analysis of customer product reviews using machine learning”, IEEE *International conference on intelligent computing and control (I2C2)*, pp. 1-5, 2017.

[19] Kumar, KL Santhosh, Jayanti Desai, and Jharna Majumdar, “Opinion mining and sentiment analysis on online customer review”, IEEE *International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1-4, 2016.

[20] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh, “Sentiment analysis on product reviews using machine learning techniques”, *Cognitive Informatics and Soft Computing,* pp. 639-647, 2019.

[21] Kumari, Upma, Arvind K. Sharma, and Dinesh Soni, “Sentiment analysis of smart phone product review using SVM classification technique”, IEEE *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 1469-1474. 2017.

[22] Chauhan, Chhaya, and Smriti Sehgal, “Sentiment analysis on product reviews”, IEEE *International Conference on Computing, Communication and Automation (ICCCA)*, pp. 26-31, 2017.

[23] Janhavi, N. L., Jharna Majumdar, and Santhosh Kumar, “Sentiment Analysis of Customer Reviews on Laptop Products for Flipkart”, *International Research Journal of Engineering and Technology (IRJET), Vol. 5*, no. 03, pp. 629-634, 2018.

[24] Pujari, Chetana, and Nisha P. Shetty. “Comparison of classification techniques for feature-oriented sentiment analysis of product review data”, *Data Engineering and Intelligent Computing*, Springer, pp. 149-158, 2018.

[25] Shivaprasad, T. K., and Jyothi Shetty, “Sentiment analysis of product reviews: a review", *International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 298-301, 2017.

[26] Alsolamy, Afnan Atiah, Muazzam Ahmed Siddiqui, and Imtiaz Hussain Khan, "A Corpus Based Approach to Build Arabic Sentiment Lexicon", *International Journal of Information Engineering & Electronic Business*, Vol. 11, No. 6, 2019.

[27] Kumar, Akshi, and Teeja Mary Sebastian, "Sentiment analysis on twitter", *International Journal of Computer Science Issues (IJCSI)*, Vol. 9, No. 4, pp. 372, 2012.

[28] Darwich, Mohammad, Shahrul Azman Mohd, Nazlia Omar, and Nurul Aida Osman, "Corpus-Based Techniques for Sentiment Lexicon Generation: A Review", *Journal of Digital Information Management*, Vol.17, No. 5, pp. 296, 2019.

[29] Chathuranga, P. D. T., S. A. S. Lorensuhewa, and M. A. L. Kalyani, "Sinhala sentiment analysis using corpus-based sentiment lexicon", IEEE *19th international conference on advances in ICT for emerging regions (ICTer)*, Vol. 250, pp. 1-7, 2019.

[30] Abdulla, Nawaf A., Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based", IEEE *Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pp. 1-6, 2013.

[31] Cordeiro, Cheryl Marie, "A corpus-based approach to understanding market access in fisheries and aquaculture: a systematic literature review", *International Journal of Economics and Management Engineering,* Vol. 13, no. 10, pp. 1324-1333, 2019.

[32] Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh, "Optimization of sentiment analysis using machine learning classifiers", *Human-centric Computing and information Sciences,* Vol. 7, no.1, pp.1-12, 2017.

[33] Neethu, M. S., and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques*",* IEEE *Fourth international conference on computing, communications and networking technologies (ICCCNT)*, pp. 1-5, 2013.

[34] Le, Bac, and Huy Nguyen, "Twitter sentiment analysis using machine learning techniques", *In Advanced Computational Methods for Knowledge Engineering,* Springer, pp. 279-289, 2015.

[35] Hemalatha, I., GP Saradhi Varma, and A. Govardhan, "Sentiment analysis tool using machine learning algorithms", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* Vol. *2*, no. 2, pp. 105-109, 2013.

[36] Brar, Gurshobit Singh, and Ankit Sharma, "Sentiment analysis of movie review using supervised machine learning techniques", International Journal of Applied Engineering Research, Vol. 13, no. 16 , pp. 12788-12791, 2018.

[37] Sihwi, Sari Widya, Insan Prasetya Jati, and Rini Anggrainingsih, "Twitter sentiment analysis of movie reviews using information gain and naïve bayes classifier," IEEE International Seminar on Application for Technology of Information and Communication, pp. 190-195, 2018.

[38] Brar, Gurshobit Singh, and Ankit Sharma, “Sentiment analysis of movie review using supervised machine learning techniques”, *International Journal of Applied Engineering Research, Vol. 13*, no.16, pp. 12788-12791, 2018.

[39] Mitra, Ayushi, “Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset)”, *Journal of Ubiquitous Computing and Communication Technologies*, Vol. 2, no. 03, pp. 145-152, 2020.

[40] Basari, Abd Samad Hasan, Burairah Hussin, I. Gede Pramudya Ananta, and Junta Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization", *Procedia Engineering* 53, pp.453-462, 2013.