**SENTIMENT ANALYSIS ON**

**MOVIE REVIEWS IN TAMIL LANGUAGE**

Mini Project Report Submitted to the Department of Computer Applications, Bharathiar University in partial fulfilment of the requirements for the award of the degree of,

**MASTER OF SCIENCE IN DATA ANALYTICS**

Submitted By,

**VIGNESH R**

**(21CSEG37)**

Under the Guidance of,

**Prof. Dr. V. BHUVANESWARI, M.C.A., M.Phil., Ph.D.,**

Professor, Department of Computer Applications.

****

**DEPARTMENT OF COMPUTER APPLICATIONS**

**BHARATHIAR UNIVERSITY**

**COIMBATORE – 641 046**

**DECEMBER 2022**

**DECLARATION**

**DECLARATION**

I hereby declare that the Mini-project work titled **“SENTIMENT ANALYSIS ON MOVIE REVIEWS IN TAMIL LANGUAGE”** submitted to the Department of Computer Applications, Bharathiar University, Coimbatore is a record of original project work done by **VIGNESH R (21CSEG37)** under the supervision and guidance of **Prof**. **Dr. V. BHUVANESWARI, M.C.A., M.Phil., Ph.D.,** Professor, Department of Computer Applications, Bharathiar University, Coimbatore and this project work has not previously formed the basis of the award of any Degree/Diploma/Associate ship/Fellowship or similar title to any candidate of any University.

Place: Coimbatore Signature of the Candidate

Date:  **(VIGNESH R)**

Countersigned by Guide

**CERTIFICATE**

**CERTIFICATE**

This is to certify that the Mini-Project titled **“SENTIMENT ANALYSIS ON MOVIE REVIEWS IN TAMIL LANGUAGE”** submitted to Bharathiar University in partial fulfilment of the requirement for the award of the degree of the **Master of Science in Data Analytics** is a record of the original work done by **VIGNESH R (21CSEG37)** under my supervision and guidance and this project work has not formed the basis for the award of any Degree/Diploma/Associate ship/Fellowship or similar title to any candidate of any University.

Place: Coimbatore

Date:

Project Guide Head of the Department

Submitted for the university Viva-Voce Examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_

Internal Examiner External Examiner

**ACKNOWLEDGEMENT**

**ACKNOWLEDGEMENT**

Union is Strength. It gives me a great pleasure to acknowledgement with gratitude to personalities, without their help the completion of this project work would not been possible.

I express my respectful thanks to **Prof**. **Dr. T. DEVI., M.C.A., M.Phil., Ph.D., (UK),** Professor and Head, Department of Computer Applications, Bharathiar University, Coimbatore, for permitting me to carry out my mini project work.

I really deem it a special privilege to convey my prodigious and everlasting thanks to my guide **Prof. Dr. V. BHUVANESWARI, M.C.A., M.Phil., Ph.D.,** Professor, Department of Computer Applications, Bharathiar University, for her valuable guidance and suggestions to this Mini-Project work.

I extended my sincere thanks to **Mrs. P. SUDHANDRADEVI and Ms. M. LISSA,** Ph.D.,Research Scholars, Department of Computer Applications, Bharathiar University, for helping me by providing new ideas and support to complete my project.

Finally, I express my thanks to my dear parents and my dear friends for their support and encouragement for the successful completion of this project. I am highly obliged to those who have helped me directly and indirectly in making this project a successful one.

**CONTENT**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **DESCRIPTION** | **PAGE NO** |
|  | Synopsis | 1 |
| 1. | Introduction | 3 |
| 2. | Literature Review | 6 |
| 3. | Methodology & Frameworks | 11 |
|  | 3.1 Data Source  3.1.1. YouTube Comments Scraping  3.2 Lexicon Based  3.2.1. Text Pre Processing  3.2.2. Corpus Creation  3.2.3. Sentiment Scoring  3.2.4. Proposed Methods  3.3 Algorithm Based  3.3.1. Model Validation  3.3.2. Vectorization  3.3.3. Proposed Models  3.4 Evaluation Methods  3.5 Deployment |  |
| 4. | Results & Discussions | 38 |
| 5. | Conclusion | 45 |
|  | Future Enhancements |  |
|  | References |  |

**SYNOPSIS**

**SYNOPSIS**

Tamil (தமிழ்) is a classical language and one of the main languages of the Dravidian language family. Nowadays, more people are commonly using their mother tongue to express their opinions and views on social media. Especially Indian languages are often used in social media messages. However, there are not as much work in sentiment analysis has been done for Indian languages like Hindi, Tamil, Kannada, etc. In this work, we focus on Tamil-language YouTube comments. our objective is to classify the sentiment of the Tamil movies based on YouTube comments in Tamil using Tamil SentiWordNet (TSWN). The primary sentiments categorization of Tamil films is identified using domain-specific ontology. In contextual semantics, the sentiment of a word may flip based on the neighboring word. So, we have used unigram, bigrams, and trigrams approaches for understanding sentiment of the comments in Tamil.

The sentiment scoring has been applied by lexicon based and algorithm- based methods. In lexicon-based method we further proceeded into five new approaches. Among those word count method and manual scoring method are based on the actual positive and negative corpus. Vader sentiment intensity analyzer, text blob and Afinn are based on the translated positive and negative corpus. In algorithm-based method the text document has been vectorized and it’s performance has been verified using various classification algorithms like Logistic Regression, Multinomial Naive Bayes, Random Forest and Support Vector Machine. Their performances were noted and compared.

.

**INTRODUCTION**

**CHAPTER - I**

**INTRODUCTION**

**1.1 Movie Reviews**

The Tamil cinema industry, which known as Kollywood, has drawn more attention in recent years. This is due to fact that they started producing world-level films in these years. So, reviewing Tamil movies will be important work in the field of Tamil natural language processing. The Internet performs a major role in gathering the reviews. Movie reviews are significant way to measure the performance of a movie. While assigning a numerical rating or "stars" to a movie tells us about its quantitative success (or) failure, Group of movie reviews gives us a deeper qualitative insight into distinct aspects of the movie. A textual movie review tells us about the strong and weak points of the movie, and a deeper analysis of a movie review can tell us if the movie meets the prospects of the reviewer.

**1.2 Sentiment Analysis**

The Web has changed the world extremely, and the rise of Web 2.0 has totally transformed the situation as people can now express their thoughts and feelings digitally. For example, if someone wants to purchase a cell phone, they can use the Internet to read buyer reviews about the product, and if somebody plans to watch a movie, they can read reviews about the film. The internet has given users the choice to express themselves in the form of reviews or blogs, which can be positive or negative. Users’ reviews are a kind of data in which some unseen patterns are analysed to predict the behaviour of clients and identify new opportunities in the market. This process is called sentiment analysis and is known as "opinion mining." Sentiment analysis, also called opinion mining, is the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards things such as products, services, administrations, individuals, issues, events, topics, and their attributes. It represents a large problem space.

**1.3 Sentiment Analysis on Tamil Movie Reviews**

Tamil is one of the first-born languages in the world. This language is spoken by around 78 million people in the world. It is permitted as an official language in Tamil Nadu (India), Singapore, and Sri Lanka. It is also one of the ancient Dravidian languages, which follows Subject Object Verb (SOV) pattern The order may change to construct the sentences in Tamil. It is probable to construct sentences in Tamil using only verbs, subjects and verbs, or subjects and objects. Over the last span, less attention has been paid to sentiment analysis of Tamil-language reviews. Sentiment analysis is prevalent in the English language, but in Indian languages it’s rare. Sentiment analysis aims to expose the author's attitude toward a particular theme from the written text. Other terms used to denote this research area include "opinion mining" & "subjectivity detection". It used for natural language processing and machine learning methods to find statistical & linguistic patterns in the text that reveal attitudes. It has gained fame in recent years due to its fast applicability in business environments, such as summarising feedback from product reviews, discovering collaborative recommendations, or assisting in election campaigns. Developments in Text to Speech, Syntactic Parsing, Character Recognition, Spell Check, and other technologies are examples of the advancement of artificial intelligence and natural language processing in Tamil language. The approaches to bilingual data bases, electronic Tamil libraries, Tamil on mobile devices, Tamil search engines, and processing sizable Tamil data bases are emerging recently. Some go into detail about computer-assisted Tamil teaching techniques, Tamil blogging, and Wikipedia content development in Tamil. In comparison to previous years, the web is now enriched with non-English languages as well. There are many scenarios where the equivalent words may be used in several contexts, and context-dependent word mapping is still a hard task that is error-prone and needs manual efforts to find the accurate polarity of words.

**LITERATURE REVIEW**

**CHAPTER - II**

**LITERATURE REVIEW**

Tripathi and S. k. Trivedi (May 2006) had described SENTIWORDNET, a lexical resource in which each WORDNET synset is associated with three numerical scores. These scores define how objective, positive, and negative the terms contained in the synset are. The method based on the quantitative analysis of shines associated with synsets. It was endowed with a web-based graphical user interface and is freely available for research purposes.[9]

Kumar, Anand M., (12, Sept. 2010) had developed the Novel approach to morphological analyser is based on sequence labelling and training by kernel methods. It captures the non-linear relationships and various morphological features of Tamil language in a better and simpler way. The efficiency of our system is compared with the existing morphological analysers which are available in net. Regarding the accuracy our system significantly outperforms the existing morphological analyser and achieves a very competitive accuracy of 95.65% for Tamil language.[2]

Taboada, Maite ,( June 2011) offer a method for extracting sentiment from text that is lexicon-based. The Semantic Orientation CALculator (SO-CAL) makes use of dictionaries containing words that have been annotated with their semantic orientation (polarity and strength), as well as intensification and negation. SO-CAL is used to do this. They demonstrate that SO-CAL performs consistently across domains and with completely new data. Additionally, they go over the dictionary-making process and how we use Mechanical Turk to ensure the consistency and dependability of our dictionaries.[5]**\**

The goal of the book of Liu, Bing (2012) is to give an in-depth introduction to this fascinating problem and to present a comprehensive survey of all important research topics and the latest developments in the field. This book is suitable for students, researchers, and practitioners who are interested in social media analysis in general and sentiment analysis in particular. Lecturers can readily use it in class for courses on natural language processing, social media analysis, text mining, and data mining.[6]

Mathan, thangarasu, and Manavalan(7 July 2013) has proposed a light stemmer for Tamil text is to handle inflectional morphology word. This stemmer removes suffixes from a word to get stem word. From the computational result it is prove light stemmer approach is more suitable stemmer for Tamil language compare to a rule-based suffix stripping stemmer.[3]

Tripathi and S. K. Trivedi, ( 2016) were used Bayesian Classifier has been used in this study for testing feature selection mechanics. This classifier is trained on the words/features of the corpus extracted using five feature selection algorithms (Chi-square, Info-gain, Gain-Ratio, One-R and relief attribute) and a comparative study have been performed amongst them. The classifier and feature selection approaches were evaluated by two different metrics (F-Value, and False Positive).[10]

Se, Shriya & Ravi, Vinayakumar & Kumar, M. & Kp, Soman (2016). were presented a proposal for classifying the Tamil movie reviews using supervised algorithms, namely SVM, Maxent classifier, Decision tree and Naive Bayes. Feature are extracted and are given to the system for better classification and to improve the accuracy of the system. The TamilSentiwordnet (word) features are influencing the file level sentiment analysis in a greater level. The SVM classifies better than other classifiers.[19]

R. Naidu, S. K. Bharti, K. S. Babu and R. K. Mohapatra (2017) were proposed a two-phase sentiment analysis for Telugu news sentences using Telugu SentiWordNet. Initially, it identifies subjectivity classification where sentences are classified as subjective or objective. Objective sentences are treated as neutral sentiment as they don’t carry any sentiment value. Next, Sentiment Classification has been done where the subjective sentences are further classified into positive and negative sentences. With the existing Telugu SentiWordNet, our proposed system attains an accuracy of 74% and 81% for subjectivity and sentiment classification respectively.[11]

Thilagavathi, R., and Mrs. K. Krishnakumari(2018) were identified Differentiating key sentences from trivial ones may be helpful to improve sentiment analysis. Inspired by this observation in this paper they propose a novel framework to estimate the sentiment polarity of reviews by virtue of opinion lexica and key sentences automatically extracted from unlabeled data. This framework cannot only overcome the problem of excessive dependence on external resources, but also is able to capture the overall sentiment polarity of reviews. [13]

Thavareesan, Sajeetha, and Sinnathamby Mahesan(2018) have discovered Performance of SA model depends on the pre-processing steps such as negation handling techniques and stop words removal. Word level feature representation using TF-IDF and Word2vec gives notable improvement compared to feature representation techniques using character level and word level feature representation using presence of words, BoW or TF. SVM and RNN give best results when compared to other classifiers.[18]

Abdelrahman, S.( 2019), has gathered The Tamil Tweets from Twitter API and are translated into English Tweets using Google Translator. Sentiment Analysis is employed to determine the polarity of English and Tamil tweets, which is therefore bilingual. Words are classified using the Support Vector Machine to get the sentiment of the words.[14]

Jagdale, Rajkumar & Shirsath, Vishal & Deshmukh, Sachin.( 2019) Introduced New opinion mining algorithms proposed in their research to collect customer opinions from social media and classify those using basic, hybrid and ensemble classification techniques. Dataset size gets reduced after eliminating redundant features. Preprocessing has also been done after collection of reviews to remove unwanted words, punctuations, semicolon, tabs etc. The proposed method applies POS tagger to separate all verbs, nouns, pronouns, adverbs and adjectives.[15]

Ramanathan, Vallikannu‘s (2019)objective is to classify the sentiment of the Tamil movies based on Tamil tweets using Tamil SentiWordNet (TSWN). They proposed Term Frequency - Inverse Document Frequency (TF-IDF) method to find the sentiment polarity of the Tamil movie dataset. This method provides baseline for our research. Domain specific ontology is used to identify the primary sentiment categorization of the Tamil movies.[16]

Abhishek, Kumar (2020) aims to use Sentiment Analysis on a set of movie reviews given by reviewers and try to understand what their overall reaction to the movie was, i.e. if they liked the movie or they hated it. They aim to utilize the relationships of the words in the review to predict the overall polarity of the review. Logistic Regression Model with feature set using mixture of Unigrams and Bigrams is best. Apart from this, one can also use a Naïve Bayes’ Classifier or a SGD classifier as they also provide good accuracy percentage. One peculiar thing to note is low accuracy with Random Forest classifier. This might be because of over-fitting of decision trees to the training data. Also, low accuracy of kNN Classifiers shows us that people have varied writing styles and kNN Models are not suited to data with high variance.[8]

Chakravarthi, Bharathi Raja, (2020) created a gold standard Tamil-English code-switched, sentiment-annotated corpus containing 15,744 comment posts from YouTube. In their paper, they describe the process of creating the corpus and assigning polarities. They preprocessed the comments by removing the emoticons and applying a sentence length filter. They want to create a code-mixed corpus of reasonable size with sentences that have fairly defined sentiments which will be useful for future research.[4]

Novendri, Risky’s (2020) study seeks to examine and categorize the types of comments made by YouTube users to understand how the sentiments of users can impact the first day revenue. Such a conclusion can act like a predictive model for us and tell the movie makers, distributors that how the movie would perform financially the very next day.[12]

In Chakravarthia , Bharathi‘s(2020) shared task The teams detected offensive language using various machine learning and deep learning classification models.This paper has analysed those benchmark systems to find out how well they accommodate a code-mixed scenario in Dravidian languages, focusing on Tamil and Malayalam. Teams used methods ranging from Bag of Words, TF-IDF to BERT-based models to represent the data and applied conventional machine learning algorithms, deep neural networks and transformer networks for prediction.[1]

In their Paper Regina and P. Sengottuvelan(2021) includes sentiment analysis of movie reviews using feature-based opinion mining and supervised machine learning. In this paper, the main focus is to determine the polarity of reviews using nouns, verbs, and adjectives as opinion words. Reviews will be Classified into two different categories positive and negative. Reviews of Open Movie Database is used as source data set and Natural Language Processing Toolkit for Part of Speech Tagging. This paper also contains some facts about the classification of data on basis of polarity.[7]

Ramanathan, Vallikannu & Thirunavukkarasu, Meyyappan & Thamarai, S (2021 ),used TF-IDF method to find the accuracy based on keywords. To improve the performance, domain specific ontology has been applied to get the result. In the proposed method we have created our own Tamil movies sentimental-bearing terms ontology to categorize the movies. To analyze the semantic meaning between the words contextual semantic sentiment analysis is applied. Considering the neighbouring word would change the polarity of the tweet. When we used contextual semantic sentiment analysis, it deals negation problem also. In this paper they used Tamil SentiWordNet with adjectives to classify the sentiment.[17]

**METHODOLOGY AND FRAMEWORK**

**CHAPTER - III**

**METHODOLOGY AND FRAMEWORK**

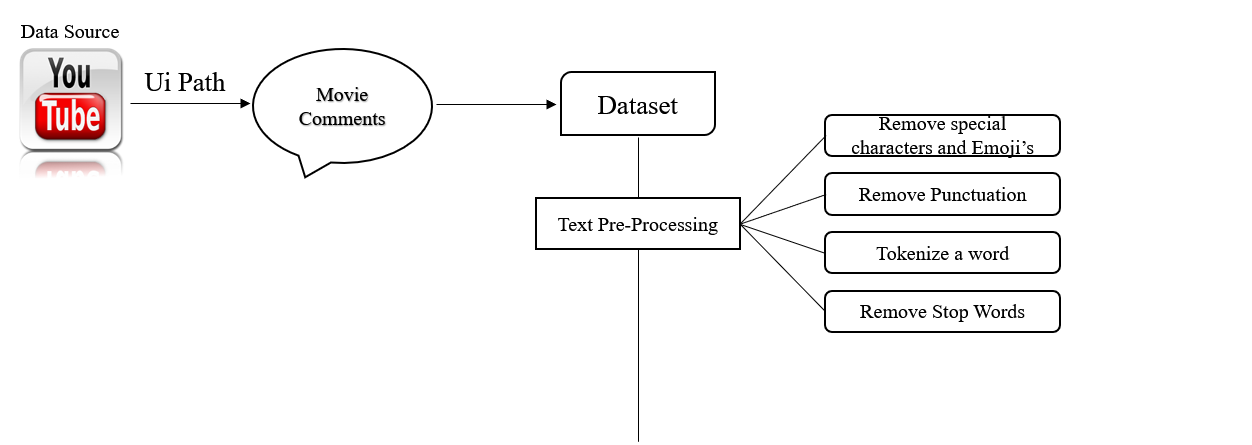
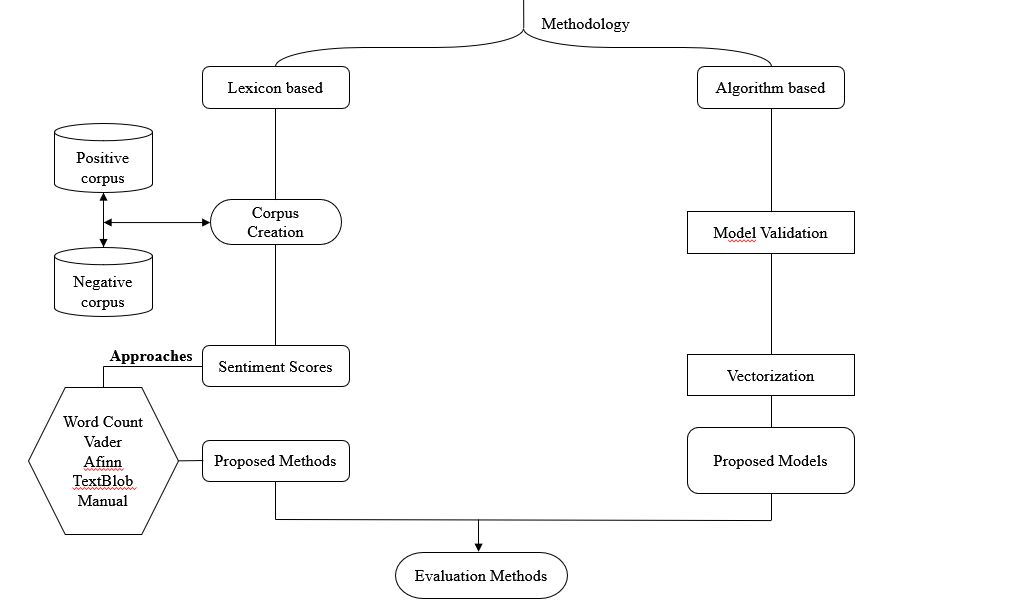
**3.1 Problem Statement**

There exist very few systems that calculate the sentiment associated with Tamil text, as sentiment analysis is very difficult for the Tamil language due to the different complexity associated with Tamil text. Well-annotated standard corpora are still not available for the Tamil language. The Tamil language lacks efficient resources such as a parser and tagger, which are required for sentiment extraction. TamilSentiWordNet (TSWN), like the well-known English SentiWordNet, is available but has a limited number of adjectives and adverbs that need to be improved for greater accuracy.

Further there are more new words were used in movie reviews. So, identifying them and understanding their sentiment has become more complex. Also, there are no packages available for analysing and calculating the sentiment of the movie reviews. So, we can’t directly analyse the sentiment of the Tamil Language text.

**3.2 Objective**

Our objective is to detect whether the movie will succeed or not based the reviews on YouTube given by the people who watched the movie. In precise we work on the movie reviews which are written in direct Tamil [தமிழ்] language. We proceeded with algorithm based and lexicon-based methods for calculating the sentiment of the movie reviews. The concept is to create a corpus from the scraped YouTube reviews and translating them into English. Then to do sentiment scoring by various approaches to do sentiment classification. Finally, to predict the success of the movie.

**3.3 FRAMEWORK**

**Fig. 1. Flowchart Methodology**

**3.3.1 Data Source**

A) Web Scrapy Using UI Path Studio

For data collection we have used UiPath studio for scrapping the reviews of the Tamil Movies which were commented on YouTube reviews.

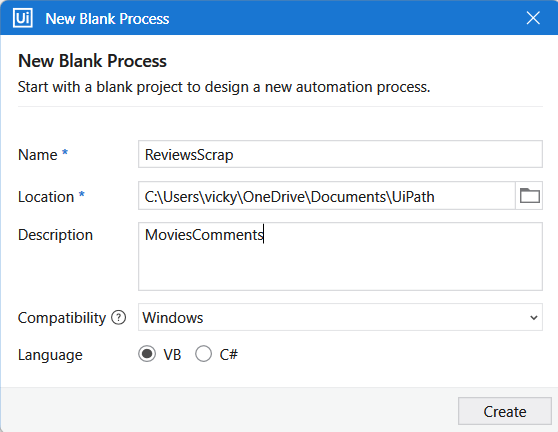
* UiPath can automatically log into the browser, **extract data** what you select to scrapping, **filter** and **transform the data into the required format. UiPath is very much helpful in scrapping the data which are present in structured format.**
* **We have scraped twelve recent movie’s comments for corpus creation and training the model.**

**PROCESS OF UIPATH**

* Select the Site and the Data Source.

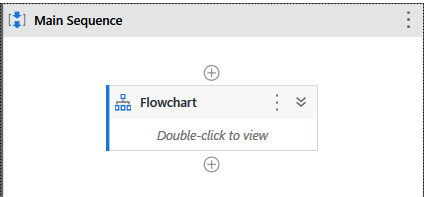
We Select the YouTube Movie review link in order to scrap the comments of those movies.

* we created a Blank process and opened new Sequence in the UiPath shown on Fig. 2.



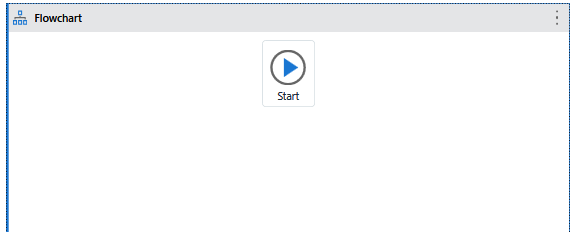
**Fig. 2. Process Begin**

* Then we created a flowchart file by Drag and Drop on the main workflow shown in Fig. 3.



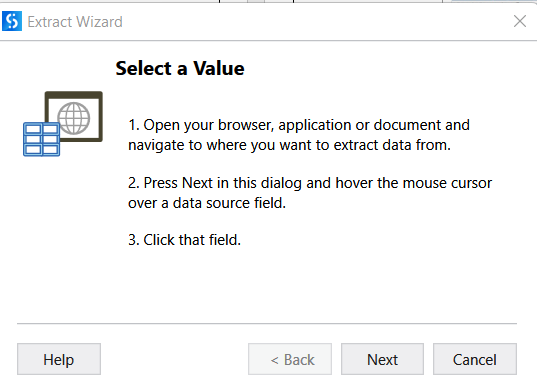
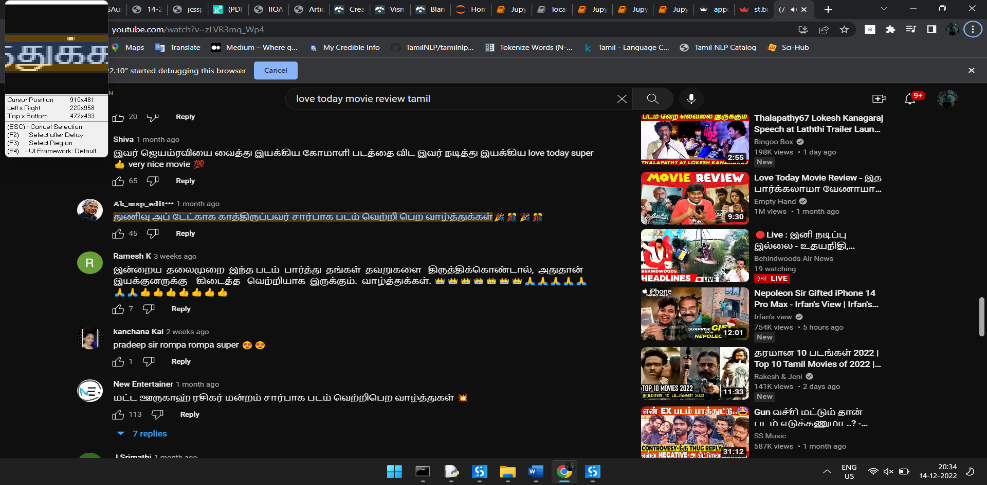
**Fig. 3. Flowchart Drop**

* We have designed the flow for scraping the contents using the data scrap button at the top of toolbar on UI path. Fig. 4, shows start the scrap process on Ui path.



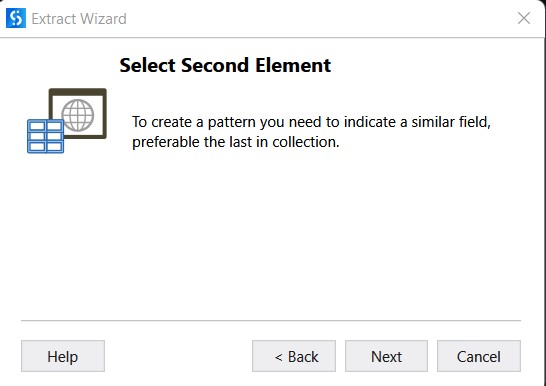
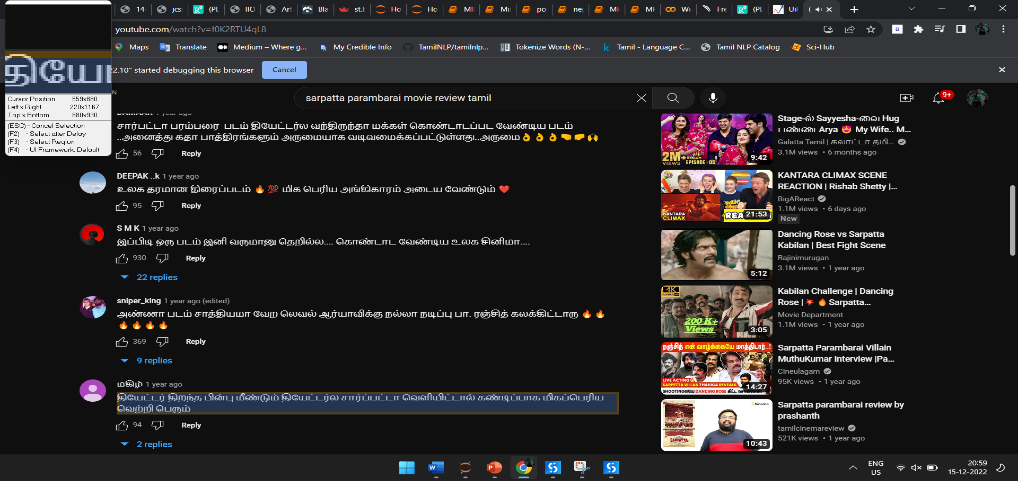
**Fig. 4. Start the Scraping**

* We chose the chrome web Browser to open YouTube and search based on movie review’s URL for scraping the comments.
* We selected the comment which is on the URL to extract by clicking the pop-up Select a Value as shown on Fig.5.



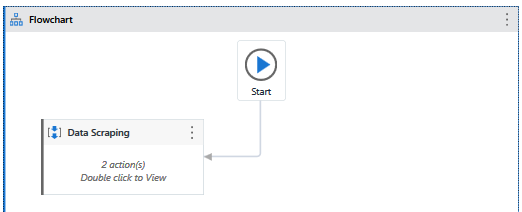
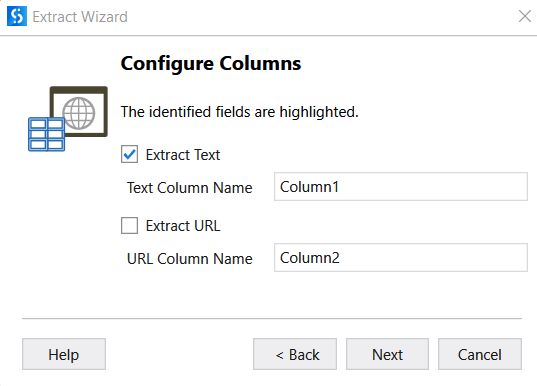
**Fig. 5. Select a Review comments**

* Then selected the second similar elements of the same page shown on Fig. 6.



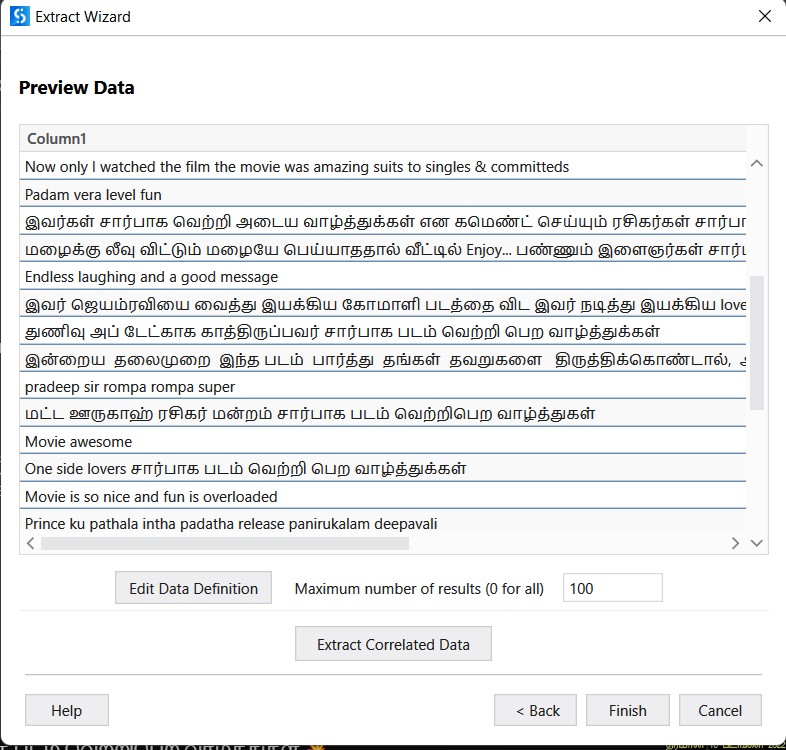
**Fig. 6. Similarity slection**

* Then the Element selector highlighter came, to select the element. Once the elements were selected, we can see the preview of data like in Fig. 8. If the data are extracted as expected then click on the finish button else, we can re-select the data as shown in Fig. 7. Connect node for flowchart to data scraping.

**Fig. 7 Connect the scrap with Flowchart**

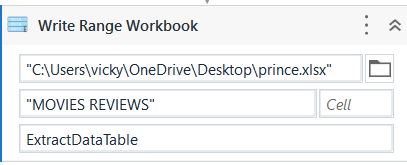
The following are the results of Data Scrapping



**Fig. 8 Preview of scraping reviews data**

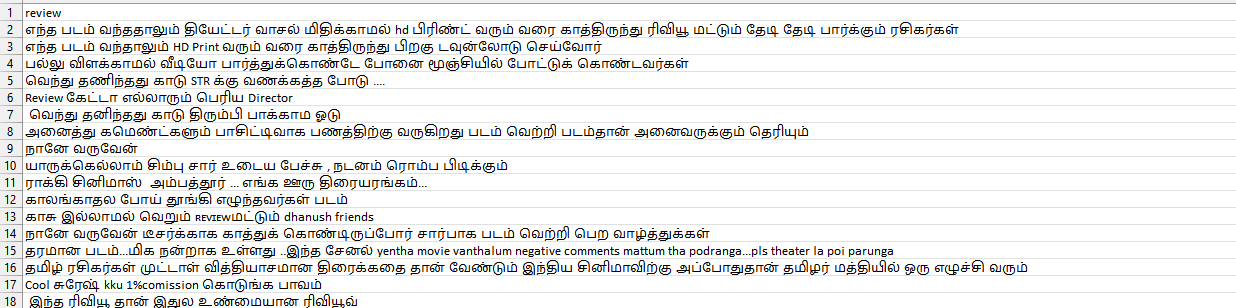
* All the obtained data of Reviews are stored in Excel using the Write Range Workbook

Fig. 9. The ExtractDataTable option helps to covert the data into excel sheet.



**Fig. 9. Extract the data to Table Format**

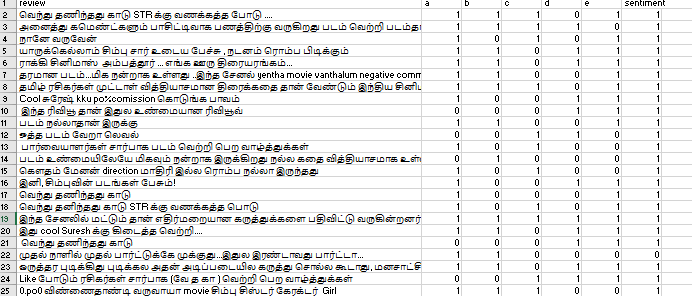
* Preview of the scraped data is on Fig. 10. Likewise same procedure is followed for scrapping the other Movies comments of Various Movies.

****

**Fig. 10. Preview of Extract data Reviews**

**Dataset**

* The comments were scraped from Annatha, Beast, Iravin nizhal, yanai, lovetoday, nenjukku needhi, master, valimai, Vikram, Vendhu thanindhadhu kaadu movie’s youtube reviews.
* After scrapping all the movie Comments, they were merged into the same excel file.
* The dataset consists of Movie Comments from the YouTube. The Comments are in both English and Tamil Languages. Using the language detector, we have detected the language of the comments.
* The dataset has been filtered to have only the Tamil comments.



**Fig. 11. Manual Sentiment Score**

* Then the comments were manually scored positive as 1 and negative as 0 based on the basic knowledge in Tamil language as shown on Fig. 12. The scores were collected from 5 people and its mode has been labelled as the actual positive and negative. Those Scoring shown on Fig. 11.

****

**Fig. 12. Manual Sentiment Scoring of Reviews**

**3.2 Lexicon Based Method**

The lexicon-based approach is extracting the sentiment from text. Lexicon-based approach uses pre-prepared sentiment lexicon to score a document by aggregating the sentiment scores of all the words in the document. This technique calculates the sentiment orientations of the whole document or set of sentences from semantic orientation of lexicons. Semantic orientation can be positive, negative, or neutral. The dictionary of lexicons can be created manually as well as automatically generated.

* Took a positive and negative comments based on the sentiment scoring. If 1 means positive comment and 0 be a negative comment.
* To create a corpus for both positive and negative words further process to implement the sentiment scoring for Movie comments.

**3.2.1 TEXT PREPROCESSING**

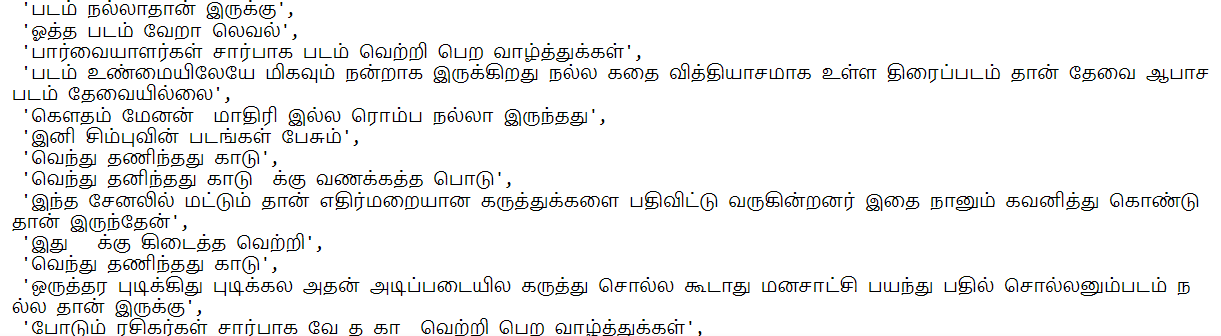
Text pre-processing is a crucial step in Natural Language Processing. Cleaning the text data in order to convert it into a presentable form that is analysable and predictable for task is known as text pre-processing. Python libraries used to handle many text pre-processing tasks.

The steps followed in text pre-processing are,

* Stripping whitespace, removing special characters and emojis.
* Removing punctuations.
* Tokenization.
* Removing the stop words.

**Stripping the empty spaces and Special Characters**

* Stripping here is done using Regex module. Regex is used to remove unwanted space and unwanted obstacles for words.
* A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern.
* RegEx can be used to check if a string contains the specified search pattern.
* Removing spaces re.sub("@[A-Za-z0-9\_] +","“, text)
* Removing special characters("#[A-Za-z0-9\_] +","“, text). Cleaned comments are shown on Fig. 13.



**Fig. 13. Remove Special characters & emoji’s**

**Removing Punctuations**

* Python string package consists of! **" # $ % & ' ( ) \* + , - . / : ; ? @ [ \ ] ^ \_ ` { | } ~** inside the punctuation module.
* All these characters in string. punctuations are removed from the comments which makes the comments free from noise. The punctuation removed text is shown in Fig. 14.



**Fig. 14. Punctuation removal**

**Tokenizing**

* Tokenisation is a technique for breaking down a piece of text into small units. A token may be a word or part of a word.
* Tokenizing the word do not contribute significantly to the information content of a sentence.
* Here tokenize the comments for breaking each sentence on the comments for better understanding.

**Removing Stopwords**

* **Stopwords are** the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc) and does not add much information to the text.
* By removing the stopwords low level information are removed and hence we can focus on the important information.
* Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training.
* Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”. Similarly in Tamil we have ஆகும், அல்லது, அவர், நான், உள்ளம், etc.
* We have used AdverTool for Tamil language to remove the stopwords occurred in the comments. This package had stopwords for some Indian languages. The stopword less tokens are shown on Fig. 15.



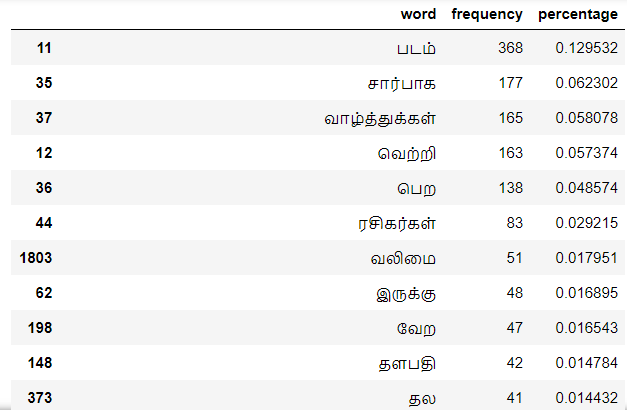
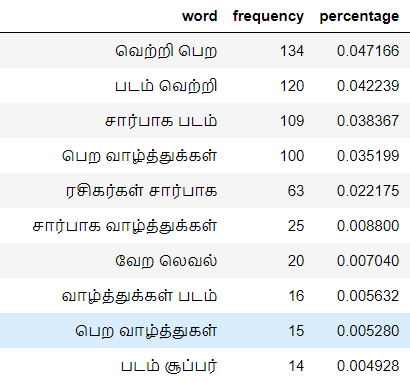
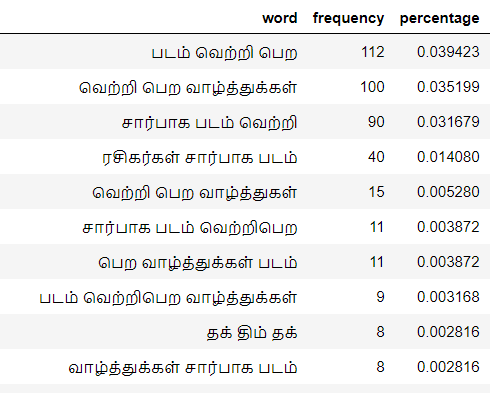
**Fig. 15. StopWords Removing for Reviews**

**3.2.2 Corpus Creation**

A corpus is the collection of an authentic text organized into datasets. Authentic here means text written by native language. A corpus can be made up of everything from newspapers, novels, recipes, radio broadcasts to television shows, movies, and tweets. In natural language processing, a corpus contains text and speech data that can be used to train AI and machine learning systems. Here we have created a corpus for identifying positive and negative reviews for the movie’s comments. we scored positive and negative word based on the sentiment score.

**A) Corpus Creation for Positive words**

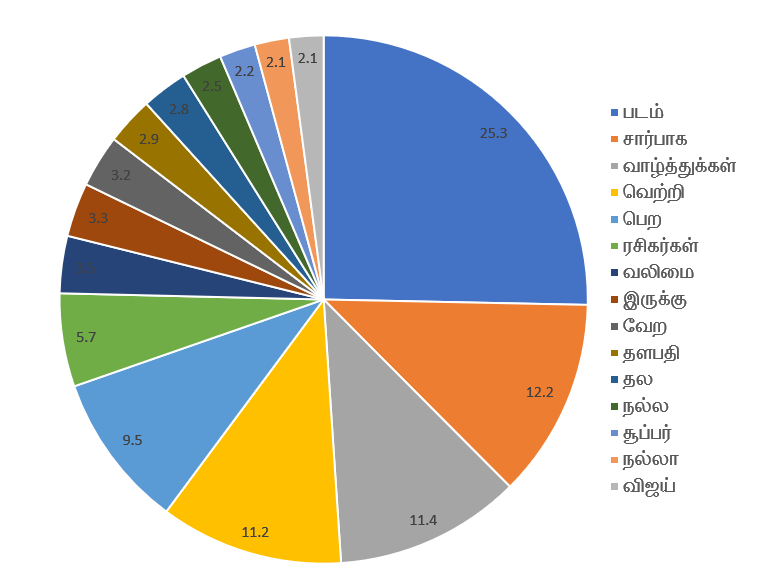
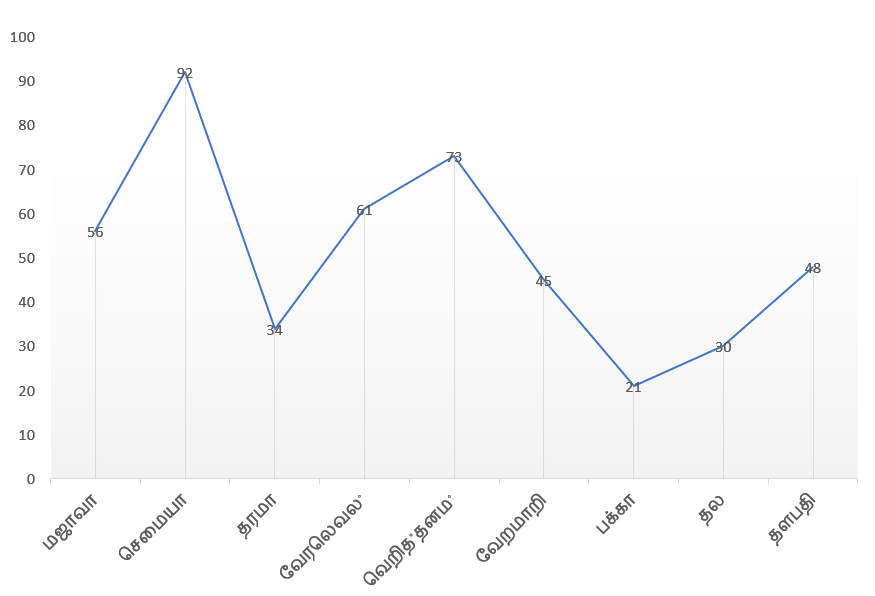
* Form the sentiment scores split the positive comments and negative comments, after doing all the pre-processing on 3.2.1 for sentiment score 1 [positive words].
* Using Counter, we have obtained the frequency of the words.
* We took the words which are most occurred on the positive comments and created a list of words for positive words as corpus.

**  **

**Fig. A) Unigram Positive Fig. B) Bigram Positive Fig. C) Trigram Positive**

**Fig. 16 Positive Corpus n-grams**

* The unigram, bigram and trigram positive words are extracted based on their combination and shown in Fig. 16. From the n-grams most occurred positive words on the text are taken by above 30 % based on the word frequency counter are shown on Fig. 17.
* Fig. 17. A) shows a most repeated Positive words occurs from the movie comments.
* Fig. 17. B) shows the new trending words and their frequency.

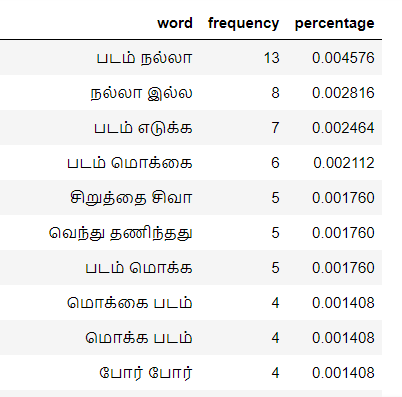
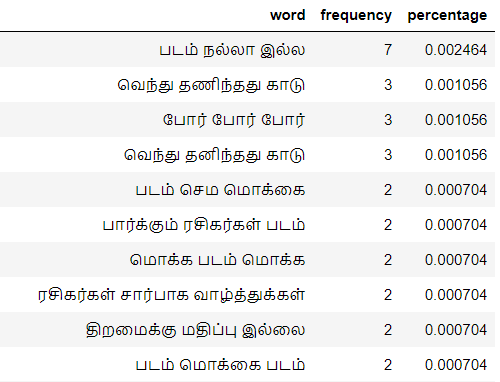
 ****

**Fig. A) Repeated Words**  **Fig. B) Trending Words**

**Fig. 17 Positive Corpus**

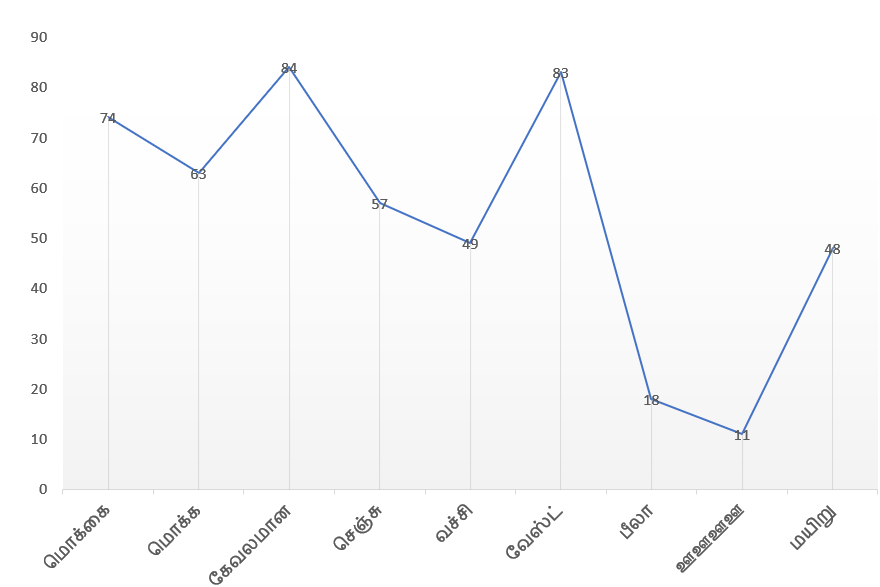
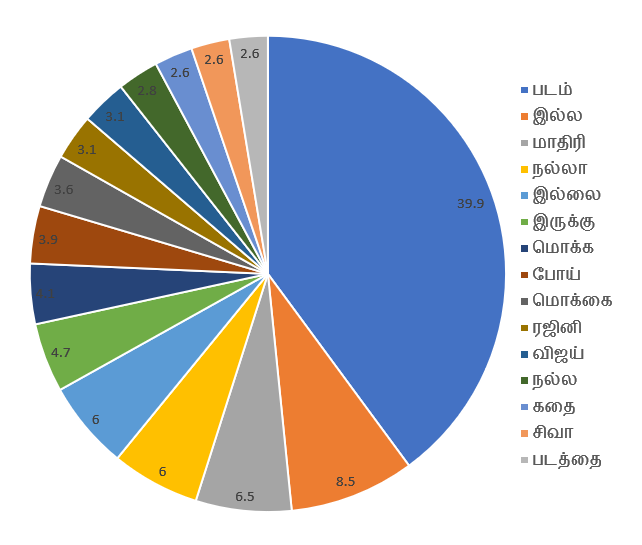
**B) Corpus Creation for Negative words**

* Form the sentiment scores split the positive comments and negative comments, after doing all the pre-processing on 3.2.1 for sentiment score 0 [positive words].
* Using Counter, we have obtained the frequency of the words.
* We took the words which are most occurred on the negative comments and created a list of words for negative words as corpus.

 **** 

**Fig. A) Unigram Negative Fig. B) Bigram Negative Fig. C) Trigram Negative**

**Fig. 18 Negative Corpus n-grams**

* The unigram, bigram and trigram negative words are extracted based on their combination and shown in Fig. 18. From the n-grams most occurred negative words on the text are taken by above 30 % based on the word frequency counter are shown on Fig. 19.
* Fig. 19. A) shows a most repeated negative words occurs from the movie comments.
* Fig. 19. B) shows the new trending negative words and their frequency.

**Fig. A) Repeated Words**  **Fig. B) Trending Words**

**Fig. 19. Negative Corpus**

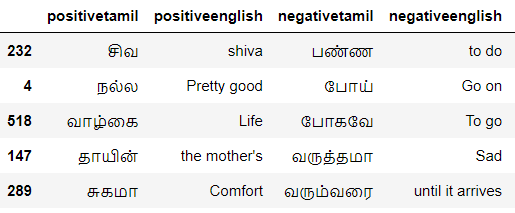
**4.3. Sentiment Scoring**

The sentiments of the reviews are categorised into positive and negative by using the libraries.

* Word Count
* Sentiment Intensity Analyzer [Vader Sentiment]
* TextBlob
* Afinn
* Manual scores

After creating a positive and negative corpus,we try to match the words for English by translating the Tamil word corpus into English. Using translator, we converted the Tamil text into English text.

* There are no packages available for doing sentiment analysis Tamil language. we can’t allocate sentiment score directly to Tamil text.
* We need to Translate a Tamil text into English text Using Translator and the words which can’t be directly translated were translated based on their synonym in Tamil Text.

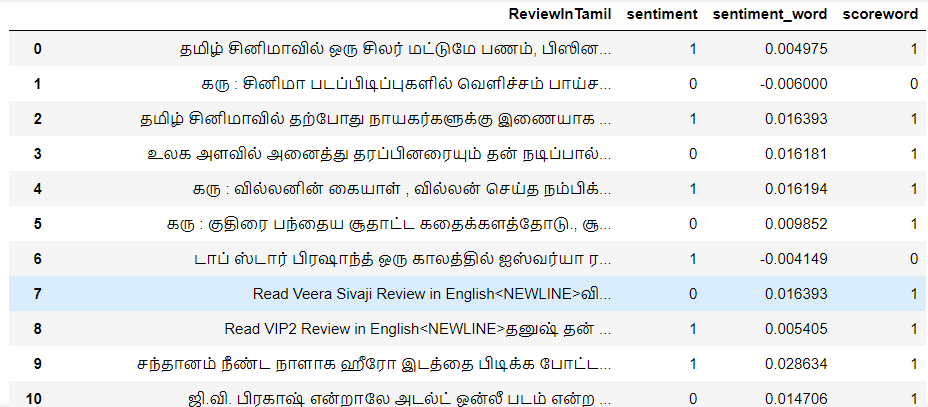


**Fig. 20. Tamil Words to English Words Translation**

Above Fig. 20, shows the Translated text corpus. further we scored each word by using different libraries and techniques.

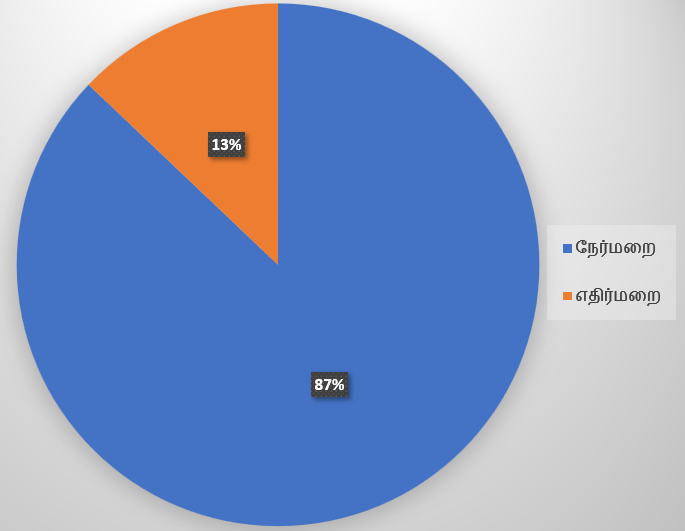
**Word Count**

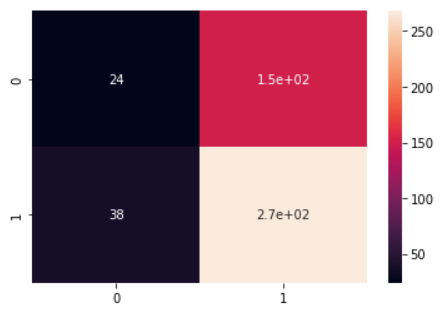
* Wordcount refers to most repeated words or text occurs in the sentence or paragraph. The words with more occurrences were used to identify the sentiment. The initial sentiment score is set to 0.
* Then if a token in the comment occurs in the positive corpus, the sentiment score will be added with 1 else if the word occurs in negative corpus the sentiment score will be subtracted with 1.
* The overall sum of the sentiment score is used to find the sentiment of the comment. If the sentiment score is greater than 0 then it is positive else it is negative. Score-word is the sentiment calculated using word count approach. Fig. 21.

****

**Fig. 21. Score by Frequency of wordcount**

* By the manual sentiment and wordcount score to cross check the different between actual sentiment score and wordcount scores.
* From Fig.22.A) basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* There are 4 classes in confusion matrix. The true positive signifies how many positive classes sample predicted correctly, true negative signifies how many negative class samples the created model predicted correctly, false positive signifies how many negative class samples created model predicted incorrectly and vice versa goes for false negative.
* The accuracy of the word count is 60.83%. The accuracy check by library accuracy score. And show the predicted outcomes**.**

****

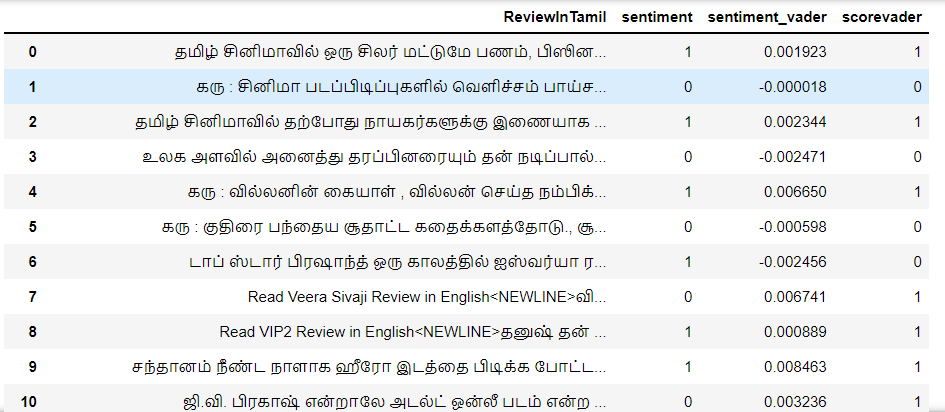
****

**Fig. A). Confusion Matrix Fig. B). Percentage for WordCount**

**Fig. 22. Classification Metrics for WordCount**

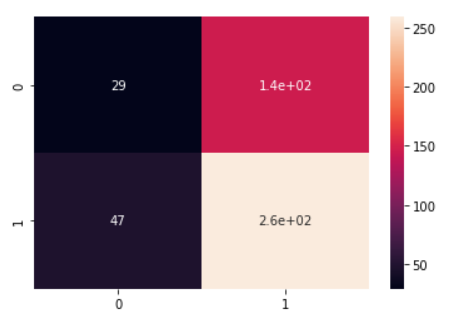
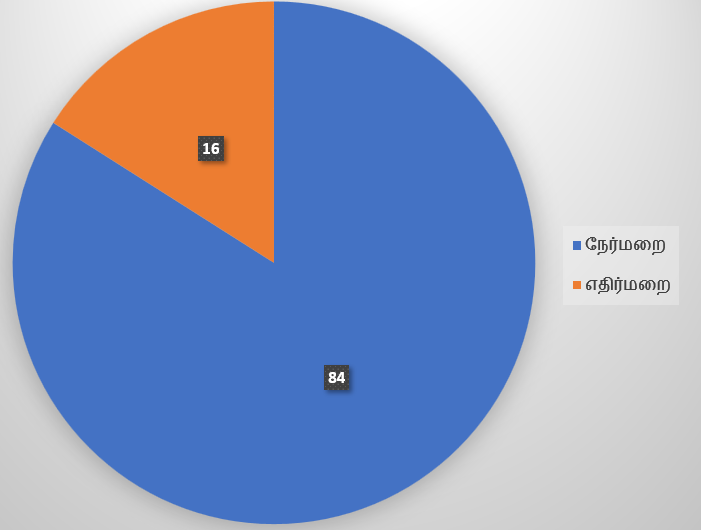
**Sentiment Intensity Analyzer**

* **VADER (Valence Aware Dictionary and Sentiment Reasoner)** is a lexicon and rule-based sentiment analysis tool that is specifically tuned to sentiments expressed in social media.
* **VADER** uses a combination of A sentiment lexicon is a list of lexical features which are generally labelled according to their semantic orientation as either positive or negative. **VADER** not only talks about the Positivity and Negativity score but also tells us about how positive or negative a sentiment.
* By the Vader sentiment scoring of text by using library. The score of Vader be in 1 to -1.
* Coverts the score by 0 and 1. If 1 be positive and 0 be negative. Append to the sentiment score and Vader positive and negative score shown on the Fig. 23.

****

**Fig. 23. Score By Vader Sentiment**

* By the manual sentiment and Vader sentiment score to cross check the different between actual sentiment score and Vader sentiment scores.
* From Fig. 24.A) basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* The accuracy of the Vader score is 61.23%. The accuracy check by a library accuracy score. And show the predicted outcomes.

****

**Fig. A). Confusion Matrix Fig. B). Percentage For Vader Sentiment**

**Fig. 24. Classification Metrics for Vader Sentiment**

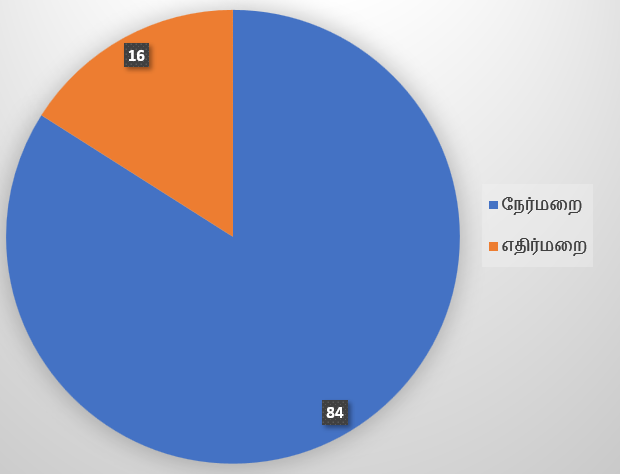
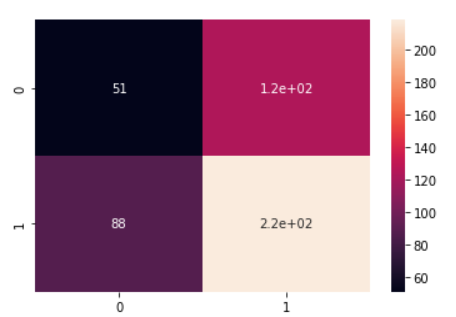
**TextBlob**

* TextBlob is a Natural Language Toolkit.
* Single sentence is given as input to TextBlob and the final output contains polarity and subjectivity.
* Polarity score lies between (-1 to 1) where -1 identifies the most negative words such as ‘மொக்கை’, ‘இல்லாமல்’ and 1 identifies the most positive words like ‘வெற்றி’, ‘வாழ்த்துக்கள்’.
* Subjectivity score lies between (0 and 1), It shows the amount of personal opinion, If a sentence has high subjectivity i.e. close to 1, It resembles that the text contains more personal opinion than factual information.
* The polarity score greater than 0 are considered as positive sentiment, less than 0 are considered as negative sentiment by the Fig. 25.



**Fig. 25. Score By Text Blob**

* By the manual sentiment and TextBlob score to cross check the different between actual sentiment score and TextBlob sentiment scores.
* From Fig.26.A) basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* The accuracy of the TextBlob is 56.04%. The accuracy check by an library accuracy score. And show the predicted outcomes.

****

**Fig. A). Confusion Matrix Fig. B). Percentage For TextBlob Score**

**Fig. 26. Classification Metrics for TextBlob Score**

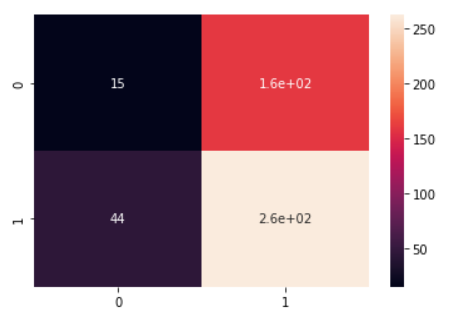
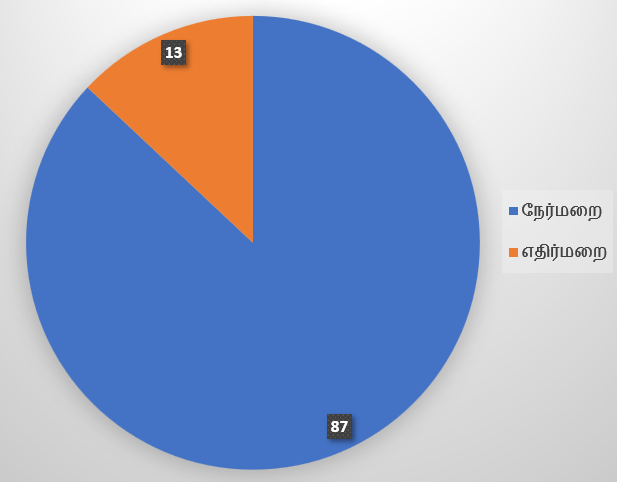
**Afinn**

* Afinn lexicon was created in 2009 for tweets downloaded for online sentiment analysis. The old version called AFINN-96 contains 1468 different words, including a few phrases. The updated version called AFINN-111 contains 2477 words.
* Score in AFINN is ranging from -5 to +5. -5 means very negative and +5 means very positive. Then if a token in the comment occurs in the Translated corpus, the sentiment score will be added or subtracted with Afinn sentiment score.
* Change the polarity of score between 0 and 1. 1 consider as positive 0 consider as negative.



**Fig. 27. Score By Afinn Sentiment**

* By the manual sentiment and Afinn score to cross check the different between actual sentiment score and Afinn sentiment scores.
* From Fig.28.A) basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* The accuracy of the Afinn is 57.74%. The accuracy check by a library accuracy score. And show the predicted outcomes.

**Fig. A). Confusion Matrix Fig. B). Percentage For Afinn Score**

**Fig. 28. Classification Metrics for Afinn Score**

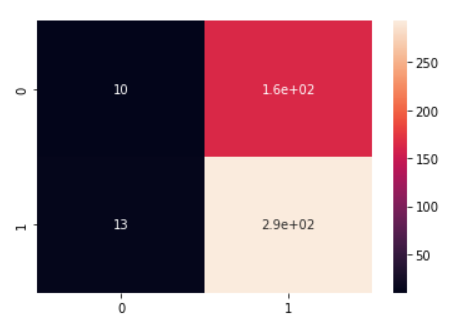
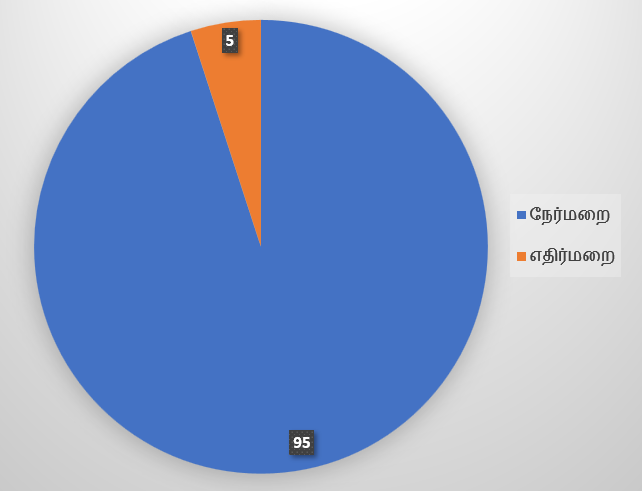
**Manual Score**

* The corpus has been scored manually in the range -5 to 5 based on their frequency of occurrences and basic knowledge in Tamil language.
* The initialsentiment score is set to 0.
* Then if a token in the comment occurs in the Translated corpus, the sentiment score will be added or subtracted with manual sentiment score.
* The overall sum of the sentiment score is used to find the sentiment of the comment.
* If the sentiment score is greater than 0 then it is positive else it is negative. Score-tb is the sentiment calculated using word count approach. Fig.29.



**Fig. 29. Score By Manual**

* By the manual sentiment and manually scored to cross check the different between actual sentiment score and Manual sentiment scores.
* From Fig.30.A) basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* The accuracy of the Manual score is 63. The accuracy check by a library accuracy score. And show the predicted outcomes.

**Fig. A). Confusion Matrix Fig. B). Percentage For Manual Score**

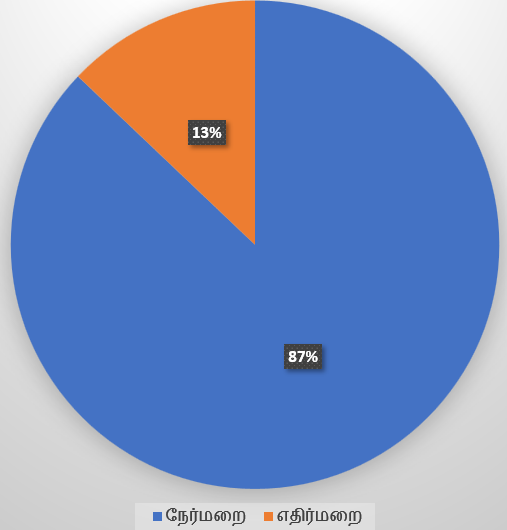
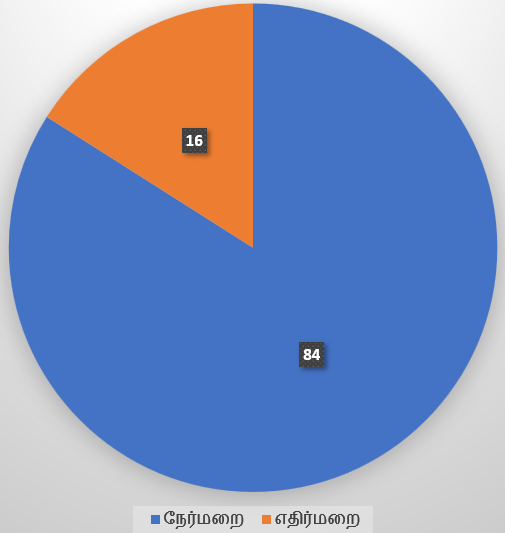
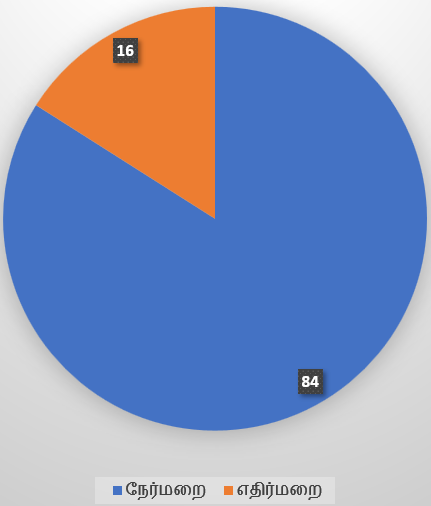
**Fig. 30. Classification Metrics for Manual Score**

**Sentiment Results**

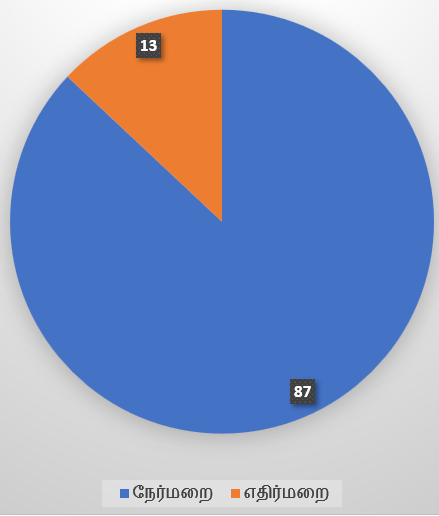
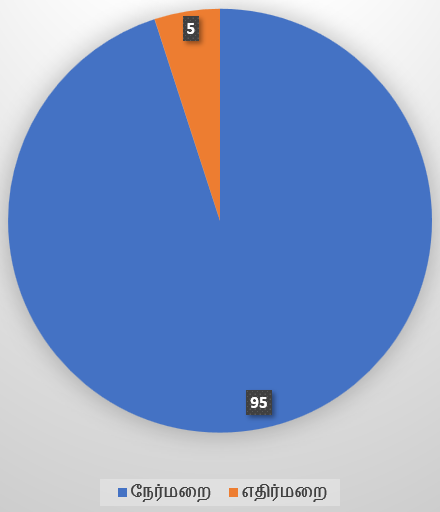
A sample dataset with 480 instances has been downloaded and tested using all the above five approaches. Their results have been represented in table 1 and Fig. 31.

**Table 1: Positive and Negative Reviews Lexicon-Based**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| WordCount  Sentiments | | Vader  Sentiments | | Afinn  Sentiments | | TextBlob  Sentiments | | Manual  Sentiments | |
| Positive | 418 | Positive | 404 | Positive | 421 | Positive | 341 | Positive | 457 |
| Negative | 62 | Negative | 76 | Negative | 59 | Negative | 139 | Negative | 23 |

**  **

**a) Wordcout Percentage b) Vader sentiment Percentage c) Afinn sentiment Percentage**

** **

**d) Text Blob sentiment Percentage d) Manual Scoring Percentage**

**Fig. 31 Percentage of different methods**

**ALGORITHM BASED**

* Machines cannot understand the characters and words. So, when dealing with text data we need to represent it in numbers to be understood by the machine.
* The Tamil movies Reviews which are collected from YouTube Comments are structured and manually tagged. Entirely, 1983 reviews about different Tamil movies are collected and are manually tagged. These reviews contain 1103 unique positive reviews and 880 negative reviews respectively.
* We need to classify positive and negative reviews by using different Machine learning algorithms.
* Classification is the process of predicting the class (also called, targets or labels or categories) of given data points. The task of classifiers is, approximating a mapping function (f) from input variables (X) to discrete output variables (y).

**Model validation**

* The train-test split procedure is appropriate when we have a costly model to train, or require a good estimate of model performance quickly.
* The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.
* We used 80% data for training and the balance 20% data for Testing.
* Training data has 1586 samples and Testing data has 397 samples.

**Vectorization**

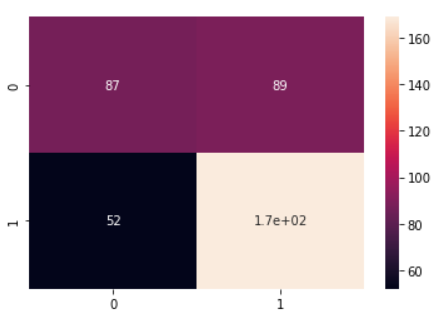
* Count Vectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.
* This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for using in further text analysis).
* Count Vectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix.
* The text document has been vectorized based on n-gram analysis. The features are made of word n-grams.

**Proposed Models**

* Sentiment classification has been done using different classification algorithms and their performance were recorded.
* In this approach we have used Naïve Bayes classifier, Logistic Regression, Random Forest Classifier and support vector machine.
* Their accuracy varied based on their classification approach.

**Naive Bayes**

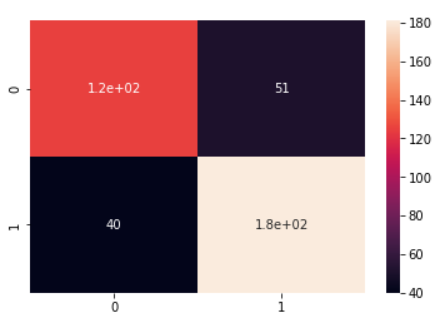
* One of the best methods for classifying documents is Naive Bayes. It has been employed frequently in the field of information recovery and most recently in studies involving machine learning.
* The multinomial template represents the integer feature to represent document whereas in Bernoulli model vector of binary feature are attained from the document.
* From Fig. 32. basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* The accuracy of the Multinominal Naïve bayes is 64%.

****

**Fig. 32. Confusion Matrix for Naïve bayes**

**Logistic Regression**

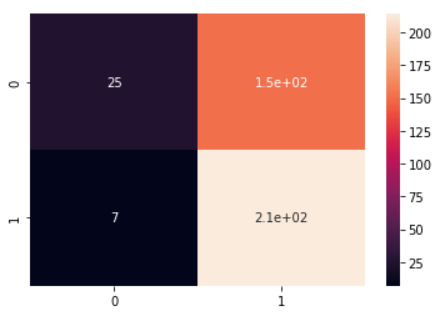
* Logistic regression itself is not a classifier but can be used for binary text classification. It simply models the probability of output in terms of input. But it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class and below the cutoff as the other for binary classification. The estimation is done through maximum likelihood estimation approach.
* From Fig. 33. basic classification metrics is the confusion matrix. Each row represents the instances in a predicted class and each column represent instance of actual class.
* The accuracy of the Logistic Regression is 77%.



**Fig. 33. Confusion Matrix for Logistic Regressio**

**Support Vector Machine**

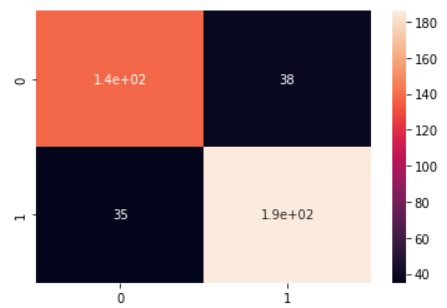
* Support Vector Machine is a classification technique used mostly in function approximation and data classification. A hyper plane created by a support vector machine divides two distinct classes.
* The SVM strives to achieve the highest separation possible between two classes while separating them. In cases where the train data are already projected, SVM prediction projects the test data.
* Based on the weight of the testing data, the SVM classifier divides the data into positive and negative categories. Create a linear SVM in this case to categories reviews as either favorable or negative. The SVM's accuracy rate is 60%.



**Fig. 34. Confusion Matrix for Support Vector Machine**

**Random Forest**

* The High Dimensional Noisy Data in Text Classification can be handled using Random Forest (RF) classifiers. Each decision tree in an RF model is trained using a different random subset of features.
* The prediction made by the RF is determined by a majority vote of all the forest's trees.
* From Fig. 35. basic classification metrics is the confusion matrix. Each tuples represents the instances in a predicted class and each attributes represent instance of actual class.
* The accuracy of the word count is 81%. The accuracy check by library accuracy score. And show the predicted outcomes

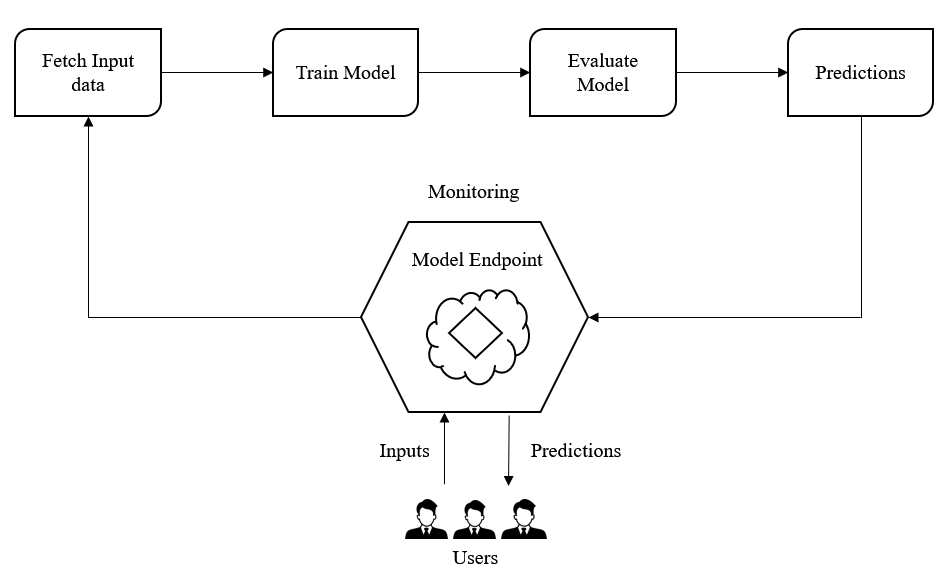
****

**Fig. 35. Confusion Matrix for Random Forest**

**3.5 Model Deployment**

* Machine learning Models deployment is the process of deploying a model in a live environment. The model can be deployed across a range of different environments and will often be integrated with the apps through an API.
* Models will usually deployment in offline or local environment, so will need to be deployed to be used with live data. The deployment techniques of model is different methods.
* Machine learning deployment must be proper manner so its efficient.
* We use a Stream Lit model, It is the easy way to build data api’s and share them. This tool makes simpler the entire model deployment Process and lets you deploy your model quickly.

**Model training & Deployment Pipeline**

****

**Fig. 36. Deployment Pipeline**

* Model Deployment generally contains 2 parts, front end and back end. Here no front-end experience has required.
* Using Stream Lit to create this frontend API in python shown on the Fig. 36. So, not need to any other programming language learning or web development techniques. Understand a python is enough.
* We have defined five types of scoring function. This function will take the data provided by users as input comments and make the predicted scoring using those function. Movie review is the input data for the model. It will return whether the comments Positive or Negative based on output of the functions.
* At the end of the app, fill the input comments, user have to click a enter button. Once enter is clicked, the function will be called and the result of the sentiment status of Movie Reviews will be displayed in the API Shown on the Fig. 37.
* The complete web app running on your local machine not a actual deployment. Deploy your model that’s create a link like local URL - <http://localhost:8501/>.
* The link will take you into the web app.

****

**Fig. 37. Model Deployment**

**RESULT AND DISCUSSION**

**CHAPTER - IV**

**RESULT AND DISCUSSION**

**Results**

The sentiment of the single comment has been predicted based on the five approaches. Those were represented in table 2

**Table 2: Sentiment of General Comments based on Lexicon-Based Approaches**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Movie’s**  **Comments** | **WordCount**  **Result** | **Vader**  **Sentiment**  **Result** | **Afinn**  **Sentiment**  **Result** | **TextBlob**  **Sentiment**  **Result** | **Manual**  **Sentiment**  **Result** |
| படம் நல்லாதான் இருக்கு | Positive | Positive | Positive | Positive | Positive |
| படம் மொக்க எதிர்பார்த்த மாதிரி இல்ல | Negative | Negative | Negative | Positive | Negative |
| படம் நல்லா இருக்கு ஆனா தியேட்டர் ல இருக்குறவங்களுக்கு புடிக்கல | Negative | Negative | Negative | Negative | Negative |
| படத்துக்குள் நிறைய கருத்துக்கள் இருக்கு..... பார்க்கும் கோணம் தவறு... | Negative | Positive | Positive | Positive | Positive |

From the table 2 we can say that the proposed manual is having more accurate results. This is similar to the accuracy scores on table 1.

Also, the algorithm’s performances were evaluated based on the classification evaluation metrics such as precision, recall, f1score and Accuracy score.

**Table 3: Classification Evolution Metrics for Algorithm-Based Approach**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scores | Multinominal  NB | Logistic  Regression | Random Forest  Classifier | SVM |
| Precision | 65 | 78 | 82 | 58 |
| Recall | 76 | 81 | 85 | 96 |
| F1 score | 70 | 79 | 84 | 73 |
| Accuracy | 64 | 77 | 82 | 60 |

Based on the accuracy we can say that Random Forest Model understands the vectorized reviews and predicts more accurately.

**Discussions**

**Movie Name: Sarpatta Parambarai**

A Sports Drama film, set during the 1970s, the film revolves around a clash between two clans namely Sarpatta Parambarai and Idiyappa Parambarai in North Chennai, young man living in Madras of the 1970s is presented with the chance to revive his boxing career and his clan. Its displays the boxing culture in the zone and also the politics being involved in it.

**Data Scrape**

The YouTube comments of sarpatta parambarai movie reviews were scraped using Uipath and the comments which are of languages other than Tamil language were removed. They were shown in Fig. 38.

****

**Fig. 38. Sarapatta Parambarai movie comments**

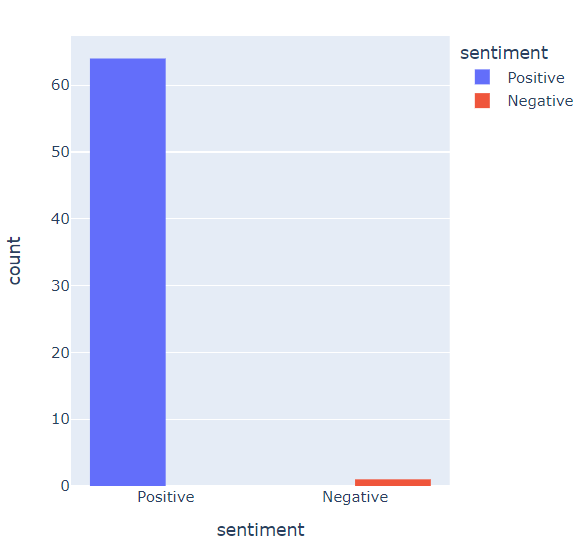
**Algorithm Based Approach**

In algorithm-based approach the model has been constructed based on the previous input data then the model is used to predict the sentiment of the comments of the reviews. The sentiment which are observed according to various algorithms on movie reviews were used to predict the occurrence of positive and negative comments. Based on their frequency the success of the movie can be predicted shows on Fig.39.

**Fig. 39. predicted output**

**Support Vector Machine**

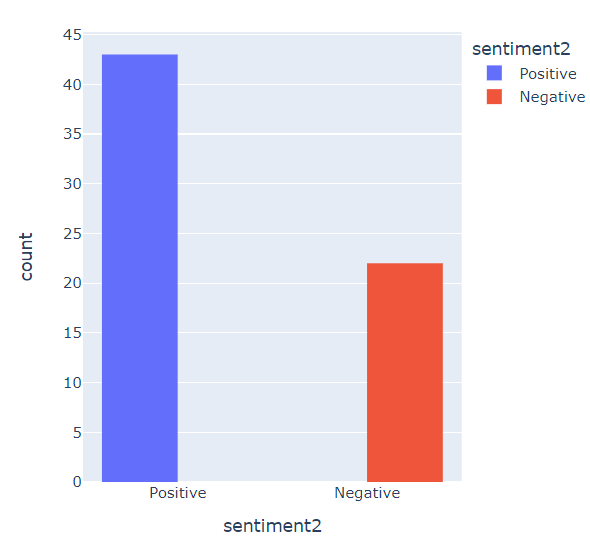
The data has been pre-processed and given as input for Support Vector Machine model and the sentiment of the comments were observed. Predicted Output shown on Fig. 40.

****

**Fig. 40. count of positive and negative comments Based on Support Vector Machine**

**Logistic Regression**

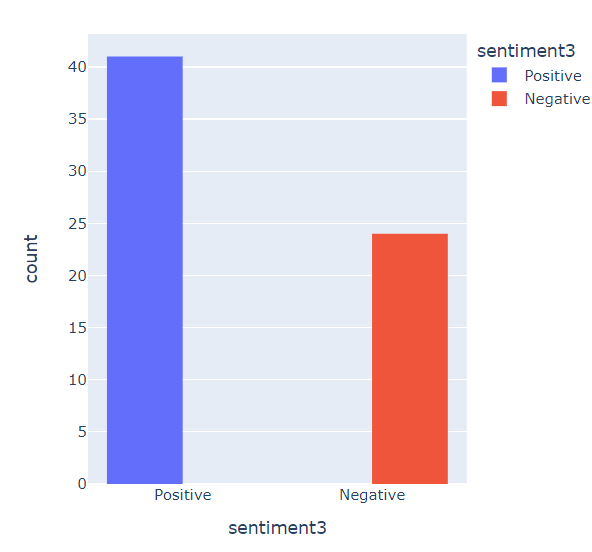
The data has been pre-processed and given as input for Logistic Regression model and the sentiment of the comments were observed. Predicted Output shown on Fig. 41.

****

**Fig. 41. count of positive and negative comments Based on Logistic Regression**

**Random Forest**

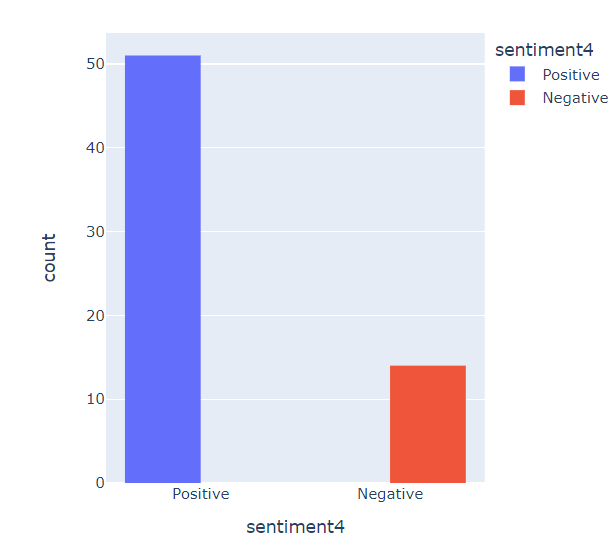
The data has been pre-processed and given as input for Random Forest model and the sentiment of the comments were observed.Predicted Output shown on Fig. 42.

****

**Fig. 42 count of positive and negative comments Based on Random Forest**

**Multinomial Naïve Bayes**

The data has been pre-processed and given as input for Multinomial Naive Bayes model and the sentiment of the comments were observed. Predicted Output shown on Fig. 43.

****

**Fig. 43. count of positive and negative comments Based on Multinomial NB**

The above four models clearly shows that the movie is have positive review among the audience. Because it has been observed to have nearly 70% of all the reviews were positive by all the algorithms.

**Lexicon – Based Approach**

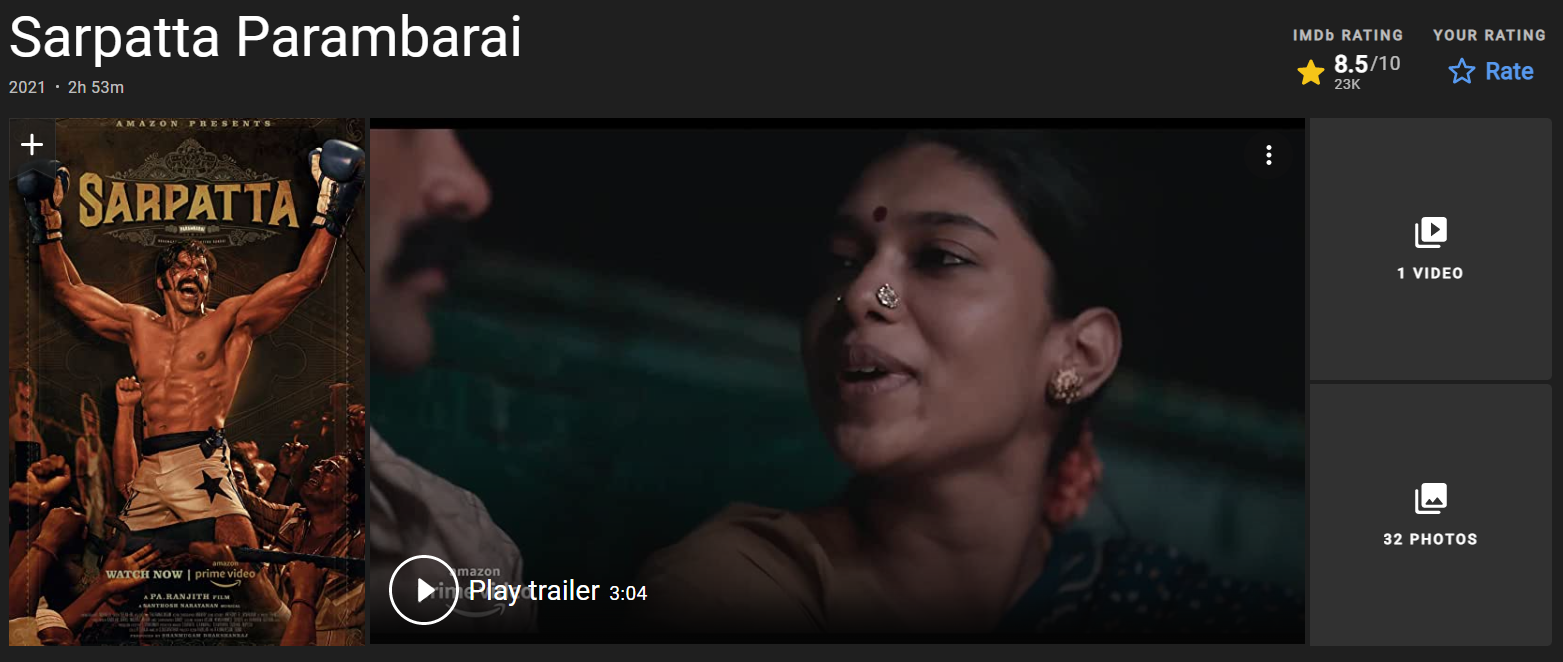
* The same filtered Tamil comments of sarpatta parambarai reviews were given as input in lexicon-based approach.
* Similarly, their outputs were observed and analyzed based on their sentiment which is shown in table 1.
* Finally, the most occurred sentiment has been used to predict the overall sentiment of that movie’s reviews.

**Table 4: Count of Positive & Negative Reviews of Sarpatta Parambarai Movie**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sentiment | Word Count  Result | Vader  Sentiment  Result | Afinn  Sentiment  Result | TextBlob  Sentiment  Result | Manual  Sentiment  Result |
| Positive | 53 | 46 | 48 | 45 | 57 |
| Negative | 12 | 19 | 17 | 20 | 8 |

* This table 3 clearly shows that the movie sarpatta parambarai is having positive reviews.

**Comparison**



**Fig. 44. IMDB Rating**

* From the above, we observed that the movie Sarpatta Parambarai is having more Positive Revies in both Lexicon-based and Algorithm-Based Approaches.
* The results were further compared with the IMDB rating which is 8.5/10 on Fig. 44. So, this shows that the prediction is correct.

**CONCLUSION**

**CHAPTER - V**

**CONCLUSION**

Five heuristic approaches have been applied to score the sentiment values. i.e) word frequency, afinn, textblob, vader sentiment and manual scoring. Among those five approaches manual scoring performs well with 63.125% accuracy whereas textblob has poor performance with lowest score of 56.456%. We presented a proposal for classifying the Tamil movie reviews using supervised algorithms, namely SVM, Random Forest, Logistic regression and Multinomial Naive Bayes. Features are extracted and are given to the system for better classification and to improve the accuracy of the system. The accuracy of different classifiers with word feature are represented in the table. The Random Forest [RF] classifies better than other classifiers with more accuracy.

**FUTURE ENHANCEMENTS**

* For future work, fine grained tagging can be done where, these reviews can be fine grained into very positive, positive, very negative, negative and neutral.
* Unsupervised way of implementing data can also be done as forthcoming work.
* More than trigram approach can be used to improve the positive and negative word corpus.
* Tokenized words may further stem or lemmatized to get precise

corpus.

**REFERENCES**

1) Chakravarthy, Bharathi et al. “Overview of the HASOC-Dravidian Codemix Shared Task on Offensive Language Detection in Tamil and Malayalam“CEUR.org/vol-3159, , December 17-21, 2020.

2) Kumar, Anand M.,et al.“A Novel Data Driven Algorithm for Tamil Morphological Generator.” International Journal of Computer Applications, vol. 6, no. 12, Sept. 2010.

3) mathan, thangarasu, and Manavalan . “Stemmers for Tamil Language: Performance Analysis.” (PDF) Stemmers for Tamil Language: Performance Analysis | Thangarasu Mathan - Academia.edu, 7 July 2013.

4) Chakravarthi, Bharathi Raja, et al. “Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text.” ACL Anthology, 30 May 2020.

5) Taboada, Maite, et al. “Lexicon-Based Methods for Sentiment Analysis.” Computational Linguistics, vol. 37, no. 2, June 2011, pp. 267–307.

6) Liu, Bing. “Sentiment Analysis and Opinion Mining.” Synthesis Lectures on Human Language Technologies, 2012.

7) Regina and P. Sengottuvelan, "Analysis of Sentiments in Movie Reviews using Supervised Machine Learning Technique," 2021 4th International Conference on Computing and Communications Technologies (ICCCT), 2021, pp. 242-246.

8) Abhishek, Kumar, et al. “SENTIMENTAL ANALYSIS FOR MOVIE REVIEWS | Abhishek | International Journal of Advanced Research in Computer Science.” SENTIMENTAL ANALYSIS FOR MOVIE REVIEWS | Abhishek | International Journal of Advanced Research in Computer Science, 25 May 2020.

9) Lone, Musharaf Zia. “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.” (PDF) SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining | Musharaf Zia Lone - Academia.edu, May 2006.

10) Tripathi and S. K. Trivedi, "Sentiment analyis of Indian movie review with various feature selection techniques," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 181-185.

11) R. Naidu, S. K. Bharti, K. S. Babu and R. K. Mohapatra, "Sentiment analysis using Telugu SentiWordNet," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 666-670.

12) Novendri, Risky, et al. “Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes.” Bulletin of Computer Science and Electrical Engineering, vol. 1, no. 1, June 2020, pp. 26–32.

13) Thilagavathi, R., and Mrs. K. Krishnakumari. “Tamil English Language Sentiment Analysis System – IJERT.” Tamil English Language Sentiment Analysis System – IJERT, 24 Apr. 2018.

14) Abdelrahman, S., et al. “[PDF] Sentiment Analysis of English and Tamil Tweets Using Path Length Similarity Based Word Sense Disambiguation | Semantic Scholar.” [PDF] Sentiment Analysis of English and Tamil Tweets Using Path Length Similarity Based Word Sense Disambiguation | Semantic Scholar, 1 Jan. 2019.

15) Jagdale, Rajkumar & Shirsath, Vishal & Deshmukh, Sachin. (2019). Sentiment Analysis on Product Reviews Using Machine Learning Techniques: Proceeding of CISC 2017. 10.1007/978-981-13-0617-4\_61.

16) Ramanathan, Vallikannu, et al. “Predicting Tamil Movies Sentimental Reviews Using Tamil Tweets | Journal of Computer Science | Science Publications.” Predicting Tamil Movies Sentimental Reviews Using Tamil Tweets | Journal of Computer Science | Science Publications, 13 Nov. 2019.

17) Ramanathan, Vallikannu & Thirunavukkarasu, Meyyappan & Thamarai, S.. (2021). Sentiment Analysis: An Approach for Analysing Tamil Movie Reviews Using Tamil Tweets. 10.9734/bpi/ramrcs/v3/4845F.

18) Thavareesan, Sajeetha, and Sinnathamby Mahesan. “Review on Sentiment Analysis in Tamil Texts.” Journal of Science, vol. 9, no. 2, Dec. 2018, p. 1.

19) Se, Shriya & Ravi, Vinayakumar & Kumar, M. & Kp, Soman. (2016). Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i45/106482.