RESEARCH ARTICLE OPEN ACCESS

Review On Sentiment Analysis of Social Media Malayalam Text

Abin Joseph [1], Ameena A.K [2], Lamia Thasni O Nysam [3], Abeera V P [4]

[1],[2],[3] Student, Computer Science and Engineering, KMEA Engineering College - Aluva

[4] Asst.prof, Computer Science and Engineering, KMEA Engineering College - Aluva

ABSTRACT

Sentiment analysis or opinion mining is a Natural Language Processing to find the emotions of public opinion from user generated text. Sentiment Analysis in social media, acquiring large importance today because people use social media platforms to share their views and opinions on relevant topics in the form of movie reviews, product slanguage has a large importance. Malayalam is a low-resource language and it does not possess a standard corpus or a sentiment lexicon. This paper aims a review of different approaches to sentiment analysis of social media Malayalam text. The learning carried out at two levels and the system classify sentences into positive, negative classes. The work includes creation of a large size annotated corpus as a primary task and then followed by training a sentence level classifier to perform sentiment analysis.

Keywords: - Sentiment Analysis, Natural Language Processing.

I. INTRODUCTION

As internet is growing bigger, its horizons are becoming wider. Social Media platforms like YouTube, Facebook, Twitter etc dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analysing interesting patterns from the infinite social media data for business-driven applications.

Natural Language Processing is the ability by which a computer program identifies what a human being has said in the exact context spoken by him. It can be also considered as a field of Artificial Intelligence. One of the most emerging field of NLP is Sentiment Analysis. It is in short a cognitive process which can extract user's feelings and emotions. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. In overall, it is the computational study that analyses people's opinions, sentiment, attitude, and evaluation from the written languages. It has wide applications and include emotion mining, polarity, classification and influence analysis. It is also otherwise called as opinion mining. This emotions can be used to analyses people attitudes towards certain topics. Precisely, it is a paradigm of categorizing sentiments into positive, negative or neutral labels.

Everyday enormous amount of data is created from social networks, blogs and other media and diffused in to the world wide web. This huge data contains very crucial opinion related information that can be used to benefit businesses and other aspects of commercial and scientific industries. Manual tracking and extraction of this useful information is not possible, thus Sentiment analysis is required.

Natural language processing (NLP) is the technology dealing with our most ubiquitous product: human language, as it appears in emails, web pages, tweets, product descriptions, newspaper stories, social media, and scientific articles, in thousands of languages and varieties. In the past decade, successful natural language processing applications have become part of our everyday experience, from spelling and grammar correction in word processors to machine translation on the web, from email spam detection to automatic question answering, from detecting people's opinions about products or services to extracting appointments from your email. The greatest challenge of sentiment analysis is to design application-specific algorithms and techniques that can analyse the human language linguistics accurately.

II. RELATED WORKS

Bo Pang *et.al [1]* proposed the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, they find that standard machine learning techniques.Definitively outperform human-produced baselines. However, the three machine learning

methods they employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. They conclude by examining factors that make the sentiment classification problem more challenging. They showed that SVM out performs other two classifiers.

Turney et. al [2] proposed an unsupervised technique for sentiment classification. They used the semantic orientation and Point wise Mutual Information- Information Retrieval (PMI−IR) method for SA of 410 reviews collected from different domains. The algorithm has three steps:

1. Extract phrases containing adjectives or adverbs.
2. Estimate the semantic orientation of each phrase.
3. Classify the review based on the average semantic orientation of the phrases. The algorithm attains an average accuracy 74%. It appears that movie reviews are difficult to classify.

The limitations of this work include the time required for queries and, for some applications, the level of accuracy that was achieved. The former difficulty will be eliminated by progress in hardware. The latter difficulty might be addressed by using semantic orientation combined with other features in a supervised classification algorithm.

Pramod K.V et. al *[3]* SA of Malayalam tweets using NB, SVM, and RF are proposed in this work. Four different features like BOW, TF IDF, Unigram with Senti-wordnet, and Unigram with Senti-wordnet including negation words, are considered for feature vector formation of the input dataset. All the classifiers with the last two features have shown better accuracy compared with other features. RF classifier with Unigram with Senti wordnet including negation words, got the highest accuracy, 95.6%.

Wang *et.al [4]* proposed a capsule model-based recurrent neural networks for sentiment analysis. The key idea of RNN-Capsule model is to design a simple capsule structure and use each capsule to focus on one sentiment category. Each capsule outputs its active probability and the reconstruction representation. The objective of learning is to maximize the active probability of the capsule matching the ground truth and to minimize its reconstruction representation with the given instance representation. At the same time, the other capsules' active probability needs to be minimized, and the distance between their reconstruction representations with the instance representation needs to be maximized.They show that this simple capsule model achieves state of the art sentiment classification accuracy without any carefully designed instance representations or linguistic knowledge.

They also show that the capsule is able to output the words best reflecting the sentiment category. The words well reflect the domain specificity of the datasets, and many words carry sentiment tendencies within the context defined by the data. Such words are not included in any sentiment lexicon, but these words become extremely useful in real applications for decision makers to further understand the quality of their products and services.

Liu *et.al* [5] proposed attention-based sentiment reasoner for aspect-based sentiment analysis. They have applied attention mechanisms for assigning importance for different words in a sentence. The AS Reasoner model has experimented on four different datasets of Chinese and English. Aspect-based sentiment analysis (ABSA) is a powerful way of predicting the sentiment polarity of text in natural language processing. However, understanding human emotions and reasoning from text like a human continues to be a challenge. Liu *et.al* [5] they propose a model, named Attention-based Sentiment Reasoner (AS-Reasoner), to alleviate the problem of how to capture precise sentiment expressions in ABSA for reasoning. AS-Reasoner assigns importance degrees to different words in a sentence to capture key sentiment expressions towards a specific aspect, and transfers them into a sentiment sentence representation for reasoning in the next layer.

Nairet *et.al [6]* used both linear SVM and CRF approaches for the SA of Malayalam movie review. Hyperparameter tuning was not done in their work. They combined both machine learning method and rule based method. The classification is performed at Sentence level and SVM or CRF learning is used in machine learning part. New tag set is defined for film domain. Training data is created using this tag set. The learned model will label the words with newly defined tag set. The output of machine learning part is then given to rule based system which determine the overall sentiment polarity of the sentence based on number information of annotated tags. More specifically the machine learning part helps to determine the contextual polarity of individual words and rule based part determine sentence level sentiment polarity on the basis of frequency and relative position of sentiment words in the sentence. The efficiency of the system would be reduced if the text contain sarcasm, connectives, conjunctions etc. Because the rules are very crisp and viable to over fitting. The training corpus consists 30,000 tokens.

They concluded that SVM learning obtained better

accuracy compared to CRF in sentiment identification. The dataset is created by collecting texts from online Malayalam movie reviews.

Table 2.1 Summary of related work on the SA of Malayalam.

Summary of related work on the SA of Malayalam.

References	Dataset	Sentiment classification methods
Soumya et al. [7]	Malayalam Tweets	Deep neural network architectures like RNN, LSTM, Bi-LSTM, GRU and CNN models
Kasthoori et al. [10]	Malayalam Online Newspaper	Machine Learning Method ART classifier for domain identification and Fuzzy logic for polarity Classification
Rahul et al. [9]	Malayalam text from social media	CRF and SVM
Kumar et al. [8]	Malayalam Tweets	CNN and LSTM
Ashna et al. [11]	Malayalam Reviews	Lexicon based approach
Thulasi et al. [12]	Malayalam Movie Reviews	Aspect based analysis using Viterbi and HMM model
Nair et al. [6]	Malayalam Movie Reviews	SVM and CRF
Anagha et al. [13]	Malayalam film Reviews	Fuzzy logic
Jayan et al. [14]	Malayalam film Reviews	CRF combined with rule based approach
Nair et al. [15]	Malayalam Movie Reviews	Rule based approach
Mohandas et al. [16]	Malayalam Movie Reviews	SO-PMI-IR

SA has been done in different Indian languages like Bengali, Hindi, Punjabi, Manipuri, Kannada, Tamil and Malayalam. SA done in the Malayalam language is summarized in Table 2.1.

The paper by Deepu et.al [17] discuss Hybrid approach to sentiment analysis of Malayalam movie reviews. They combined both machine learning method and rule based method. The classification is performed at Sentence level and SVM or CRF learning is used in machine learning part. New tag set is defined for film domain. Training data is created using this tag set. The learned model will label the words with newly defined tag set. The output of machine learning part is then given to rule based system which determine the overall sentiment polarity of the sentence based on number information of annotated tags. More specifically the machine learning part helps to determine the contextual polarity of individual words and rule based part determine sentence level sentiment polarity on the basis of frequency and relative position of sentiment words in the sentence. The efficiency of the system would be reduced if the text contain sarcasm, connectives, conjunctions etc. Because the rules are very crisp and viable to over fitting. The training corpus consists 30,000 tokens. They concluded that SVM learning obtained better accuracy compared to CRF in sentiment identification. The dataset is created by collecting texts from online Malayalam movie reviews.

M. Neethu et.al [18] discussing a lexicon based approach to extract different moods or different levels of sentiment from Malayalam text. Here the classification of text performed into more refined classes. Sentences are tagged with appropriate moods like sad, angry, happy and neutral. A reference word set which act as sentiment lexicon is created manually. Which contains desirable and undesirable words. Using statistical methods like Point wise Mutual Information (PMI) and Latent Semantic Analysis (LSA) is used to measure the semantic orientation of words with previously stored desirable and undesirable word set. Which then used to determine the target classes of individual words in a sentence. Only adjectives and adverbs are considered because these POS categories are used to express the subjectivity. Which is then used to calculate the semantic orientation. The dataset is domain dependent which is collected from Malayalam novels.

Malayalam is a highly agglutinative language making the pre-processing step more challenging compared with other languages. A significant issue in SA of Malayalam is the unavailability of the tagged datasets. All the works mentioned in Table 2.1 have used their own manually created datasets.

III. THEORETICAL BACKGROUND

There are three broad classification methods in sentiment analysis: machine learning based approach,lexicon based approach and hybrid based approach. Machine learning based sentiment classifiers depends on annotated training data. This training data is usually derived from feature words in order to classify new data. The outcomes of the machine learning classifier are based on feature selection methods. Most of the existing studies focus on machine learning methods. Although there are several numbers of machine learning methods, support vector machine, Naive Bayes and maximum entropy are the standard classifiers used in most of the research studies. Fig 3.1 represents different sentiment analysis approach.

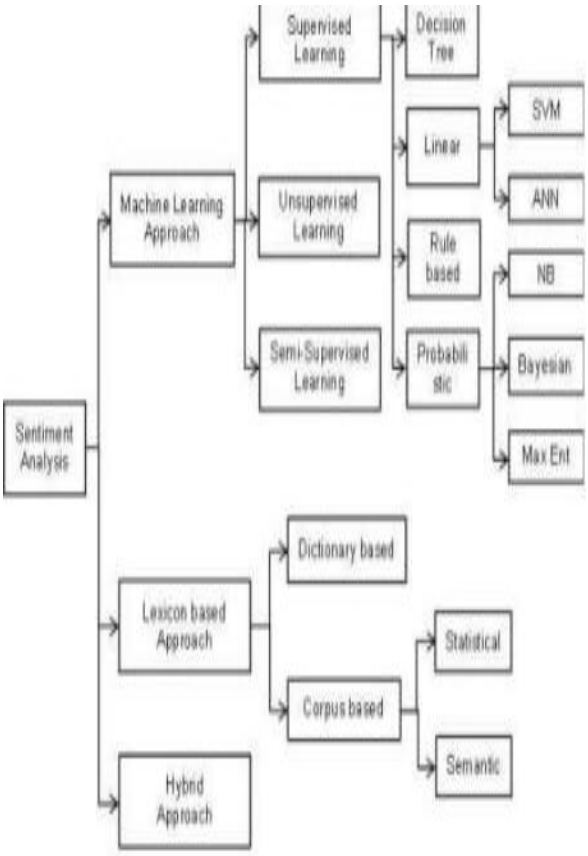

Fig 3.1:SA approaches

(1) Machine learning based sentiment analysis

A classification of related research works specifically to Naive Bayes (NB) classifier , Linear regression, K-Nearest Neighbors (kNN), Support Vector Machines (SVM),K-Means clustering,Random forest, Decision tree etc. The selection of feature words or feature vectors is an important part when using machine learning classifiers for sentiment analysis. Proper engineering of contextual features can provide higher informational value and reduce chances of noise. In order to achieve this, different sources for feature engineering are often used. Most of the reviews proposed the method for polarity classification using machine learning algorithms.

i) Supervised Learning: In supervised learning,two pre-annotated datasets are required,training set and test set. The training set is used to train our classifier while test set is used to evaluate the performance of the classifier.The first step is to collect the data for thetraining set and then classifier is trainedaccordingly with the help of the chosen techniques.The most commonly used techniques to train classifierincludeNaive Bayes classifier, Support VectorMachine, Maximum entropy model,etc. The main disadvantage of supervisedlearningmethod is that it requires asignificant amount of annotated dataset.

ii) Unsupervised Learning: The problems like human annotation requirement, domaindependency, and multi-language applicability can be solved with the help of unsupervised learning techniques.It uses different clustering algorithms like K-Mean clustering, to classify input data into classes. Semantic Orientation and Pointwise mutual information are also utilized for the unsupervised classification in sentiment analysis. In semantic orientationmethod,two arbitrary seed words(poor and excellent) are selectedin conjunctionwith vast text corpus. Then the semantic orientation of thephrases is calculated with their association with these seedwords. The average of thesemanticorientation of all such phrases can determinethe overall sentiment of the document. Supervised learninghas been foundto perform better than Unsupervised learning.

(2) Lexicon based sentiment analysis

Lexicon or SentiWordNet development is an important task in Lexical based approach, which is “structure that keeps information about words and their synonyms or related meaning” where each word is described as “lexical items”. Using this lexicon or WordNet, the total polarity of the user sentence or text is then calculated by a weighted count of all lexical items.

Lexicons are constructed using sentiment bearing or polar words. Further those polar words are split into two or possibly three categories (positive, negative and neutral) based on their polarity to build the lexicon. The lexical resources and knowledge in a particular domain is needed for lexicon-based sentiment analysis. The sentiment scores of a given text or reviews are calculated using the polarity of the words or phrases in the lexicon-based method. In most of the machine learning algorithms, unigrams or N-grams are used as features for training the classifiers. However, unigrams in the lexicons are used to assign polarity, thus the overall polarity of the text is then calculated as the total of unigram scores.

(3) Hybrid model based sentiment analysis

Hybrid approach is combined form of both machine learning based methods and lexicon based methods.Further, there is a method referred as linguistic rule-based approach which is generally combined with a lexicon based sentiment classification. A classification of The hybrid approach uses the syntactic features of the words, word phrases,

negations, and the construction of the text to identify the semantic orientation of the text or sentence. Parts-of-Speech (POS) tagging is one of the methods used in the linguistic-based approach to identify grammatical category of the words. Several POS patterns or tagset can be used as features to be assigned to the sentence. POS tags contain combinations of nouns, adjectives or verbs or any other part of speech. Those tags can then be used to specify either a specific polarity or a specific sentiment topic. The three different sentiment classification approaches can be used individually or combined with one another. For example, N-grams model and linguistic-based approaches can be combined in such a way that POS tags are used as features for training the machine learning sentiment classifier. Further machine learning algorithms can be combined with Sentiment analysis, which is in demand because of its efficiency.

People all around the world nowadays working on different topics of sentiment analysis. As the amount of data increasing day by day data mining and sentiment analysis became more popular to people. Different companies need sentiment analysis for collecting customer information. Because this is the easiest way to find their customers and target them for their business.

IV. METHODOLOGY

Overall steps involved in this sentiment analysis can be listed as dataset collection, preprocessing, feature extraction, model development and model assessment.Fig 4.1 shows the overall steps involved in the sentiment analysis.

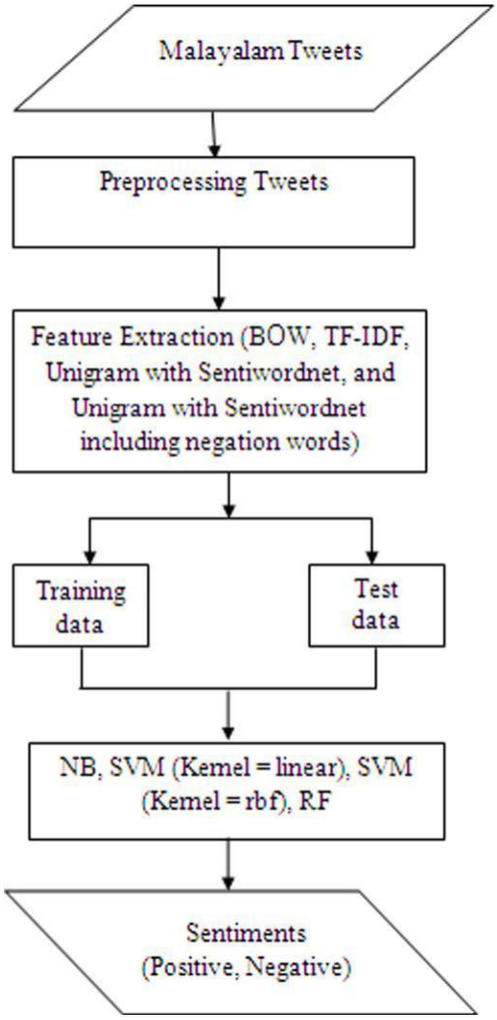

Fig 4.1:Steps of sentiment analysis

(1) Data set collection

Data collection is defined as the procedure of collecting, measuring and analyzing accurate datas for research using standard validated techniques.

The malayalam text dataset are retrieved from social media platform using web scraping technique. The various pre-processing phases are required to clean the dataset before proceeding to sentiment analysis.

(2) Preprocessing

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the

data mining process

a) Tokenization:

The tokenizing stage is the stage of cutting the input string based on each word that makes it up.

b) Normalisation:

Normalisation is the next step where uniformity in data is achieved by converting the text into standard form, correcting the spelling, deleted the repeated letters etc.

c) Removing stop words:

Excluding unnecessary words which include articles, prepositions which does not benefit in ascertaining the polarity of the text is the next step, stop word removal.Fig 4.2 represents the list of stopwords.

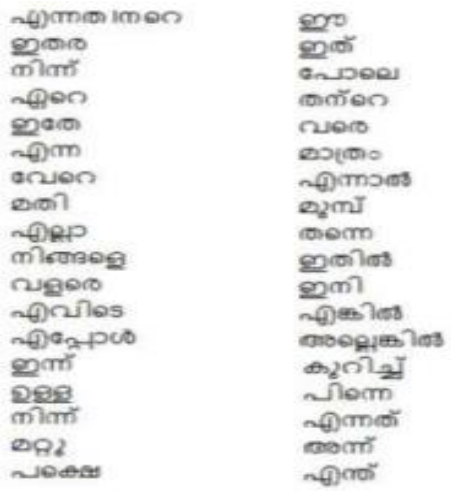

Fig 4.2 : List of stopwords.

d) POS Tagging:

POS tagging is the mode of labelling different parts of speech in a sentence. This step is used in identifying various facets from a sentence that are generally conveyed by nouns or noun phrases whereas sentiments are transpired by adjectives.

e) Stemming

The stemming technique is needed in addition to reducing the number of different indexes of a document, also to do a grouping of other words that have similar basic words and meanings but have different forms because they get different affixes. The stemming process in Indonesian texts is different from stemming in English texts. In English texts, the only process required is the process of removing suffixes. While in the Indonesian-language text all affixes both suffixes and prefixes are also omitted.Fig 5.3 represents some examples of stemming.

Fig 5.3 Stemming

f) Lemmatization

Lemmatization is the algorithmic process of finding the lemma of a word depending on their meaning. Lemmatization usually refers to the morphological analysis of words, which aims to remove inflectional endings.

(3) Feature Extraction

The machine recognize data in numbers only, so we want to represent each text data in numbers. After cleaning the data with pre-processing steps, convert or map the text or words to real valued vectors is called word vectorization or word embedding. Feature map or matrix is created by this feature extraction technique where in a document is broken down into sentences that are further broken into words, then map to numbers.

In the feature matrix, each row represents a sentence and column represents features as a word in the dictionary, and the values present in the cells

represent the count of the word in the sentence. Different methods are exist for feature extractions such as BOW, N gram, TF-IDF and Word Embedding.

a) Bag of words:

One of the most straight forward method is 'Bag of Words' (BOW), in which a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. A bag-of words is a representation of text that describes the occurrence of words within a document. It involves two things: a vocabulary of known words and a measure of the presence of known words. In which a fixed-length vector of the count is defined where each entry corresponds to a word in a pre-defineds dictionary of words. A count of 0 is assigned if the word in a sentence is not present in the pre-defined dictionary, otherwise a count of frequency of the word it appears in the sentence. That is why the length of the vector is always equal to the words present in the dictionary.

b) N-gram:

N-gram is a contiguous sequence of n items from a given sample of text or speech. In an n-gram vector representation, the text is represented as a collaboration of unique n-gram means groups of n adjacent terms or words. The value of n can be any natural number. In n-grams, word order is impotant, whereas in BOW it is not important to maintain word order. During the NLP application, n-gram is used to consider words in their real order so we can get an idea about the context of the particular word.

c) TF-IDF:

TFIDF, Term frequency-inverse document frequency, is another method used for feature extraction. TFIDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

In this method text represents in matrix form, where each number represents how much information these terms carry in a given document. Term frequency is the number of times a word w appears in a document divided by the total number of words W in the document, and IDF is log (total number of documents (N) divided by the total number of documents in which word w appears (n)).

d) Word embedding:

Another feature extraction method is word embedding based on neural networks. In this method, the words with the same semantics or those related to each other are represented by similar vectors. This is more popular in word prediction as it retains the semantics of words. We are planning to use TF IDF as feature extraction for the malayalam sentiment analysis.

(4) Model Develeopment

For the model development we have several approaches as machine learning based approach, lexicon based approach and hybrid based approach. In this project we are planning to use machine learning based approach. Machine learning algorithms include such as NB, SVM and RF have been applied for predicting the sentiment of Malayalam texts. The selection of hyper parameter is most challenging for accurate prediction of data.

a) Naive Bayes Classifier: NB predicts the sentiment of the test dataset as positive or negative using a Multinomial NB classifier. This classification is done based on Bayes' theorem.

b) Linear Regression: It is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

c) Logistic Regression: It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s).

In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1.

d) Support Vector Machine: SVM is a supervised machine learning algorithm proposed by Vapnik in 1992. SVM finds the linear separator with maximum marginal distance using support vectors in high dimensional space. Both linear and RBF kernel functions have been used.

e) K-Nearest Neighbors (kNN): K-nearest neighbors (kNN) is a supervised learning algorithm that can be used to solve both classification and regression tasks.

The main idea behind kNN is that the value or class of a data point is determined by the data points around it.

f) K-Means clustering: Clustering is a way to group a set of data points in a way that similar data points are grouped together. Therefore, clustering algorithms look for similarities or dissimilarities among data points. Clustering is an unsupervised learning method so there is no label associated with data points. Clustering algorithms try to find the underlying structure of the data.

g) Random Forest: RF is a supervised machine learning algorithm created by Tin Kam Ho in 1995. It builds multiple decision trees and merges together for the classification of data. It searches for the best feature among a random subset of features. In this project we are planning to use NB algorithm.

h) Decision Tree: It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

Nowadays, deep learning algorithms such as Convolutional Neural Networks (CNNs),Long Short Term Memory Networks (LSTMs),Recurrent Neural Networks (RNNs),Generative Adversarial Networks (GANs),Radial Basis Function Networks (RBFNs),Multilayer Perceptrons (MLPs) are also used for sentiment analysis model development.

V. EXPERIMENTAL RESULT AND DISCUSSIONS

The results of works proposed by Bo Pang *et.al [1]*produced via machine learning techniques are quite good in comparison to the human generated baselines .In terms of relative performance, Naive Bayes tends to do the worst and SVMs tend to do the best, although the differences aren't very large.They were not able to achieve accuracies on the sentiment classification problem comparable to those reported for standard topic-based categorization, despite the several different types of features they tried. Unigram presence information turned out to be the most effective; in fact, none of the alternative features they employed provided consistently better performance once unigram presence was incorporated. Interestingly, though, the superiority of presence information in comparison to frequency information in our setting contradicts previous observations made in topic-classification work .A human would easily detect the true sentiment of the review, but bag-of-features classifiers would presumably find these instances difficult, since there are many words indicative of the opposite sentiment to that of the entire review. Fundamentally, it seems that some form of discourse analysis is necessary using more sophisticated techniques than our positional feature. Hence, they believe that an important next step is the identification of features indicating whether sentences are on-topic (which is a kind of co-reference problem); they look forward to addressing this challenge in future work.

Turney et. Al *[2] t*hat semantic orientation is also useful for summarization of reviews. A positive review could be summarized by picking out the sentence with the highest positive semantic orientation and a negative review could be summarized by extracting the sentence with the lowest negative semantic orientation.This paper introduces a simple unsupervised learning algorithm for rating a review as thumbs up or down. The algorithm has three steps: (1) extract phrases containing adjectives or adverbs, (2) estimate the semantic orientation of each phrase, and (3) classify the review based on the average semantic orientation of the phrases. The core of the algorithm is the second step, which uses PMI-IR to calculate semantic orientation (Turney, 2001).In experiments with 410 reviews from Epinions, the algorithm attains an average accuracy of 74%. It appears that movie reviews are difficult to classify, because the whole is not necessarily the sum of the parts; thus the accuracy on movie reviews is about 66%. On the other hand, for bank and automobiles, it seems that the whole is the sum of the parts, and the accuracy is 80% to 84%.The limitations of this work include the time required for queries and, for some applications, the level of accuracy that was achieved. The former difficulty will be eliminated by progress in hardware. The latter difficulty might be addressed by using semantic orientation combined with other features in a supervised classification algorithm.

Pramod K.V et. Al *[3]* SA of Malayalam tweets using NB, SVM, and RF are proposed in this work. Four different features like BOW, TF IDF, Unigram with Senti-wordnet, and Unigram with Senti-wordnet including negation words, are considered for feature vector formation of the input dataset. All the classifiers with the last two features have shown better accuracy compared with other features. RF classifier with Unigram with Senti wordnet including negation words, got the highest accuracy, 95.6%.

The classification or sentiment analysis (SA) is regarded as a specific case of test classification. Despite the number of classes in sentiment analysis being small, the process of sentiment classification is complex than the traditional topic text classification R Prabowo et.al*[19]*. The classification in topic text classification depends on the use of keywords; however, it does not work efficiently in case of sentiment analysis. The nature of the problem defines other difficulties in sentiment analysis. The negative sentiment may sometimes be represented in a sentence without using any notable negative words. In addition, there is a fine line between whether a sentence should be considered subjective or objective. Identifying the opinion holder, the person who voices sentiments in the text is the most complex task in sentiment analysis. The sentiment analysis greatly relies on the subject or field of the data. The words may sometimes have a positive sentiment in a particular field, and the same words may have another polarity sentiment in a different field [19].

Neural network and deep learning shows more accuracy than usual machine learning algorithms.

VI. CONCLUSION & FUTURE SCOPE

In this paper, we mainly focus on the basics of sentiment /opinion mining and its levels. There are various approaches and methods to identify sentiment from content. In this paper, our examination represents machine learning procedures. From various classification methods, Sentiment Analysis indicates the results into positive, negative and neutral scores. The study shows that machine learning methods, such as SVM, Naive Bayes, and neural networks have the highest accuracy and can be considered as the baseline learning methods as well as in some cases lexicon-based methods are very effective. Neural network and deep learning shows more accuracy than usual machine learning algorithms. In future work, discovering the result of various other combinations of text data and other on prediction accuracy can be done. More work in the future is needed to improve performance measures.

REFERENCES

[1] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural LanguageProcessing-Volume.10Association for Computational Linguistics,2002.

[2] Peter D. Turney, semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational

Linguistics, 2002.

[3] Soumya, S., and K. V. Pramod. "Sentiment analysis of malayalam tweets using machine learning techniques." ICT Express 6.4 (2020): 300-305.

[4] Yequan Wang, et al., Sentiment analysis by capsules, in: Proceedings of the 2018 World Wide Web Conference, 2018.

[5] Ning Liu, et al., Attention-based sentiment reasoner for aspect-based sentiment analysis, Hum.-Cent.Comput. Inf. Sci. 9 (1) (2019) 35.

[6] Deepu S. Nair, et al., Sentiment analysis of malayalam film review using machine learning techniques, in:2015 International Conference on Advances in Computing, Communications and Informatics,

ICACCI,IEEE, 2015.

[7] S. Soumya, K.V. Pramod, Sentiment analysis of malayalam tweets using different deep neural network models-case study, in: 2019 9th International Conference on Advances in Computing and Communication,ICACC, IEEE, 2019.

[8] S. Sachin Kumar, M. Anand Kumar, K.P. Soman, Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets, in: International Conference on Mining Intelligence and Knowledge Exploration, Springer, Cham, 2017.

[9] M. Rahul, R.R. Rajeev, S. Shine, Social Media Sentiment Analysis for Malayalam, 2018.

[10] V. Kasthoori, B. Soniya, V. Jayan, Domain-independent sentiment analysis in malayalam, in:Computational Intelligence: Theories, Applications and Future Directions-Volume II, Springer, Singapore,2019,pp. 151–160.

[11] M.P. Ashna, Ancy K. Sunny, Exicon based sentiment analysis system for malayalam language, in: 2017 International Conference on Computing Methodologies and Communication, ICCMC, IEEE, 2017.

[12] P.K. Thulasi, K. Usha, Aspect polarity recognition of movie and product reviews in Malayalam, in: 2016 International Conference on Next Generation Intelligent Systems, ICNGIS, IEEE, 2016.

[13] M. Anagha, et al., Fuzzy logic based hybrid approach for sentiment analysis of malayalam movie reviews, in: 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, SPICES, IEEE, 2015.

[14] P. Jayan, Deepu S. Nair, S. Elizabeth Jisha, A subjective feature extraction for sentiment analysis in Malayalam language, Int. J. Eng. Sci. 14 (2015) 1–4.

[15] Deepu S. Nair, Jisha P. Jayan, Elizabeth Sherly, Sentima-sentiment extraction for malayalam, in: 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI, IEEE,2014.

[16] Neethu Mohandas, Janardhanan P.S. Nair, V. Govindaru, Domain specific sentence level mood extraction from malayalam text, in: 2012 International Conference on Advances in Computing and Communications, IEEE, 2012.

[17] Nair, S. Deepu, J.P. Jayan, R.R. Rajeev, E. Sherly, "Sentiment Analysis of Malayalam film review using machine learning techniques", In Advances in Computing, Communications and Informatics (ICACCI), International Conference on, pp.2381- 2384, IEEE, 2015.

[18]Mohandas, Neethu, J.P.S. Nair, V. Govindaru, "Domain specific sentence level mood extraction from malayalam text", Advances in Computing and Communications (ICACC), International Conference on. IEEE, 2012

[19]R. Prabowo and M. Thelwall." Sentiment analysis: A combined approach". Journal of Informetrics , 3(2): 143-157,b2009.