IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

A Study on Sentiment Analysis in Malayalam Language

Ashna M.P [1], Ancy K Sunny [2]

PG Student, Computer Science & Engineering, Vimal Jyothi Engineering College, Kannur, India [1]

Assistant Professor, Computer Science & Engineering, Vimal Jyothi Engineering College, Kannur, India[2]

Abstract: Sentiment Analysis is a natural language processing task that mines opinion information from various text forms such as reviews, news, and blogs and classifies themby their polarity as positive, negative or neutral. With the rise of information being communicated via regional languages like Malayalam, comes a promising chance of mining this information. The works are very less in dialectal languages like Malayalam even though so many are there for universal languages like English. Mining sentiments in Malayalam comes with a lot of issues and challenges. As compared to English, Malayalam is a free order and morphologically rich language, which adds complexity while handling the user-generated content. This paper gives anoverview of the sentiment analysis works that has been donein the Malayalam language and challenges faced in these works.

Keywords: Sentiment Analysis, Natural Language Processing, Malayalam Sentiment Analysis.

I. INTRODUCTION

Sentiment analysis is one of the most active research areas in Natural Language Processing (NLP). Sentiment analysis or Opinion Mining is used to extract the opinions that appear on the web to evaluate the attitude and judgment of the opinion holder about a particular area. It is widely used to analyze reviews, social media, and blogs for sentiments and views expressed on products, services, individuals, and organizations. Polarity detection is the most common form of sentiment analysis, i.e., determining whether the sentiment expressed in a review is positive, negative or neutral.

The social media, blogs, forums and e-commerce websites encourage people to share their opinions and feelings publically. People's views and experiences are precious information in the decision-making process. Nowadays it is normal to look at reviews before buying a new product or watching a new movie. However, reading a review fully is a time-consuming task and more than that most of them does not provide a final verdict. So it is desirable to have an automated sentiment analysis system that identifies the sentiment expressed in a review.

Malayalam is a language spoken in India, predominantly in the state of Keralaand was designated as a Classical Language in India in 2013. With the rise of information being communicated via regional languages like Malayalam, comes a promising chance of mining this information. The works on sentiment analysis are very less in dialectal languages like Malayalam even though so many are there for universal languages like English. Mining sentiments in Malayalam comes with a lot of issues and challenges. Malayalam is morphologically rich and is a free order language as compared to English, which adds complexity while handling the user-generated content.

This paper is an attempt to study the concepts of Sentiment Analysis and compare various sentiment analysis works in Malayalam. The rest of the paper is organized as follows: an overview of the sentiment analysis techniques is presented in section II and III. Section IV discusses various Sentiment Analysis works carried out in Malayalam and in section V comparison of those works are discussed. Section VI contains challenges present in Sentiment Analysis research. Finally, the paper concluded in section VII.

II. LEVELS OF SENTIMENT ANALYSIS

Depending upon which type of data is to be processed sentiment analysis can be performed mainly at three levels. They are sentence level, document level, and aspect level [1].

A. Sentence Level Sentiment Analysis

In sentence-level sentiment analysis, each sentence in the input text areanalyzed separately and classified as positive, negative or neutral. Sentence level sentiment analysis is performed in two steps. The first step is subjectivity classification, which determines whether the given sentence is subjective or objective. i.e., whether the sentence expresses an opinion or not. Sentiment classification is the second step which is performed only on the subjectivesentences. Sentiment classification classifies the subjective sentence into two categories, which is positive

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

and negative. Sentence level sentiment analysis assumes that the given sentence expresses a single opinion from a single opinion holder. This assumption is only suitable for simple sentences with a single opinion. It is not appropriate for compound sentences since more than one opinion may be expressed in a single compound sentence like "The picture quality of the phone camera is amazing,but the battery life is poor." One drawback of using theapproach mentioned above is that many objective sentences can express opinions or sentiments and conversely many subjective sentences may not express any opinions or sentiments.

B. Document Level Sentiment Analysis

Document level sentiment analysis classifies the input document into positive, negative or neutral based on the overall opinion expressed in thatdocument. For example, given a product review, the system determines the overall opinion about the product that is either positive or negative.This level of analysis assumes that each document expressesopinion about a single entity (product or service). Thus document-level sentiment analysis is not applicable to a forum or blog post because in such post the author may express opinion about multiple entities and compare them. This type of analysis is more suitable for classifying customer reviews such as product reviews or movie reviews.

C. Aspect/Feature Level Sentiment Analysis

The two approaches mentioned above works with either the whole document or each sentence. In many cases,objects may have many aspects or features. The author may write both positiveand negative opinions about different features of the object, although the overall sentiment on the object may be positive or negative.That means a positively opinionated review on a particular entity does not mean that the user has positive opinions on all aspects or features of that entity. Similarly, a negatively opinionated review does not mean that the user dislikes everything. So the sentiment classification should be performed on each feature of the entity. Aspect /Feature-based sentiment analysis focuses on twosub-tasks. First one is to identify different object features that have been commentedon, and next is to determine whether the opinions on these features are positive, negative or neutral. This method is suitable for analyzingreviews about products or posts in discussion forums related to specific product categories (such as smartphones, cameras,cars,etc.).

III.SENTIMENT ANALYSIS APPROACHES

Sentiment Analysis (SA) techniques can be broadly classified into machine learning techniquesand lexicon based techniques [1] as shown in Fig 1.

A. Machine Learning Approach

Machine learning approach works by training an algorithm with a training data set before applying it to the actual data set. Machine learning techniques first train the algorithm with some particular inputs with known outputs so that later it can work with new unknown data.

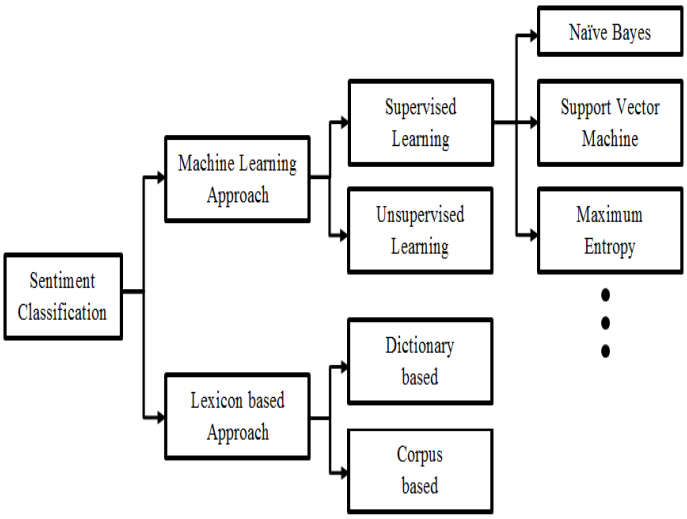

Fig.1 Classification of sentiment analysis techniques

Machine learning approaches can be further classified into two: Supervised learning and unsupervised learning.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

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 [4]. 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
Unsupervisedlearning. Butunsupervised learning benefits with the lesscost required for it because it does not need a huge set of annotated data.

A. Lexicon-based Approach

Lexicon basedapproach workson the assumption that the overall polarity of a sentence or documents is the sum of polarities of the individual phrases or words. Thismethodrelies onfinding semantic orientation(positivity,negativity or the neutrality) of the text by referring to thelexicon or dictionaries.These lexicons identify the positive ornegative label regardingthe polarity for each word and its strength related totheir meaning. These dictionaries can be created manually or automatically with the help of seed words.One of the commonly usedsentimentlexiconis SentiWordNet. Instead of using the available sentiment lexicon we can also implement one manually or by using two approaches: Dictionary based and Corpus-based.

i) Dictionary-based Approach:

Dictionary based approach relies on bootstrappingusing an online dictionary (e.g., WordNet) and a small set of seed opinion words. The methodis first tocollect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet for their synonyms and antonyms. The newly found words are added to theseed list, and next iteration starts. This iterative process stops when no newer words are found[11].

ii) Corpus-based Approach:

The corpus-based approach relies on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus [9].The procedure starts with a list of seed opinion adjective words, and by using them and a set of semantic constraints, new adjective opinion words and their orientations are identified. One of the constraints is about conjunction (AND), which states that conjoined adjectives typically have a similar orientation. Kanayama and Nasukawa [10] extended this methodology by presenting the idea of intra-sentential (within a sentence) and inter-sentential (between neighboring sentences) sentiment consistency. The intra-sentential consistency is similar to that in [9]. Inter-sentential consistency applies the idea to nearby sentences. That is, it states that the same opinion orientation (positive or negative) is usually conveyed in a few consecutive sentences and opinion changes are specified by opposing expressions such as “but” and “however”.

Lexicon based approach for sentiment analysis is easy to implement, andadvantages of this method include its nature for being domain independent.

B. Comparison of different approaches used for sentiment analysis

Each approach for performing sentiment analysis has its ownadvantagesand disadvantages as shown in TABLE I.

TABLE I PROS AND CONS OF SENTIMENT ANALYSIS APPROACHES

Approaches	Advantages	Disadvantages
Machine Learning	Yields high accuracy of classification	A classifier trained on the texts in one domaindoes not work with other domains. A Large amount of training corpus is required.
Lexicon based	Labeled corpus and learning process	Requires powerful linguistic resources which

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

	is not required. Lexicon created for one domain can be used for other domains with small changes.	are not always available.

In most of the cases the supervised machine learning approaches outperformed the unsupervised lexicon based approaches. But, the requirement of huge annotated training dataset for supervised machine learning approaches; force the researchers to accept the unsupervised methods, as it is very easy to collect unlabelleddataset.
In lexicon based approaches the performance depends up onthe lexicondictionary used. If the dictionary contains only fewwordsthen it leads to performance degradation. One of the mainchallenge in the sentiment analysis is thedetermination of polarity of the sentiment words. The polarity orientation of each wordis completely depends on the domain. The currently available sentiment lexicons such as SentiWordnet fails to capturethe context sensitivity of sentiment words. The lexicon based sentiment analysis gives low recall if it is being notused with well-built sentiment lexicon dictionary.
TABLE II summarizes various sentimentanalysis techniques alongwith the accuracy attained in their evaluation, as per thedata provided by the authors.

TABLE II ACCURACY OF DIFFERENT SENTIMENT ANALYSIS TECHNIQUES

Paper	Approach	Dataset	Technique	Accuracy
Pang et al. [12]	Supervised	Movie review	SVM	82.9%
			Naïve Bayes	81.5%
			Maximum Entropy	81%
Harb et al.[13]	Unsupervised	Movie review	Lexicon	71%
Sharma et al.[14]	Unsupervised	Product review	Dictionary based	74%
Joshi et al.[15]	Supervised	Travel Reviews	ML- based	78.14%
			MT-based	67%
	Unsupervised		Senti WordNet Dictionary	60.3%
Trilla et al. [16]	Supervised	Semeval	SVM	58.12%
		Twitter		72.76%
Zhang et al. [17]	Hybrid	Twitter tweets	ML and Lexicon	85.4%
Abbasi et al [18]	Supervised	Movie reviews	SVM	95.5%

IV. SENTIMENT ANALYSIS IN MALAYALAM LANGUAGE

In Malayalam works on sentiment analysis is still in its initial stages. Only a few works have been reported in Malayalam so far.

A. SentiMa-Sentiment Extraction for Malayalam [4]
This paper proposes a rule-based approach for extracting sentiments from Malayalam movie reviews. Sentence-level sentiment extraction is used in this paper. They state that sentence-level sentiment extraction is effective since in movie websites the user comments are just single sentences. Negation rules are used for analysing the sentiments whichreduce the chance of occurrence of errors. The proposed method firstcollects the corpus from movie websites or various blogs, newspapers, and magazines. Then the sentences are split into various tokens using sandhirules. After this, each word is checked with the pre-stored positive or negative categories of theword. If that word does not match with either positive or negative category, it can be considered as aneutral word. If theword matches to the positivecategory, then that word is marked as positiveand negative if it matches to negative category. After assigning apolarity to each word, negation rule is applied to find the overall polarity. The overall sentiment of the given text is calculated by taking thecount of positive and negative words. They achieved an accuracy of 85%.

B. Domain Specific Sentence Level Mood Extraction from Malayalam Text [3]
This paper focused onaspecific domain because different domains may use different words to express the mood. The sentiment extraction process first manually collects the corpus from Malayalam novels and then pos-tagging of the input sentence is performed toextract the adjective and adverb since they are the most emotionalbearing phrases. Then scoring the sentenceis carried out by calculating the semantic orientation of the sentence using the extracted patterns. The SO-PMI-IR formulas classify an input text into one of the two classes that indicate desirable or notdesirable [4]. The formulas have to be modifiedappropriately so that it classifies the input into one ofthe four classes:Joy, Sorrow,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

Anger or Neutral.In this paper, both the corpus creation and POStagging is done manually to reduce error and also forsimplicity. The paper concluded with an accuracy of 63%.

C. A Novel Hybrid Approach Based on Maximum Entropy Classifier for Sentiment Analysis of Malayalam Movie Reviews [8]
This paper performs sentiment classification of Malayalam movie reviews obtained from theuser as positive, negative and neutral. A hybrid approach is used here. i.e., acombination of Maximum Entropy Model which used for tagging and certain rules for handling special cases. Maximum entropy classifier is a probabilistic classifier which belongs to the class of exponential models for categorizing without knowing the prior knowledge. It selects maximum entropy from all the models that fit our training data. Maximum Entropy Classification finds out which class the review must belong to the given a context so that it maximizes the entropy of the classification system. Here also some rules are applied to handle special cases which include negation, intensifiers, dilators,etc. Here the number of tagged classes is increased to seven from three (i.e. positive, negative and neutral). The new classes are inverse negative, intensifier, dilator and special. To find out the overall polarity, the total number of positive and negative words are counted. This method gave an accuracy of 93.6%.

D. Lexical Resource-based Hybrid Approach for Cross-Domain Sentiment Analysis in Malayalam [7]
This paper proposed a Lexical Resource-based approach to extract sentiments from domain independent Malayalam reviews. The proposed method finds out the polarity of opinion words in the input text with the help of Hindi WordNet-based lexical resource file created. Machine Learning method is used for tagging certain special cases. This approach also gives a better accuracy of 93.6%.

E. Sentiment Analysis of Malayalam film review using machine learning techniques [6]
Here a sentiment analysis system for Malayalam movie reviews is implemented by using a combined approach of machine learning techniques. In this work two statistical machine learning techniques, CRF combined with rules and SVM combined with rules are used separately. They concluded by saying that SVM outperforms CRF with an accuracy 91%.

V. COMPARISON OF VARIOUS SENTIMENT ANALYSIS WORKS IN MALAYALAM

TABLE III summarizes various works on sentiment analysis in Malayalam Language along with the accuracy obtained.

TABLE III COMPARISON OF SENTIMENT ANALYSISWORKS IN MALAYALAM

Paper	Level	Dataset	Approach	Accuracy
Neethu Mohandaset al.[3]	Sentence	Malayalam Novels	Semantic Orientation with PMI-IR	63%
Deepu S. Nairet al. [4]	Sentence	Movie Reviews	Rule-based using negation rule	85%
Anagha M et al.[8]	Sentence	Movie Reviews	Hybrid approach (Maximum entropy + Rule based)	93.6%
Anagha M et al.[7]	Sentence	Movie Reviews	Lexical resource based hybrid approach	93.6%
Deepu S.Nair et al.[6]	Sentence	Movie Reviews	Machine learning (SVM , CRF)	SVM is better with 91% accuracy

It is evident from the survey that only a few works are available for sentiment analysis in Malayalam. Most of the works related to sentiment analysis are done using machine learning approach and achieved a high accuracy level of 93.6%. Since there are no annotated corpora is available for Malayalam, we required to create and label a corpus manually to perform supervised learning based sentiment analysis. Lexicon based method for sentiment polarity classification is much faster than a supervised approach while yielding similar accuracy.

VI. CHALLENGES IN SENTIMENT ANALYSIS

General challenges in sentiment analysis are:

- **Noise (abbreviations, slangs):** Noise on the web is increasing day by day. Abbreviations, slangsand emotions are commonly used by people for ease of use. But for language processing, these increases the complexity.
- **Unstructured Data:** Web contains a large amount of unstructured data.The sources of web varies from web documents, journals, books,videos, audios, imagesetc. So, this diversity in the sources of data and different formats increases the complexity

- **Contextual Information:** Actual sense of the text varies from domain to domain; this property is referred as contextual property. So, based on the context, the polarity of the word changes.
- **Lack of resources:** Lack of sufficient tools, resources, and annotated corpora lead to great struggle while doing sentiment analysis for Indian languages. Lack of standard datasets makes collection/creation of dataset a time-consuming task. Comparison of techniques applied and results obtained, is a difficult task in the absence of standard data set.
- **Morphological Variations:** Handling the morphological variations is also a bigchallenge for Indian languages. Indian languages are morphologically rich which means that lots of information are fused in words as compared to the English language where we add another word for the extra information.
- **Negation Handling:** Handling negation is a challenging task in sentiment analysis because negations can be expressed in various ways even without the use of the negative word.
- **Implicit Sentiment and Sarcasm:** Sentences may carry implicit sentiments, which meanthe opinion can be expressed without having any sentiment-bearing words in it.

VII. CONCLUSION

Sentiment Analysis has been the focus of research community from last decade. With the increase of information being communicated via regional languages like Malayalam, comes a promising opportunity of mining this information. The works are very less in dialectal languages like Malayalam even though so many are there for universal languages like English. Much of the research in Malayalam sentiment analysis has been done using different supervised learning techniques. Although the Supervised methods provide better accuracy compared to dictionary-based approach, supervised learning method cannot perform well without sufficient training examples. Since there are no annotated corpora is available for Malayalam the training data should be collected and labeled manually. Labeling training data is tedious and time-consuming. Also, the accuracy of supervised learning method is directly related to the quality of training corpus created. One solution to this problem is by using Dictionary based approach. Dictionary based approach takes less processing time compared to supervised learning techniques. Another drawback of the existing system is since sentiment analysis is a domain dependent task; a corpus once created cannot be used for other areas. However, the sentiment lexicon created can be utilized for the entire area by making small changes in the polarity of seed words.

REFERENCES

[1] Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." Mining text data. Springer US, 2012. 415-463.
[2] P. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424, 2002.
[3] Mohandas, Neethu, Janardhanan PS Nair, and V. Govindaru. ”Domain Specific Sentence Level Mood Extraction from Malayalam Text. ” Advances in Computing and Communications (ICACC), 2012 International Conference on. IEEE, 2012.
[4] Nair, D. S., Jayan, J. P., Rajeev, R. R., Sherly, E. ”SentiMa-Sentiment extraction for Malayalam.”Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE,2014.
[5] Anagha M, Raveena R Kumar, Sreetha K, Rajeev R.R,P.C.Raghu Raj,”Lexical Resource based Hybrid Approach forCorpus basedSentiment Analysis in Malayalam” ,International Journal of Engineering Sciences,2014.
[6] Nair, Deepu S., Jisha P. Jayan, and Elizabeth Sherly. "Sentiment Analysis of Malayalam film review using machine learning techniques." Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on. IEEE, 2015.
[7] Anagha, M., Raveena R. Kumar, K. Sreetha, R. R. Rajeev, and PC ReghuRaj. "Lexical Resource Based Hybrid Approach For Cross Domain Sentiment Analysis in Malayalam." An International Journal of Engineering Sciences, Special Issue iDravadian , December 2014.
[8] Anagha M, Raveena R Kumar ,Sreetha K and P C Reghu Raj. “A Novel Hybrid Approach on Maximum Entropy Classifier for Sentiment Analysis of Malayalam Movie Reviews", International Journal Of Scientific Research,Volume : 4 Special Issue June 2015 ,ISSN No 2277 -8179.
[9] V. Hatzivassiloglou and K. McKeown, “Predicting the semantic orientation of adjectives,” Proceedings of the Joint ACL/EACL Conference, pp. 174–181, 1997.
[10] H. Kanayama and T. Nasukawa, “Fully automatic lexicon expansion for domain-oriented sentiment analysis,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 355–363, July 2006.
[11] M. Hu and B. Liu, “Mining and summarizing customer reviews,” Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168–177, 2004.
[12] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.
[13] A. Harb, M. Planti, G. Dray, M. Roche, Fran, o. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, Cergy-Pontoise, France, 2008.
[14] Sharma, Richa, Shweta Nigam, and Rekha Jain. "Mining of product reviews at aspect level." arXiv preprint arXiv:1406.3714 (2014).
[15] Joshi, Aditya, A. R. Balamurali, and Pushpak Bhattacharyya. "A fall-back strategy for sentiment analysis in hindi: a case study." Proceedings of the 8th ICON (2010).
[16] Trilla, Alexandre, and Francesc Alias. "Sentence-based sentiment analysis for expressive text-to-speech." IEEE transactions on audio, speech, and language processing 21.2 (2013): 223-233.
[17] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B.Liu,“Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis”, Technical report, HP Laboratories, 2011.
[18] A. Abbasi, H. Chen, and A. Salem, “Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums,” In ACM Transactions on Information Systems, vol. 26 Issue 3, pp. 1-34, 2008.