Sentiment Analysis on Malayalam Reviews: A Survey

Deepa Mary Mathews[1], Sajimon Abraham[2]

[1]*School of Computer Sciences, Mahatma Gandhi University, Kerala, India*

[2] *School of Management and Business Studies, Mahatma Gandhi University, Kerala, India*

ABSTRACT:

The exponential boom of social data in social media and the challenge to take the decisions in business using computers formulate Sentiment Analysis a demanding and attention-grabbing research problem. Sentiment Analysis is a scientific approach that is inseparably vault to the field of Affective Sciences which concerns the individual suppositions, feelings or notions and thereby discovers the cognitive deeds of humans. Communicating the slants in their own native language can be considered as the most comfortable way of expressing the viewpoints. This leads to the necessity of Sentiment Analysis in various local dialects. This paper summarizes the major contributions in the field of Sentiment Analysis that can be utilized for looking into Malayalam audits.

Keywords: *Lexicon based, Malayalam, POS Tagging, Sentiment Analysis, Stemming*

1. Introduction

Social Data is being generated at an unprecedented rate every second. People utilize online networking to express their sentiments and standpoints about a specific event. Retrieving and evaluating real time social data is gaining popularity because the dynamic trends and outlooks are updated right away on such platforms. The viewpoints marked by the users are in unstructured format and analyzing these texts gained attention of many researchers as it has a significant role in decision making. In the last decade, the viewpoint analysis gained lots of significance and many researchers came out with different strategies for performing sentiment analysis. Standpoints of the social users on diverse topics communicated in their own mother tongue leads to the necessity of mining the sentiments in various dialects. Though the percentage of these dialects compared to English dialect is low, the analysis of these kind of information need to be done to identify many valuable frequent patterns at a regional level. This type of analysis gained a lot of popularity as it contains recommendations and suggestions. Compared to the high-resource languages such as English, sentiment analysis task in low-resource language suffers mainly due to the absence of annotated corpus and the tools to extract features. The present paper brings out the works being carried out in Malayalam language in the field of Sentiment Analysis.

The rest of the article is structured as follows. A brief description about Sentiment Analysis and the significance of Sentiment Analysis in Malayalam language is depicted in Section 2 and 3. Major Contributions in POS Tagging,

Stemming and Sentiment Analysis in Malayalam dialect are summarized in section 4. Finally the article is concluded in the section 5.

2. Sentiment Analysis

Sentiment Analysis is the process of identifying the polarity of the raw data. The sentiment of the Malayalam reviews marked in the social media is extracted and the whole opinion is categorized into mainly three classes - positive, neutral and negative. The two commonly used approaches to perform Sentiment Analysis are Supervised and Unsupervised approach. In the supervised approach, the previous experiences are used to classify the test set. The classification can be done based on Lexicon based approach or by using Machine Learning algorithms. In the unsupervised approach, no labeled set is provided with; instead the Sentiment Orientation (SO) of opinion words is determined. Each sentiment bearing word is associated with a sentiment score. Different strategies like point wise mutual information method can be used to calculate the score.

Sentiment Analysis can be carry out based on different levels of granularity. In Document level Sentiment Analysis, the polarity or the orientation of the reviews are calculated based on the overall opinion expressed in the whole document which consists of the reviews about the same issue. In the Sentence level Sentiment Analysis, each review sentence is analyzed and classified based on their polarity. If the reviews consists of comments about different aspect or features of an object, then aspect based Sentiment Analysis is more suitable. The two components of Machine Learning approach are pre-processing to clean the data, and the classification of Sentiments.

3. Significance of Sentiment Analysis in Malayalam Language

In India, nearly 30 official local languages are there and more than 35 million people spreading along the regions of Kerala, Pondicherry and Lakshadweep are having Malayalam as their local dialect. It is a language of the Dravidian family with a rich literary tradition. Mainly Malayalam was influenced by Sanskrit which was brought into Kerala by Brahmins. Malayalam absorbed a lot from Sanskrit, not only in the lexical level, but also in the phonemic, morphemic and grammatical levels of language. Lots of individuals used to convey viably in their local dialect hence creating huge amount of information. This information can well be routed to extract many valuable patterns like the customers buying pattern, product feedback, and so on at a regional level. The aforementioned work is lagging in these local dialects leads to the necessity of having opinion mining in regional languages too.

The overall sentiment of User Generated Content (UGC) can be determined by the polarities of different sentiment words used in UGC. The term that expresses the viewpoints or sentiments is called as lexicons. The set of these vocabularies are usually used to convey positive or negative sentiments. For example, 'നന്നായി', 'പ്രശംസിച്ചു', 'നടേടങ്ങ□', 'ആക□ഷണീയമായ', 'ആക□ഷകമായ', 'സമ്പന്നമായ', 'മനോഹരമായ', 'അഭിനന്ദിക്കുക' etc are positive Malayalam sentiment words, and 'മോശമായ', 'നിന്ദിക്കുക', 'അബദ്ധം', 'ചീത്തപറയുക', 'നിഷ്ഠൂരമായ' etc are negative Malayalam sentiment words.

4. Major Contributions

Since significantly high percentage of the data generated is in the Universal language English, all these works were focused on the reviews in the English dialect. Standpoints of the social users on diverse topics communicated in their own mother tongue leads to the necessity of mining the sentiments in various dialects. Though the percentage of these dialects compared to English dialect is low, the analysis of these kind of information need to be done to identify many valuable frequent patterns at a regional level.

4.1. Pre-processing of the Corpus

The amount of data to be considered for performing sentiment analysis can be considerably reduced by the pre-processing phase [34]. This step is required to speed up the process while dealing with the large corpora. There has been a lot of works in Indian languages for developing stemmers, taggers, translators etc and the contributions in the pre-processing phases needed for doing sentiment analysis in Malayalam language is depicted here. The major works in the area is grouped to have a better outlook and is listed in Table 1, 2 and 3.

Table 1: Summary of major works in POS Tagging

Year	Author Details	Methodology & Outcome	Accuracy
2018	[1] Kumar.S, *et al*	Deep Learning Based Part-of-Speech Tagging for Malayalam Twitter Data	
2018	[2] Ajees A.P *et al*	POS Tagger for Malayalam using Conditional Random Fields	91.2%
2016	[3]Kumar S.S *et al*	POS tagger using Epic framework in Scala.	87.35%
2015	[4] D. Muhammad Mubarak *et.al*	An approach using multithreaded technology to do Parts Of Speech Tagging in Malayalam	
2014	[5] Rinku T S *et.al*	Analysis on various Approaches used for tagging and chunking in Malayalam"	
2013	[6] Robert, et.al	IB1 algorithm implemented with TiMBL tagger tool	
2011	[7] Jisha P.J et.al	Statistical approach to do tagging and chunking	Tagging 91%, Chunking 92%
2010	[8] P J Antony, et.al	SVM Based Part of Speech Tagger	
2010	[9] Rajeev R et.al	Part of speech tagging for Malayalam using SVM Tool and TnT tagger based on Hidden Markow model.	SVMTool 88%, TnT tagger 80%
2009	[10] Manju K, et.al	Stochastic approach using word frequencies and bigram statistics	

Table 2: Summary of major works in Stemming/Lemmatization

Year	Author Details	Methodology & Outcome	Accuracy
2018	[11] Dhanya, P. M. et.al	A Tree based Malayalam Lemmatizer named Vriksh is developed using Suffix Replacement dictionary	87%
2016	[12] C. Balasankar, et.al	Multi level inflection handling stemmer using iterative suffix-stripping	
2014	[13] Nisha M, et.al	Machine Learning Approach for Malayalam Root Word Identification	92%
2013	[14] Prajitha, U., et.al	One pass Suffix stripping, used Suffix dictionary, scans from right to left for the longest match	86%
2013	[15] Vasudevan, N.,et.al	Semi supervised stemming through stem set minimization	80%
2013	[16] Pragisha K., et.al [4]	Three pass Suffix Stripping Rule based system	97%
2012	[17] Meera Subhash, et.al	Rule based approach for root word identification in Malayalam language.	90%
2012	[18]Vasudevan, N.,	Proposed a probabilistic model that minimizes the stem distributional entropy for stemming	

4.1.1. POS Tagging

In Natural Language Processing (NLP), one of the well-studied problems under constant exploration is part-of speech tagging or POS tagging or grammatical tagging. The task is to assign labels or syntactic categories such as noun, verb, adjective, adverb, preposition etc. to the words in a sentence or in an un-annotated corpus. The contributions in this area are listed in Table 1. The SVMTool is a software package that can be used for Malayalam POS tagging.

4.1.2. Stemming

Malayalam is highly agglutinative in nature and hundreds of inflections are possible for each word. Stemming which is a pre-processing step to have a better recall, is the process of removing the affixes from inflections and to return the root form. Related works in this area is listed in Table 2.

4.2 Sentiment Classification System

Sentiments can be analyzed either using Lexicon based methods or using Machine Learning algorithms. Cross Validations and various evaluation metrics can be used to evaluate the performance of the method. These types of applications and analysis help the people to communicate and work in their own mother tongue rather than

depending on other languages. The articles that mentions and explains the Sentiment Analysis on Malayalam dialect are summarized in Table 3.

Table 3: Summary of major works on Sentiment Analysis in Malayalam language

Year	References	Major steps of Methodology	Datasets Used	Remarks
2019	Kasthoori V.,et.al [19]	Domain-Independent Sentiment Analysis	Malayalam online newspapers	Sentence-level sentiment analysis using machine learning method and fuzzy logic
2018	R. Jayakrishnan.,et.al [20]	SVM classifier is used for sentence level multi-class emotion detection	Malayalam Reviews	uses different syntactic features such as n-gram, POS related, negation related, level related features etc
2017	Kumar,et.al [21]	Deep learning methods such as CNN and LSTM;	12922 tweets;	Explores the effect of ReLU, ELU and SELU; cross-validation to support it; LSTM with SELU depicts 98.24% accuracy.
2017	M. P. Ashna., et.al [22]	Lexicon based sentiment analysis system	Malayalam Reviews	Sentence Level - 87.5%, Document Level -90% accuracy
2016	P.K. Thulasi, et.al [23]	Aspect polarity recognition; POS tagging is done using TnT tagger	Malayalam movie and product reviews	Sentence level aspect based; trigram model, TnT tagger using Viterbi algorithm for second order Markov models; 84.7% accuracy
2016	Dhanaraj V, et al [24]	Sentiment Analysis using YamCha and Fuzzy Logic	Malayalam movie reviews	A hybrid method consisting of machine learning and fuzzy rules; YamCha is used as the machine learning tool; 93.1 % precision
2015	M. Anagha et al. [25]	Hybrid Approach Based on Maximum Entropy Classifier	Movie Reviews	Polarity of opinion words in the input text with the help of Hindi WordNet - based lexical resource file created.
2015	Deepu S. Nair, et.al [26]	Hybrid approach	Movie Reviews	Sentence level; comprising of machine learning techniques and rule based approach; 91%

				accuracy
2015	Anagha, M. et al. [27]	Fuzzy logic based hybrid approach	Malayalam film reviews	Sentence level; Machine Learning is used for tagging and Fuzzy Logic for sentiment classification; 91% accuracy
2015	Jayan, P.,et.al [28]	Subjective feature extraction for sentiment analysis	Malayalam film reviews	Machine learning techniques CRF combined with a rule based approach; 82% accuracy
2014	Anagha, M.,et,al [29]	Cross domain sentiment analysis	Multi domain reviews	Hybrid approach; TnT tagger for tagging; Rule based approach for Classification; 93.6 % accuracy
2014	Deepu S. Nair, et.al [30]	Sentima	Movie Reviews	Sentence level ; 85% accuracy
2014	Indhuja K, et.al [31]	Fuzzy Logic Based	Product review documents, SFU corpus	extends the feature based classification with various linguistic hedges; uses fuzzy functions to emulate the effect of modifiers, concentrators and dilators; 85.58% accuracy
2014	Anagha, M. et al. [32]	Malayalam SentiWordnet		Lexical Resource for Cross Domain Sentiment Analysis and Opinion Mining
2012	Neethu, Mohandas,et.al [33]	Domain specific sentence level mood extraction	Malayalam Reviews	SO-PMI-IR formula classifies an input text into one of the two classes that indicate desirable or not desirable

Table 4: Main Themes used in Sentiment Analysis

Main Theme	**References**
Deep Learning	[21]
Fuzzy Logic	[19], [24], [27], [31]
SVM Classifier	[20],[24]
Lexicon Based	[22]
Markov Model	[23]
Max. Entropy	[25]
Hybrid - Machine Learning and Rule based	[21], [23], [24], [25], [26], [27], [28], [29]

Rule Based Approach	[30],[32]

5. Conclusion

The article focuses on the review of the various methods developed for POS Tagging, Stemming and Opinion Mining in Malayalam language. The main themes and algorithms used in the major contributions in the area of 'Sentiment Analysis of Malayalam language' as described in Table 3 is depicted in Table 4. The table shows that major contributions in Sentiment Analysis is using a hybrid approach which makes use of Machine learning algorithms along with Rule based approach to dealt with the special nature and inflections of the Malayalam dialect. The survey clearly indicates that less works are being done in this area. The observation through the literature survey is that the accuracy rate of the existing methods is not excellent, so the improvement is required to make them more efficient. The article will be useful for the researchers to get an overview of the major contributions and the methodologies used for Sentiment Analysis in Malayalam dialect

References

[1]. Kumar, S., M. Anand Kumar and K.P. Soman. "Deep Learning Based Part-of-Speech Tagging for Malayalam Twitter Data (Special Issue: Deep Learning Techniques for Natural Language Processing)" *Journal of Intelligent Systems, 0.0 (2018)*

[2]. Ajees, A. P., and Sumam Mary Idicula. "A POS Tagger for Malayalam using Conditional Random Fields.", *International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018)*

[3]. Kumar, S. S., M. Anand Kumar, and K. P. Soman. "Experimental analysis of Malayalam POS tagger using Epic framework in Scala." *ARPN J. Eng. Appl. Sci 11 (2016).*

[4]. D. Muhammad Noorul Mubarak, Sareesh Madhu, S A Shanavas, "A New Approach To Parts Of Speech Tagging In Malayalam" , *International Journal Of Computer Science & Information Technology (Ijcsit) Vol 7, No 5, October 2015*

[5]. Rinku T S, Merlin Rajan, Varunakshi Bhojane , "Various Approaches Used for Tagging and Chunking in Malayalam", *International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May -2014, ISSN 2229-5518*

[6]. Robert Jesuraj and P. C. Reghu Raj, "MBLP approach applied to POS tagging in Malayalam Language", *Proceedings of NCILC, pp. 5-8, CUSAT, Cochin, 2013*

[7]. Jisha P Jayan, Rajeev R R, "Parts Of Speech Tagger and Chunker for Malayalam – Statistical Approach", *Computer Engineering and Intelligent Systems, ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online), 2011*

[8]. P. J. Antony, Mohan, S. P., and Dr. Soman K. P., "SVM Based Part of Speech Tagger for Malayalam", in *Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conference on, Kochi, Kerala, 2010, pp. 339-341*

[9]. Rajeev, R. R., Jisha P. Jayan, and Elizabeth Serly. "Tagging Malayalam Text with Parts of Speech-TnT and SVM Tagger Comparison*." (2010).*

[10]. M. K., S. S. and S. M. Idicula, "Development of a POS Tagger for Malayalam - An Experience," 2009 *International Conference on Advances in Recent Technologies in Communication and Computing, Kottayam, Kerala, 2009, pp. 709-713.*

[11]. Dhanya, P. M., A. Sreekumar, and M. Jathavedan. "Vriksh: A Tree based Malayalam lemmatizer using Suffix Replacement dictionary." *International Journal of Emerging Technologies in Engineering Research 6.1 (2018): 31-42.*

[12]. C. Balasankar, T. Sobha and C. Manusankar, "Multi-level inflection handling stemmer using iterative suffix-stripping for Malayalam language," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 530-534.*

[13]. Nisha M, Reji Rahmath K, P. C. Reghu Raj, "A Machine Learning Approach for Malayalam Root Word Identification", in *proceedings of NC CLAIR-2014. pp. 1-4*

[14]. Prajitha, U., C. Sreejith, and PC Reghu Raj. "LALITHA: A light weight Malayalam stemmer using suffix stripping method." *Control Communication and Computing (ICCC), 2013 International Conference on. IEEE, 2013.*

[15]. Vasudevan, N., and Pushpak Bhattacharyya. "Optimal stem identification in presence of suffix list." *International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, 2012.*

[16]. Pragisha, K., and P. C. Reghuraj. "STHREE: Stemmer for Malayalam using three pass algorithm." *Control* Communication and Computing (ICCC), 2013 International Conference on. IEEE, 2013.

[17]. Meera Subhash, M. Wilscy, and S. A. Shanavas. A rule based approach for root word identification in Malayalam language. *International Journal of Computer Science & Information Technology, 4(3), 2012.*

[18]. Vasudevan, N., and Pushpak Bhattacharyya. "Optimal stem identification in presence of suffix list." *International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, 2012.*

[19]. Kasthoori V., Soniya B., Jayan V. (2019) Domain-Independent Sentiment Analysis in Malayalam. In: Verma N., Ghosh A. (eds) *Computational Intelligence: Theories, Applications and Future Directions - Volume II. Advances in Intelligent Systems and Computing, vol 799. Springer, Singapore*

[20]. R. Jayakrishnan, G. N. Gopal and M. S. Santhikrishna, "Multi-Class Emotion Detection and Annotation in Malayalam Novels," *2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2018, pp. 1-5.*

[21]. Kumar, S. Sachin, M. Anand Kumar, and K. P. Soman. "Sentiment Analysis of Tweets in Malayalam Using Long Short-Term Memory Units and Convolutional Neural Nets." *International Conference on Mining Intelligence and Knowledge Exploration. Springer, Cham, 2017.*

[22]. M. P. Ashna, Ancy K Sunny, "Lexicon based sentiment analysis system for malayalam language", *Computing Methodologies and Communication (ICCMC) 2017 International Conference on, pp. 777-783, 2017.*

[23]. P.K. Thulasi, K. Usha, "Aspect polarity recognition of movie and product reviews in Malayalam", *Next Generation Intelligent Systems (ICNGIS) International Conference on, 2016.*

[24]. Dhanaraj V, Reshma R, Sreetha S and Binu R "Sentiment Analysis for Malayalam movies reviews using YamCha and Fuzzy Logic " in *NCILC 2016 conference held at CUSAT , pp. 1-4.*

[25]. Anagha M, Raveena R Kumar, Sreetha K, P C Reghu Raj, “A Novel Hybrid Approach Based on Maximum Entropy Classifier for Sentiment Analysis of Malayalam Movie Reviews”, *International Journal of Scientific Research, Vol : 4, Issue : 6 June 2015*

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

[27]. Anagha, M. et al.: Fuzzy logic based hybrid approach for sentiment analysis of Malayalam movie reviews. In: *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015. IEEE (2015)*

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

[29]. Anagha, M., et al. "LEXICAL RESOURCE BASED HYBRID APPROACH FOR CROSS DOMAIN SENTIMENT ANALYSIS IN MALAYALAM." (*2014).*

[30]. Deepu S. Nair, Jisha P. Jayan, Rajeev R. R, Elizabeth Sherly, "SentiMa - Sentiment extraction for Malayalam", Advances in Computing Communications and Informatics (ICACCI 2014 International Conference on, pp. 1719-1723, 2014.

[31]. Indhuja K, P C Reghu Raj "Fuzzy Logic Based Sentiment Analysis of Product Review Documents", published in *ICCSC at LBS Technologies, December 2014*

[32]. Anagha M, Raveena R Kumar, Sreetha K and Naseer C, "Malayalam SentiWordnet: A Lexical Resource for Cross Domain Sentiment Analysis and Opinion Mining", in *proceedings of NC CLAIR-2014. pp. 21-23*

[33]. Neethu, Mohandas, J. P. S. Nair, and V. Govindaru. "Domain specific sentence level mood extraction from malayalam text." *Advances in Computing and Communications (2012).*

[34]. Deepa Mary Mathews, Sajimon Abraham "Effects of Pre-processing Phases in Sentiment Analysis for Malayalam Language." *International Journal of Computer Sciences and Engineering 6.7 (2018): 361-366.*