JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

Social Media Sentiment Analysis For Malayalam

M. Rahul[1*], R.R. Rajeev[2], S. Shine [3]

[1]Computer Science and Engineering, Govt. Engineering College, A P J Abdul Kalam Technological University, Palakkad
[2]e-Governance and Development, International Centre for Free and Open Source Software, Technopark, Trivandrum
[3]Computer Science and Engineering, Govt. Engineering College, A P J Abdul Kalam Technological University, Palakkad

Corresponding Author: rahulgullan9@gmail.com, Tel.: +91-9544434428

Available online at: www.ijcseonline.org

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 reviews, political discussions etc. The user generated text collected from social media can help machines to summarize and take intelligent decisions in different domains. Sentiment analysis in Malayalam language has a large importance. Malayalam is a low-resource language and it does not possess a standard corpus or a sentiment lexicon. This work presents a machine learning approach to sentiment analysis in Malayalam language using the CRF and SVM. The learning carried out at two levels and the system classify sentences into positive, negative and neutral 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, CRF, SVM, NLP

I. INTRODUCTION

Emergence of social media in the last decade changed the way of expressing and sharing human emotions, attitudes and opinions to public. The rapid growth in information and communication technology encouraged people to participate and feel involved in what happening in the world. The reachability of this information across the world greatly influence the business, politics, education, culture, careers, innovation, share market, medicine, engineering, disaster management etc. Listening and analyzing social media posts helps to extract valuable information. Using this information people can take intelligent decisions in many fields. Sentiment analysis or opinion mining in social media determine the emotions of public opinion towards events or entities. More formally, Sentiment Analysis can be defined as a Natural Language Processing (NLP) which includes methods to identify the subjectivity such as emotions and opinions hidden in the user generated text [1]. The very basic approach is to classify the text into positive, negative and neutral classes. More advanced methods involve classification of texts into multiple target classes along with intensity of opinion associated with it. Nowadays people use social media platforms to express their opinions on events, entities, films etc. Sentiment analysis in social media is an important application because it helps to access public opinion on relevant topics by analyzing huge volume of user generated text. Researches in Sentiment analysis increased in recent years and it is a very exciting problem today.

Texts in the social media are very informal and very difficult to process. This natural language text is very difficult for a machine to understand. Sarcasm, contextual meaning, slang etc. will make the text even more difficult to process. The methods and tools for sentiment analysis are very advanced in European, Latin and Chinese language, also these languages possess large number of resources and tools available to perform various NLP applications. It is obvious that social media users mostly prefer their native languages to communicate and drop social media posts. Since most of the NLP tools and resources available today are highly language specific, the same kind should be developed for native languages also or else it would not possible to effectively process native languages. Sentiment analysis in native languages is very rare due to scarcity of resources. Malayalam is such a low-resourced language. No standard corpus and sentiment lexicons are available for performing sentiment analysis. Moreover, this language belongs to Dravidian family and large number of morphological inflections and agglutination make it very complex [2].

Based on the unit of text used for classification sentiment analysis can be divided into document level, sentence level and attribute level [3]. Document level sentiment analysis classify the whole document into corresponding sentiment classes. The challenge is that the document may contain many contradictory sentences. That makes, finding overall polarity becomes difficult. Sentence level sentiment analysis depends on semantic orientation of context dependent words and classify the whole sentence into corresponding classes. Attribute level or aspect level sentiment analysis provides sentiment towards each individual aspects. That is people may give different opinion for aspects or features of a same entity. Social media posts are more often confined to single sentence or length of posts will be comparatively small. So it can be considered as a single sentence. So this paper implement sentiment analysis at sentence level.

The approach existing to solve this text classification problem are categorized into three. Which are machine learning approach, lexicon based approach and hybrid approach [4, 5]. Machine Learning approach again classified into two supervised and unsupervised. Supervised learning requires a training set which then used by different machine learning algorithms to build a classifier. This classifier will classify unseen data instances into corresponding target classes. Efficiency of the classifier is then evaluated and validated by a test set. Unsupervised learning used when labeled training data is not available. Lexicon based approach calculate sentiment polarity using semantic orientation of words or sentences with a known set and it does not require any learning process. The known set is called sentiment lexicon which contain precompiled sentiment terms. This method also divided into two, dictionary based approach and corpus based approach based on the resource in which related opinion words are searched. The challenge of lexicon based approach is that it requires resources which need to almost cover the entire target language. Obtaining such resources are expensive and creating such resources are time consuming. Hybrid approach combines both of them. The efficiency is large compared to the other two methods if used them independently. Another rarely used one is the rule based approach [6]. It identifies sentiment words in the text then based on some well-defined heuristics, classify text into corresponding sentiment classes. But compared to other approaches rule based system has poor generalization capability. This method opted if the domain is comparatively narrow. More sophisticated deep learning methods are getting wide attention today [7]. It because deep learning approach does not requires feature engineering and time consuming manual annotation of data [8]. But this method requires huge amount of data and it is very expensive to train sometimes requires more sophisticated machines. Training of models may take several weeks and demand expensive GPUs.

The present paper address sentiment analysis in Malayalam language using machine learning approach. Sentence is taken as basic unit of text for classification. That is sentence level sentiment analysis. The proposed method classify sentences into positive, negative and neutral classes. Learning is carried out at two levels. Conditional Random Field (CRF) is trained on word level to semantically label Individual words as POS (positive), NEG (negative), 0(neutral) and INT (intensifiers). Support Vector Machine (SVM) is trained on sentences which are annotated both at word level and sentence level and classify the sentences into positive, negative and neutral based on 14 contextual features extracted [9].

Rest of the paper is organized as follows, Section I contains the introduction, Section II contains the related works in Malayalam, Section III explains the dataset, Section IV contains the architecture, Section V explains the methodology, Section VI describes results and discussion and Section VII concludes research work with future directions.

II. RELATED WORK

Works on Sentiment Analysis in Malayalam language is comparatively less. It is due to weaker advancement of NLP applications in such a low resourced language. The following are the related works on sentiment analysis in low-resourced languages. Almost every works begin with the creation of manually annotated corpus followed by implementation of classification methods.

The paper by Deepu et.al [10] 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.

Another paper from the same author, Deepu et.al [11] discussing a complete rule based approach to sentiment analysis. Here the author uses pre-stored positive and negative words to tag the words in the sentence. If the word does not found in the set, assign it a neutral value. The polarity of the entire sentence then determined using predefined set of rules. The negation words are handled separately by the negation rules. Here dataset is created from film reviews by collecting Malayalam text from film review sites.

M. Neethu et.al [12] 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.

Machine learning approach, SVM used in B. Jasmine et.al [13] to address sentiment classification of product reviews. The highlight is, classification is enhanced by effectively processing objective words, negation words and intensifiers. Because presence of words from these category largely influence the sentiment words adjacent to it. While calculating score of sentiment words which are related to above mentioned POS categories, treated in different manner to assign scores based on well-defined scoring convention. So that sentiment scores intensified or diminished based on new scoring convention. Training set consists about 44000 sentences.

S. Yakshi et.al [14] proposed an unsupervised lexicon based approach to sentiment analysis. They created subjective lexicon specifically for the tweets in the target language. The polarity of the tweets are determined with the help of subjective lexicon using a predefined scoring method. Tweets are then classified into different levels of polarities like positive, extremely positive, negative, extremely negative and neutral.

III. DATASET

Statistical methods requires language resources readily available in parallel. Which is the starting point of any NLP applications and it decides the quality of work. The drawback mainly exist in the development of NLP applications in Malayalam is the unavailability of such a resources. So it is necessary to create a corpus as the primary task. Here we collected Malayalam text from various social media platforms. The collected plain texts are then refined before the manual annotation. The dataset annotated at two levels. Initially words are tagged with BIS Malayalam part of speech tag set. Sentences are analyzed and annotated with target classes like positive, negative and neutral. At word level, sentiment words are identified in sentences and annotated with semantic tags such as 'POS' for positive, 'NEG' for negative, '0' for neutral and 'INT' for intensifiers. The dataset consists 1286 sentences. The annotated corpus is used as a training set for both CRF and SVM models. The sentence given below is an instance of dataset used for training.

‘’ഇന്നലെ/N_NN/0 ഒഴിവ്/N_NN/0 ആയത്/V_VM_VNF/0 വളരെ/RP_INTF/INTവലിയ/JJ/INT അപകടം/N_NN/NEG ആയിരുന്നു/V_VM_VF ./RD_PUNC/#NEG’’

IV. ARCHITECTURE

The figure 1 shows the architecture of the proposed model. The entire model can be divided into two levels based on the unit of text learned by machine learning algorithms for classification. The first learning model trained at word level and used for semantic tagging. CRF learning is used in this level. Sentence level classification is performed at second level using SVM learning. While creating the dataset for training, each sentences are tagged in two manner. Sentiment words in the sentence are identified and tagged with well-defined semantic tags. Which gives the contextual information of the individual words. The whole sentence is then tagged with corresponding polarity class. That is negative, positive or neutral. This annotation helps to keep same data set for both the levels of learning. For CRF learning, at word level, part of speech is taken as the feature. For SVM learning, sentence level polarity is taken and contextual features are extracted to train the classifier.

If we give a sentence as input, it will pass through two levels of learning. The first one predict the semantic tags of the individual words. Using this word level contextual information the sentence is then classified into its corresponding target class by the SVM model.

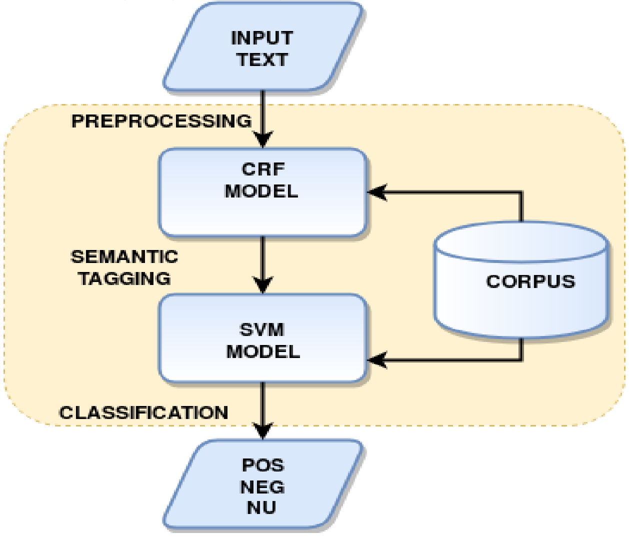

Fig1. Architecture of proposed system

V. METHODOLOGY

General procedures of sentiment analysis can be organized as data collection, text preparation, sentiment identification, sentiment classification and visualization. Data collection involves obtaining raw data from social media sites, blogs or community forums. But these user generated text is highly unstructured. It may contain non-Malayalam texts, words that never belongs to Malayalam vocabulary, influence of slang etc. These issues are managed in the text preparation step. Malayalam text is collected from different online sources. The collected texts are then refined in the text preparation step using basic programming in python and regular expressions. Which include removal of unwanted punctuations, non-Malayalam characters, numbers etc. Word completion and spelling error corrections are done manually. Textual contents which are irrelevant to sentiment analysis are also removed. Sentiment identification steps involves determining subjectivity of sentences. The subjective words are to be then tagged with appropriate semantic tags. In the sentiment classification step the sentences are to be classified into corresponding target classes like positive, negative and neutral. The final step is the visualization of classification model in the form of charts or graphs along with various quality measures like accuracy, F-measure, precision, recall etc. The numerical value of classification can be represented in the form confusion matrix or tables so that it becomes very intuitive.

Machine Learning algorithms are used for sentiment identification and sentiment classification. The methodology consists two levels of learning. It make use of word level semantic labelling followed by the sentence level classification. Word level tagging can be considered as sequence labelling problem and sentence level classification is a text classification problem. CRF is a popular supervised learning algorithm well suited for sequence labelling problems. The SVM is efficient for text classification problem. Here we using CRF for word level tagging and SVM for sentence level classification.

CRF is extensively used in many NLP applications in which word sequence or sentence sequence has large influence in predicting labels. That is the current word or sentence sometimes depends on surrounding sequence of words. It helps to capture dependencies among the observations (words). So this learning method is widely used in applications like Part of Speech (POS) tagging, Named Entity Recognition (NER), etc. The CRF can be formally stated as an undirected graphical models [15] which is trained to maximize a conditional probability. A liner chain CRF looks similar to a finite state machine used for sequence labelling.

The strength of sentiment in the text mainly determined by certain part of speech categories. So part of speech of individual words in the sentences are considered as a feature. Also three semantic tags are defined. 'POS' for positive sentiment, 'NEG' for negative sentiment, '0' for neutral and finally 'INT' for intensifiers. Some special words in the sentences which come under the category of adjectives or adverbs will enhance the subjectivity of nearby words or intensify the sentiment of nearby words [16]. When an intensifier followed by a positive word its positivity will increase. As well as when an intensifier followed by a negative word its negativity will increase. Usage of intensifiers are very common in Malayalam posts. Identifying intensifiers are hence very important in determining the overall polarity of the sentence.

വളരെ(Very) മനോഹരമായ(beautiful)
ചിത്രം(film).

The given sentence in Malayalam language contains an intensifier 'വളരെ (very)' which intensified the word 'മനോഹരമായ (beautiful)' to make the word more positive. In effect the sentence became more positive. To address such a sequence label problem CRF classifier is trained on manually labelled dataset.

The SVM provides better model compared to Maximum entropy or CRF for sentiment analysis. It is a non-probabilistic supervised learning method well suited for text classification problems. It is why because SVM can handle large number of features. More specifically very large dimensional input space. Normally text features will be very large in number it may sometimes exceeds more than ten thousands of features. SVM finds linear separators in the higher dimensional space efficiently for text classification problems. Previous works in sentiment analysis pointed out that SVM can provide better result compared to other supervised learning methods. Sentence level sentiment analysis can be efficiently handled by SVM.

A number of features are to be extracted from the training data for training the classifier. We choose contextually relevant fourteen number of features to train the classifier. These features include number information of sentiment bearing words and information about their relative positions. A linear classifier built on the extracted features.

The table given below describes features extracted for training. We calculated numerical values of these features from training data.

Table 1. Features selected for SVM learning

Features	Description
n_pos	Number of positive words
n_neg	Number of negative words
n_int	Number of intensifiers
n_int_pos	Number of intensifiers followed by positive words
n_int_neg	Number of intensifiers followed by negative words
n_int_int	Number of intensifiers followed by intensifiers

Features	Description
n_pos	Number of positive words
n_neg	Number of negative words
f_pol	First polarity of sentence
l_pol	Last polarity of sentence
c_a_pos	Number of positive terms after conjunction
c_b_pos	Number of positive terms before conjunction
c_a_neg	Number of negative words after conjunction
c_b_neg	Number of negative words before conjunction
c_a_int	Number of intensifiers after conjunction
c_b_int	Number of intensifiers before conjunction

VI. RESULTS AND ANALYSIS

Both the model trained on the dataset with 1286 sentences and approximately 1635696 tokens. The test set consists of 373 sentences. The training set is used for both word level and sentence level training. Various classification parameters are measured for both the levels and the overall system.

The table 2 given below shows the recall, precision and f-measure result for word level semantic labelling. The performance of this classifier can be increased by adding more sentences to the dataset. It is required to extract more contextual features, it because accuracy of semantic tagging is very important for the sentence level processing. This is because, features are to be extracted from first level output for prediction by the SVM. The challenge in the word level tagging is, addressing large number of morphological inflections. The dataset should contain all possible inflections and agglutination. That will helps to improve the accuracy of prediction at the lexical level.

Table 2. Precision, recall and F-measure of word level classification

	Precision	Recall	F-measure
0	0.5392	0.6572	0.5923
INT	0.6667	0.5452	0.5998
NEG	0.5768	0.6618	0.6163
POS	0.5833	0.6572	0.6180

The table 3 given below shows the recall, precision and f-measure result for sentence level classification. The model obtained comparatively large F-measure value. It indicates the relevance of features, selected for SVM training, best decide the sentence level polarity.

Table 3. Precision, recall and F-measure of sentence level classification

	Precision	Recall	F-measure
POS	0.774194	0.806723	0.790123
NEG	0.798701	0.836735	0.817276
NU	0.842593	0.758333	0.798246

The table 4 given below shows the recall, precision and f-measure result for the overall system. The test set consists 373 sentences, equal number of sentences in each class.

Table 4. Precision, recall and F-measure of overall classification system

	Precision	Recall	F-measure
POS	0.4814	0.5421	0.5099
NEG	0.5063	0.5580	0.5308
NU	0.6757	0.5316	0.5950

The F-measure value is reduced here, it because two classifiers are work in a sequential manner. So performance of individual classifiers highly influence the overall performance.

VII. CONCLUSION

Sentiment analysis is considered to be a text classification problem. It has a large importance in many fields like social media analytics, recommendation systems, electronic commerce etc. The proposed system classify user generated Malayalam texts in the social media into different sentiment classes. Here sentiment analysis is performed at sentence level using supervised learning methods, SVM and CRF. The classification problem solved at two levels. The word level semantic tagging followed by the sentence level sentiment classification. The word level semantic labelling addressed using the CRF learning. Contextual features are extracted and trained using the SVM for sentence level classification.

The accuracy obtained for the entire system is 52.75 percentage for the training set consists of 1286 sentences. The accuracy of individual classifiers greatly influence the performance of overall system. It is due the fact that, two classifiers work in a sequential manner. To obtain better accuracy it is required to populate the dataset with more sentences (dataset should contain maximum number of words from Malayalam vocabulary). Since there is no efficient stemmer in Malayalam language, it is required that the dataset should contain all the inflected forms of words and all possible agglutinated words. In future, classification can be enhanced by adding more contextual features at both the levels of the learning.

REFERENCES

[1] Kaur, Amandeep, V. Gupta, "*A survey on sentiment analysis and opinion mining techniques"*, Journal of Emerging Technologies in Web Intelligence, Vol.5, Issue.4, pp.367-371, 2013.

[2] V. Jayan, V. K. Bhadran, *"Difficulties in processing malayalam verbs for statistical machine translation"*, International Journal of Artificial Intelligence and Applications (IJAIA), Volume.6, Issue.3, 2015.

[3] Liu, Bing, "*Sentiment analysis and opinion mining"*, Synthesis lectures on human language technologies, Vol.5, Issue.1, pp.1-167, 2012.

[4] Medhat, Walaa, A. Hassan, H. Korashy, "*Sentiment analysis algorithms and applications: A survey"*, Ain Shams Engineering Journal, Volume.5, Issue.4, pp.1093-1113, 2014.

[5] Jain, P. Anuja, P. Dandannavar, "*Application of machine learning techniques to sentiment analysis*", Applied and Theoretical Computing and Communication Technology (iCATccT), 2nd International Conference on. IEEE, 2016.

[6] Chikersal, Prerna, P.Soujanya, E. Cambria, "*SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning*", In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp.647-651, 2015.

[7] Kumar, S. Sachin, M.A. Kumar, 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.

[8] Ain, T. Qurat, M. Ali, A. Riaz, A. Noureen, M. Kamran, H. Babar, A. Rehman, *"Sentiment analysis using deep learning techniques: a review"*, Int J Adv Comput Sci Appl 8, 2017.

[9] Zhang, Kunpeng, Y. Xie, Y. Cheng, H. Daniel, D. Downey, A. Agrawal, L. Wei-keng, A. Choudhary, "*Sentiment identification by incorporating syntax, semantics and context information*", In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, pp.1143-1144, 2012.

[10] 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.

[11] Nair, S. Deepu, J.P. Jayan, R.R. Rajeev, E. Sherly, "*SentiMa-sentiment extraction for Malayalam*", Advances in Computing, Communications and Informatics (ICACCI), International Conference on. IEEE, 2014.

[12] 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.

[13] K. Bhaskar, Jasmine, Sruthi, P. Nedungadi, "*Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers*", Recent Advances and Innovations in Engineering (ICRAIE), IEEE, 2014.

[14] Sharma, Yakshi, V. Mangat, K. Mandeep, "*A practical approach to Sentiment Analysis of Hindi tweets*", Next Generation Computing Technologies (NGCT), 1st International Conference on. IEEE, 2015.

[15] Sutton, Charles, A. McCallum, "*An introduction to conditional random fields*", Foundations and Trends® in Machine Learning, Vol.4, Issue.4, pp.267-373, 2012.

[16] Benamara, Farah, C. Carmine, P. Antonio, R.R. Diego, S. Venkatramana, Subrahmanian, "*Sentiment analysis: Adjectives and adverbs are better than adjectives alone*", In ICWSM, 2007.

Authors Profile

Mr. Rahul M pursued Master of Technlogogy in the field of Computer Science and engineering with specialization in Computational Linguistics from GEC, Palakkad, A P J Abdul Kalam Technological University, Kerala, India in year 2018. His area of interests are Natural Language Processing, Machine Learning and Computational linguistics.

Dr. Rajeev R.R pursed Master of Science in 2003, Master of Philosophy in 2006 and Doctor of Philosophy in 2015 from University of Kerala, India. He is currently working as a Programme Head at ICFOSS Kerala, India. He published more than 15 research papers in reputed international journals. His research work focuses on Machine Learning, Computational Linguistics and Natural Language Processing. He has more than 4 years of teaching experience in various reputed institutions like IIITM-K.

Mr. Shine S pursued Bachelor of Technology in Computer Science from University of Kerala, India in the year 1998 and Master of Technology from NIT Calicut in the year 2009. He has more than 10 years of teaching experience in various engineering colleges.