IMPROVED SENTIMENT ANALYSIS FOR DRAVIDIAN LANGUAGE-KANNADA USING DECISION TREE ALGORITHM WITH EFFICIENT DATA DICTIONARY

Ranjitha P¹ and Bhanu K N²

Department of computer science, Amrita Vishwa Vidyapeetham, Mysuru, India
¹ ranjitha_p@asas.mysore.amrita.edu
² knbhanu@asas.mysore.amrita.edu

Abstract. The trend of exploring the micro blogging services is being hit door to door. The creation and sharing of short text messages is at its top gear. Reviews, being one of the fastest growing online networking services, which receives around millions of reviews per day, has emerged as an invaluable source of the sentiment analysis. Sentiment analysis refers to recognize the sentiment, polarity, opinion or emotion of a text. Websites like news Kannada, prajavani, wedunia.com, and oneindianews have reviews written in Kannada which has a higher preference than the common languages. In the current system, research are carried in sentiment analysis for regional language such as Hindi, Bengali, Kannada etc., are very few compared to English language. Here we propose a system which is computationally identifying and categorizing opinions expressed in a piece of text written in Kannada language, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Keywords: Natural language processing, Sentiment analysis, data dictionary, Feature extraction, Tokenizing, decision tree.

1 Introduction

In the current system, research works done in sentiment analysis for regional language such as Hindi, Bengali etc., are not too many when compared to English language. These languages were analyzed by equating lexicon words using part of speech tagging technique. This represents emotions as adjectives pattern suffixed word. These semantic words are rated through machine learning approach. Machine learning includes algorithms such as rule based algorithm, maximum entropy etc.
India being multi-linguistic nation sentiment analysis in English forms a disadvantages. For analyzing the documents or piece of text that are of particular regional language such as Kannada language, it required to have sentiment analysis by using datasets in Kannada. Websites like news Kannada, prajavani, wedunia.com, and oneindianews have reviews in Kannada which has a higher landing than the ordinary languages. Even though Sentiment analysis in Kannada with English translation produces similar results with generic set of sentiment words but does not consider the domain of the text which is the feature of an entity called subjective in sentiment analysis. So our proposed model uses the Kannada words without translating to English and trains the model using efficient data dictionary by decision tree classifier. Here Data dictionary is the one where we store set of words which is labelled as positive/negative words and each word is assigned with certain score (polarity).

2. Related Work

Sentiment analysis is a current trending topic for the research. There are many unsupervised algorithm which is used to classify a review as recommended or not recommended. As an initial step, the text data is POS tagged in order to identify the adjectives, noun and adverbs and then the semantic orientation of these words are carried out. Here PMI-IR algorithm is used for calculation of semantic orientation between words. [2]

A novel approach for machine translation of simple sentences from Kannada to English language is introduced. In the first step method execute lexical analysis to process input transliterated Kannada text to recognize and to classify tokens using newly constructed source language lexicon KannadaWordNet. Then, newly designed recursive descent parser for the domain validates the transliterated Kannada text with the syntactic constructs/grammar rules. [3]

Tools like Paraphrase generator are very essential component of any NLP application. Paraphrase generator tool for south Indian language of Kannada language using dictionary lookup approach. By using the morphological analyzer, dictionary look up and morphological generator approach, the Paraphrase are generated for Kannada language. [5]

Polysemy word are the word which has multiple meanings. The same word can be treated differently at different instance. Consider a text from Kannada paragraph, a person with no prior knowledge of Kannada language cannot distinguish the difference in the meaning of a polysemy word. It can be concluded that in a huge document with multiple polysemy words cannot be summarized very easily. [6]
In Hindi language sentiment lexicon called Hindi-Senti Word Net (H-SWN) was introduced using English Senti Word Net (SWN) and English- Hindi Word Net linking. The sentiment analysis of documents was analyzed using sentiment annotated corpora, Machine translation and resource based sentiment analysis methods. [8]

Sentiment Analysis in Indian Languages is a trending topic for research area. Word Net, POS tagger, Senti Word Net and many other resources are available for very few languages. A Senti Word Net development for Indian Languages like Bengali, Hindi and Telugu is tried using Word Net, dictionary based and Corpus based methods in a Hindi sentiment lexicon called Hindi-SentiWordNet (H-SWN) was developed using English SentiWordNet (SWN) and English- Hindi WordNet linking. The sentiment analysis of documents was performed using sentiment annotated corpora, Machine translation and resource based sentiment analysis methods. [9]

3 Methodology

Fig. 1. Workflow of the proposed work

Sentiment analysis follows the steps as shown above in Fig 1. Data is collected from various websites like prajavani, wedunia.com, and oneindianews is considered as the dataset for training. Collected dataset is first subjected to the pre-
processing. Pre-processed data is taken for data dictionary generation. Fig. shows the data dictionary of the proposed system. Once the data dictionary is ready, it is trained by a decision tree. Classification requires fore-going learned labeled data in a particular form known as trained data, which are attributes and its values along with target function called Class Label. Text is tokenized during pre-processing of data set and attaches tags to each tokenized words using part of speech taggers [1]. These tagged words are identified as features or opinion words according to role in their text. Decision tree algorithm for Kannada sentiment analysis is used to train and test the data and classify into different labels. Decision tree algorithm for Kannada sentiment analysis is explained in the algorithm 1.

Algorithm 1: Decision tree algorithm for Kannada sentiment analysis

**Input:** Randomly chosen Kannada review over internet.
**Output:** classifies the review into positive / negative review.

Let R be the review
S be the set of words.
\[ S = \{ w_1, w_2, w_3 \ldots \ldots \ldots \ldots w_n \} \]
\[ P_c = \text{Positive counter}, N_c = \text{Negative counter}, P_v = \text{Polarity values}, P_{dd} = \text{Positive word} \]
In the data dictionary, \( N_{dd} = \text{Negative word in data dictionary}, V_p = \text{Polarity value of Words}, P_r = \text{Positive}, R_v = \text{Negative Review} \).

1: Tokenize R
   Set \( P_c \) to zero and \( N_c \) to zero.
2: **For each** \( W \) in \( S \)
3: **If** \( W \) is found in \( P_{dd} \)
   Increment \( P_c \) to value of \( V_p \)
4: **Else** \( W \) is found in \( N_{dd} \)
   Increment \( N_c \) to value of \( V_p \)
**End If**
**End For**
5: **If** \( P_c > N_c \)
   Return \( P_r \)
**Else** return \( N_r \)
**End If**
Figure 2 Steps to create data dictionary.

Fig 2 shows the steps to create the data dictionary for Kannada sentiment analysis. In the first step reviews are randomly chosen from the internet. Chosen review subjected to the tokenization. Later each word is labelled as positive or negative and assigned polarity values for each word and stored it in the data dictionary. Table 1 shows the sample data dictionary where each word as word type and polarity score.

Table 1. Sample data dictionary words

| Kannada words | Word Type | Polarity score |
|---------------|-----------|----------------|
| ದುಂಡಾಳಿ      | Negative  | 2              |
| ಪ್ರತಾಪಾಳಿ    | Positive  | 3              |
| ಹೊಸ್ಟ್ನಗಿ     | Negative  | 3              |
| ಹೆಜಿ     | Positive  | 2              |
| ಹಾರಾತ್ತಕ    | Positive  | 4              |
4 Results and discussion

Figure. 3 Adding the word to data dictionary

Fig 3 shows the creation of data dictionary manually. Here we have created data dictionary which consists of more than 500 words for each positive and negative category. Each word will be given one score.

Figure. 4 shows the list of words in dictionary
Fig 5 shows the working of the proposed system. When the test data is given the trained system or model. First it will list number of positive word and negative word in the given text data and later it will classify the text into positive or negative. The trained model is providing up to 85% of accuracy.

Fig 6 showing the test data is classified to positive
Table 2. Comparison on different algorithms.

| Methods                               | Accuracy |
|---------------------------------------|----------|
| Naïve bayes                           | 65%      |
| Negator Algorithm                     | 53%      |
| Decision tree using data dictionary   | 85%      |

Table 2 shows the result by comparing different algorithm. By translating kannada to english using Naïve bayes they have achieved 65% of accuracy[11]. Using Negator algorithm for kannada sentiment analysis [11]resulted in 53% of accuracy. Finally our model resulted in the accuracy of 85%. By comparing this three methods for sentiment analysis decision tree algorithm is giving more accurate result.

Table 3 Comparison on Precision and Recall

| Method                               | Precision | Recall |
|--------------------------------------|-----------|--------|
| Decision tree                        | 0.78[1]   | 0.79[1] |
| Decision tree with efficient data dictionary | 0.85    | 0.84   |

Table 3 shows the comparison on precision and recall between two methods. Using decision tree results in precision of 0.78 and recall of 0.79[1]. By manipulating decision tree with efficient data dictionary we can achieve 0.85 precision and 0.84 recall.

5 Conclusion and future work

Features based opinion mining is a challenging task in Kannada language and there are only few tools available such as Training data set specific to a domain, part of speech Tagger. Certain words in Kannada like “ಕಥೆಗುಂದಾದ,” “ಎನುವದಕೊೂ,” “ಮನೆಯೊುಂದರಲ್ದ,” “ಚಚೆೆಯಾಯಾಗಟತವೆ” on machine translation produced ambiguous text, which forms one of the major drawbacks and result in improper results. Thus sentiment analysis in regional language gives better accurate results compared to Machine Translated English Language. Here by creating the efficient data dictionary Accuracy of 85% can be achieved. Only few analyses is done in Indian regional languages, hence
this work can be extended to other regional languages especially in south Indian languages for the enhanced version of sentimental analyzer.

References

[1] Rohini V, Merin Thomas and Dr. Latha C A 2016 Domain based Sentiment Analysis in Regional Language-Kannad *Int. J. of Engg. Research & Tech. (IJERT)* Vol 4 Issue 22
[2] .Hemalatha, Dr. Saradhi Varma G P and Dr. Govardhan A Sentiemnt analysis tool using machine learning algorithm *Int. J. of emerging Trends and Tech. in computer science* vol 2 Issue 2
[3] Peter Turney D 2002 Thumbs Up and Thumbs Down? Semantic orientation applied to Unsupervised Classification of Reviews”,Institute for Information Technology, National Research Council of Canada Ottawa,Proceeding of Soth Annual Meeting of the Association for computational Linguistics Philadelphia
[4] Solamki Yogesh Ganeshbhai,Bhumika and Shah K 2015 Feature based opinion mining:A Survey Surat India
[5] Aditya Joshi, Balamurali A R and Pushpak Bhattacharyya, 2010 A Fall-back Strategy for Sentiment Analysis in hindi: a Case Study”,IIT Bombay.Proceedings of ICON 2010: 8th *Int. Conf. on Natural Language Processing*
[6] Anil Kumar K M, Rajasimha N 2015 Analysis of user’s sentiment for kannada web documents”,Sciencedirect, Procedia computer science *Eleventh Int. Multi-Conf. on Information Processing* Vol 54 pp 247- 256
[7] Ankit singh and Md.Enayat Ullah Aspect based Sentiment Analysis ,Indian Institute, Kanpur CS365A http://goo.gl/qdTfKh.
[8] Jeevanandam Jotheswaran and Dr.Y.S.Kumaraswamy Opinion mining using decision tree based feature selection through manhattan hierarchical cluster *measure J. of theoretical and applied science* Vol 58
[9] Tien-thanh Vu,Huyen-Trang Pham, Cong-To Luu 2011A Feature-Based Opinion Mining Model on Product Reviews in Vietnamese *Vietname National University, Hanoi*(VNU), *College of technology,Semantic Methods SCI*381 pp.23-33,springerlink.com
[10] Prabhu Palanisamy, Vineet Yadav and Harsha Elchuri, Serendio:Simple and Practical Lexicon Based approach to Sentiment analysis Serendio Software Pvt Lid Guindy Chennai
[11] Remya Govindan1 and Rosna Haroon P 2017 A Survey on Sentiment and Emotion Classification in Indo-Dravidian Languages Imperial Journal of Interdisciplinary Research (IJIR) Vol 3 Issue.