

Corpora Based Classification to Perform Sentiment Analysis in Kannada Language

Shankar R, Suma Swamy



Abstract: In this modern era, the users' opinions play an uncanny role in understanding how well a product has satisfied the customer requirements, so that the producer can change the product to suit the customers' demands and these reviews also help the new consumers to decide on whether to purchase the product or not. Analysis of a particular entity's feelings in terms of positive, negative or neutral polarization is known as 'Sentiment Analysis'. Sentiment Analysis is a sub-domain of opinion mining. Here the analysis is focused on the mining of emotions and opinions of the people towards a specific topic. The emotions and opinions are collected in the form of organized, semi-organized or amorphous data. As the world is slowly progressing towards regional languages, this article talks about extracting the opinions of a product in Kannada and performing analysis about these reviews and classifying them accordingly. The dataset or the corpus is scarce as it is not English. The limited corpus is being collected via website – <https://gadgetloka.com> through an API. However, extracting inclusive opinion manually from huge amorphous data would be a tedious task. An automated system called 'Sentiment Analysis or Opinion Mining' can solve this problem, which can analyze and extract the observation of the user throughout the reviews. In this classifier of review analysis, the process classifies the review via corpus, which is a huge collection of pre-defined data. The API that has been used is Python-Beautiful Soup via utf-8 text recognition method to parse Kannada characters. The reviews are converted to text sentence and each word of the sentence are broken down. Data mining task is done to find the sentiment of each word by comparing it with two stored files named as good.txt and bad.txt. Further, the analyzed result is given through text output as Positive, Negative or Neutral sentiments based on their weights.

Keywords: classification, corpora, kannada lexicon, opinion mining in kannada.

I. INTRODUCTION

Sentiment Analysis (SA) is a prominent category in the line of textual analysis as the results have shown us that there exists a lot of scope for application and implementation. For example, one can Forecast Sentiment, detect partisanship, summarize text, break down and summarize Sentiment etc. In India, there exists plenty of regional languages like Hindi, Kannada, Telugu, Tamil, Malayalam etc. There exists many

studies and research interests in the area of SA in these diverse languages. However, SA of Kannada text has not been extensively explored, especially for the purpose of analysis of products. In this paper, a case study for mobile product reviews that have been entered in Kannada is proposed. This is viable because there exist many user-generated reviews of products in Kannada, available online. Sentiment Analysis (SA) is a methodology of computation that involves characterizing, identifying, and extracting contents with embedded sentiment, such as emotions, opinions and attitudes, subjective impressions in the text, speech or databases. This SA uses concepts from Natural Language Processing (NLP) and Machine Learning (ML) algorithms. In this paper, the corpus-based technique is used to construct a sentiment lexicon. It relies on morphological patterns present in large corpus and produces the relative words with a high accuracy. Sentiment analysis (SA) also can be observed as a manifestation of an NLP problem. However, while NLP requires understanding of the context in a large corpus, SA can utilize only a few key phrases from the text and yet provide us with adequate results. Lastly, while existing research points to certain applications like detection of reasoning, this alone may not be enough to cover most use cases. Hence, it is required to do much deeper research in SA and NLP.

II. LITERATURE SURVEY

S. Parameshwarappa et.al, discussed about generating a Kannada corpus tool for Program Execution and Reporting Language (PERL) by using the web logs. Kannada Corpus Construction algorithm is used to generate a raw corpus. The web is crawled for downloading the corpus using the seed URL's. Only about how to extract the seed words from the given corpus is discussed [1]. However, no concentration on the sentences in Kannada language is done and how to tokenize these sentences and search for a word in the generated corpus is done. Jayashree R discussed about how to retrieve the useful information from the large dataset. The sentiment classification is performed in many ways they are: word level, sentence level and passage level etc. In the paper [2], a sentence level classification is used for NLP applications such as query and replying and summarization systems. The objective of sentence level classification is to classify the text according to the sentimental polarities of opinions. A separate list is used to maintain stop words. The stop words do not give any meaning to the sentence and restricted the word which is appearing only once in the document. A supervised learning method is used to classify the sentence to positive or negative.

Manuscript published on January 30, 2020.

* Correspondence Author

Shankar R*, Department of CSE, BMS Institute of Technology and Management, VTU, Bengaluru, India. Email: shankarbmg@gmail.com

Suma Swamy, Department of CSE, Sir M. Visvesvaraya Institute of Technology, VTU, Bengaluru, India. Email: sumaswamy10@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

K-fold cross validation is used to validate their performance [2]. The multi-label classification and how this approach can be used in online customer reviews in Kannada blogs are not discussed and no method to recommend the product based on the reviews is discussed. Deepamala N has discussed about the polarity detection of the given document by applying sentiment

analysis technique to classify the document with positive polarity and negative polarity. The polarity lexicon method was used, that was manually created for each word to assign a polarity score between the +5 (very positive) to -5 (very negative) to classify the sentence as positive or negative. The authors also concentrated on the suffixes that matches with the stemmed word to classify the sentence as positive or negative. A file with the toggled words is maintained. Results of their Polarity Lexicon algorithm with other machine learning algorithms like finding Maximum Entropy and Naïve Bayes are compared to check the accuracy in the results. But accuracy is less compared to the other two algorithms as it uses a manual approach for assigning a polarity score for each word present in the corpus [3]. Similar levels on Hidden Markov Model are being used in [8].

Shankar R, Suma Swamy et al, have discussed on an elaborate study on Sentiment Analysis in different Indian Dialects. This is an extensive study on various pedagogy followed in Indian languages pertaining to opinion mining [4].

III. PROPOSED METHODOLOGY

Figure 1 shows the proposed system in which two corpora related to the Kannada language are maintained. One corpus contains the positive words like ಅದ್ಭುತವಾಗಿದೆ, ಕೊಳಬಹುದು etc and negative corpus contains the words like ಕೊಳುವುದುಬೇಡ, ಕೆಟ್ಟಮೊಬೈಲ್ವಾಗಿದೆ etc. The given sentence is converted into the tokens and each word is compared with positive and negative corpus. If the word is present in the positive corpus, the positive score is incremented by one; similarly, if the word is present in negative corpus, the negative score is incremented by one. If both the scores are equal, then it is classified under the neutral review. The corpus collection itself is a huge task in Kannada language and the reviews of various mobile products are scraped through web API called Beautiful Soup via Python 3. This scraped data from gadgetloka.com, mobile.gizbot.com/Kannada is fed to the classifier to predict the sentiments. The system uses a series of reviews separated by a common separator (==) and the analysis of every review would be obtained based on the corpus weights which is the count of the corresponding good and bad word count. To perform Sentiment Analysis of Mobile Product Reviews in Kannada, following objectives are defined.

- Building a Kannada Corpora/WordNet.
- Extract features from this WordNet.
- Extract words and phrases denoting opinions that refer to the target feature.
- Classify the extracted opinion as positive, negative or neutral review.

A. Building a Kannada Corpora/Wordnet

A corpus is a collection of spoken or written document in a

structured format. Most Free and Open Source corpora is available for utilization by researchers as standard data sets to develop and test their systems. In particular, a text corpus is a collection of text documents. A plain text document contains only text. Such a document can be displayed on the screen, printed and processed on any computer running any operating system using any standard text editors, without need for specific commercial software's. For most of the Indian languages, large corpora are publicly not available. There is some progressive research going on in this direction with respect to other Indian languages [5][6][7]. Collection of large and balanced corpus for such languages is challenging and difficult on its own. It poses a number of unusual challenges. Kannada language is one among them. [9] talk about this to an extent. Instead, one can turn to Wikipedia, which is a free and open source, multilingual encyclopedia project based on the web and supported by the non-profit Wikimedia Foundation. As of July 2011, it contains more than 11 million articles in Kannada that have been written in the open, with asynchronous collaboration by editors and NLP translators.

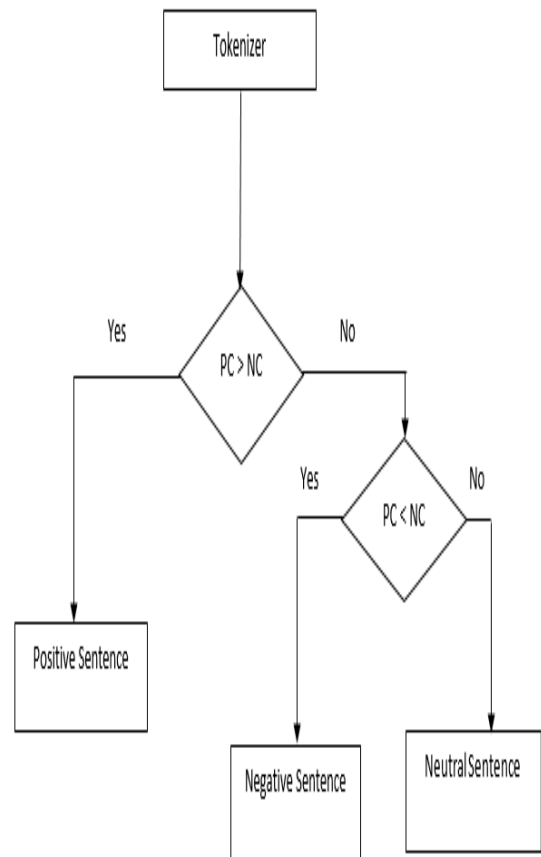


Fig. 1. Sentiment Analysis of a given Kannada sentence.

One can also turn to Websites, and blogs, which sum to provide useful corpora. Some popular blogs mostly in Kannada is Sampada, Ekanasu, and Indiblogger.

B. Extract by features from the given review

Presence of statements which would indicate polarity eg: "samsung bahala akarshakawagide" (ಸ್ಯಾಮ್ಸಂಗ್‌ನ ಹೆಚ್ಚಿನ ಆಕರ್ಷಕವಾಗಿದೆ). Idu atyantha ketta mobile aagide

(ಇದುವಿಷಯವನ್ನು ತಿಳಿಯುವುದಿಲ್ಲ). Without the knowledge of any pre-existing context of the review (regardless of classified or amorphous data), the review gives us a list of possible attributes such that, when clipped, will give us the exact feature.

Example: ನೋಟಿಂಗ್‌ನ ಮೂಲಕ ಅದನ್ನು ತಿಳಿಸುವುದು.

C. Extract words and phrases denoting opinions that refer to the target feature

Keywords like ಅದನ್ನು ತಿಳಿಸುವುದು, ಕೊಳ್ಳಬೇಕು, ಕೊಳ್ಳುವುದಿಲ್ಲ, ಕೆಟ್ಟ ಮೂಲಕ ಇತ್ಯಾದಿ etc should be extracted and matched with the corpora for further analysis. The POS tagging also help in matching with the correct words in the database. It sorts the direction of the opinion of this most recent entity model obtained from the product review. A mainstream method, such as obtaining the words with largest proximity to the target feature, does not obtain similar results when the sentence has distributed emotions and multiple attributes.

C. D. Classify the extracted review as positive, negative or neutral

This step will mark ಕೆಟ್ಟ ಮೂಲಕ ಇತ್ಯಾದಿ as a negative opinion and ಅದನ್ನು ತಿಳಿಸುವುದು as a positive opinion as an example. To add to these common concerns, there are some distinct cases. It might not be possible to get a clear-cut sort of classification, but a vaguer score, which makes it hard to assign the result a value. Some texts will not come under the positive or negative classes. This type of words can be classified as a neutral class. It rather targets the operator. Classifications of such tokens are also challenging.

IV. SYSTEM DESIGN

The figure 2 displays the architecture of sentiment analysis of the given Kannada sentence. First, the input text file is converted into the tokens and these token words are compared with the positive and negative corpus. If the words are present in the positive corpus, then the positive word is incremented count to plus one. If the words are present in the negative corpus, the negative word count is incremented to plus one. If the negative word count is lesser than positive word count, the sentence is classified as positive. Otherwise, the negative word count is greater than the positive word count, and the input text is classified as a negative sentence. Finally, if the count of both positive and negative words is same, it is classified as a neutral sentence (in the condition NWC = PWC)

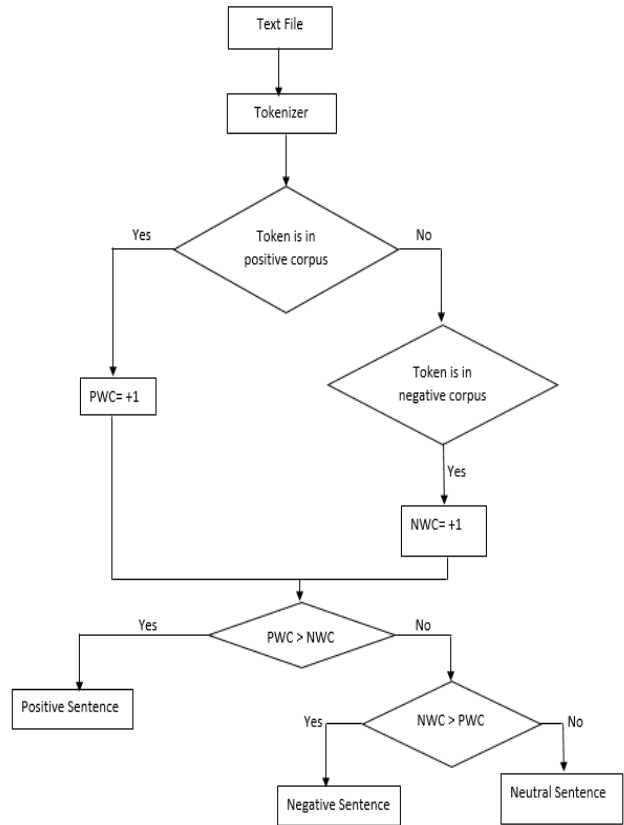


Fig. 2. Design flow of the sentiment analysis for Kannada language.

The complete process of computing the sentiment can be given as a linear function of f(x). Suppose to find the sentiment value of a sentence x consisting the word tokens as x0, x1, x2, x3 etc., can be given as per equation 1:

$$f(x) = t_0x_0 + t_1x_1 + t_2x_2 + t_3x_3 \dots t_{n-1}x_{n-1} \quad (1)$$

where,
tag(t) for each token,
where t = +1, if the token is positive
t = -1, if the token is negative
n is the number of tokens in the given sentence.

The above equation can be summarized as per the equation 2

$$f(x) = \sum_{i=0}^{n-1} tix_i$$

V. RESULTS AND DISCUSSIONS

The initial corpus having around 5016 good and bad lexicons are prepared is matched with the given reviews to analyze the sentiments. The Table 1 shows the sample corpus containing good words and bad words in Kannada language.

Table 1: The sample corpus looks like the following.

Good words sample - Kannada	
ಚೆನ್ನಾಗಿದೆ	ಸಕಾರಾತ್ಮಕ
ಚೆನ್ನಾಗಿದೆ	ಕೈಗೆಟುಕುವ
ಆಡತಡೆಇಲ್ಲದೆ	ಅಗವಾದ
ವಿಪುಲವಾಗಿವೆ	ಚುರುಕುತನ
ಹೇರಳವಾಗಿ	ಪ್ರಿಯವಾದ
ನಿಖರವಾಗಿ	ಒಪ್ಪಿಗೆಯ

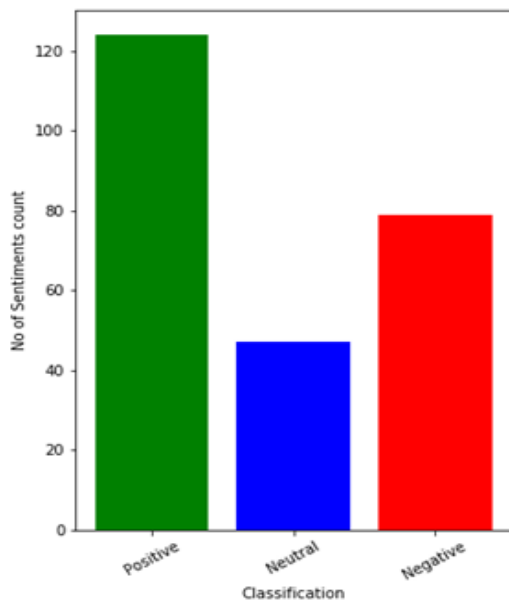


Fig. 5. Comprehensive Result of the Classification

As the bar graph shows, most of the sentences in the data consisted of positive reviews, followed by negative and least amount in neutral reviews for the mobiles. The results are based on the corpus. In addition, it is clear that the results are directly related to the size of corpus available, which in turn directly relate to the quality of analysis that would be done. Giving us a further scope of developing a larger and dynamic corpus. The above results show the necessary measures obtained from the model, by comparing it with results manually.

The table 3 and figure 6 depicts the performance measures of classifiers obtained from the models.

Table 3: Performance measures in %

Measure	Computed Performance in %
Precision	80
Recall	67
Accuracy	71
F-measure	73

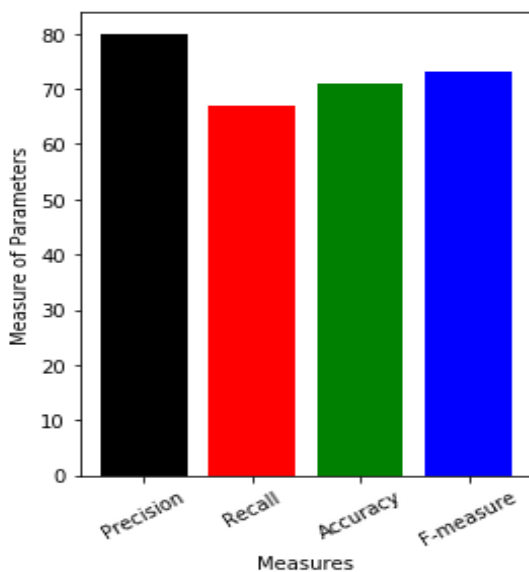


Fig 6: Performance Measures of Classifiers

The Accuracy and Precision measures gives us the insight about the data parsed. The Recall measures tells about the

high order of morphological NLP, statements like Negative Sentences and Sarcasm, which the model could not classify properly.

D. Pseudocode

The model tries to compute the linear polarity value of the sentence given to it, i.e., Positive polarity or Negative polarity and takes a final value to classify the sentence in one of the class.

- Take Input of the sentence to find sentiment.
- Normalize it and remove the stop words from the sentence.
- Tokenize the sentence.
- Initialize the result value to 0, i.e. Neutral by default
- For each token in sentence, check the polarity.
- Perform the sentiment of individual unit
- Compute the summation of the sentence.
- If the NWC > PWC: Negative polarity
- Else if PWC > NWC: Positive polarity
- Else: Neutral sentence.

E. The result is then showed as:

- The input sentence given.
- Final classification given to the sentence.
- The ratio of positive word count to negative word count.

VI. CONCLUSION AND FUTURE SCOPE

Retail industry has a huge potential if they go regional in their business. In this connection, Kannada opinions classification work has tremendous benefits. The method that is discussed in this article comprehensively classifies a given Kannada review in to positive/negative/neutral polarities. It is intended to work on generating more corpus and to classify in to polarity based on the entire document rather than a sentence, which can be performed on the hierarchy structure, computation of sentiment in the order of word to sentence, sentence to paragraph and paragraph to document. Advancements in mathematical computation could be made by providing weightage to specific tokens. As it is known, some specific words put a lot more weightage to the sentence as compared to the others, which changes the polarity of the sentence completely. This process requires an upgrade in the model of corpus to store the words in the level of weightage. The functionality should determine the level of negative or positive polarity in the sentence, by computing the result on varied distribution of weights, for better predictions.

ACKNOWLEDGMENT

The authors would like to express their heart-felt gratitude to Mr. Sahil Verma and Mr. Pranav Raikote, students of B M S Institute of Technology and Management for their support in completing this work.

REFERENCES

1. S. Parameshwarappa, V. N Narayana, G.N Bhrarathi, "A Novel Approach to Build Web Corpus", International Conference on Computer Communication and Informatics (ICCCI-2012).

2. Jayashree R, Sreekanta Murthy K, "An Analysis of Sentence level Text Classification for the Kannada Language" International Conference of Soft Computing and Pattern Recognition (SoCPaR)- 2011.
3. Deepamala N, Ramnath Kumar P, "Polarity Detection of a Kannada Document" IEEE-2015.
4. Shankar R, Suma Swamy, A Survey on Sentimental Analysis in Different Indian Dialects, International Journal of Advanced Research in Computer and Communication Engineering, Vol5, 1072-1076. 10.17148/IJARCCE.2016.54262.
5. Piyush Arora, Sentiment Analysis for Hindi Language, MS Thesis IIT-H 2013.
6. Sandeep Chandran, Bhadran V K, Santhosh George, Manoj Kumar P, "Document Level Sentiment Extraction for Malayalam", International Conference On Recent Advances In Engineering, Science & Technology (Icon 2015)
7. Anu Sharma, Sentiment Analyzer using Punjabi Language, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 9, September 2014
8. Muralikrishna H, Ananthakrishna T, Kumarasharma, HMM Based Isolated Kannada Digit Recognition System using MFCC, International Conference on Advances in Computing Communications and Informatics (ICACCI)-2013
9. Kalyanamalini Sahoo, V Eshwarchandra Vidyasagar, Kannada WordNet - A Lexical Database, TENCON 2003

AUTHORS PROFILE



Shankar R., has completed Bachelor of Engineering in Computer Science and Engineering from Visvesvaraya Technological University, Master of Engineering in Web Technologies from Bangalore University. He has 9 years of teaching experience and 1-year industry experience. His intent is to pursue R&D in the field of Sentiment Analysis. He is an active member of IUPRAI.



Dr. Suma Swamy, completed her Bachelor of Engineering in Electronics from Shivaji University in 1990, M.Tech in Electronics from Visvesvaraya Technological University in 2005 and Ph.D in Information and Communication Engineering from Anna University, Chennai in 2014. She is presently working as a Professor in the department of Computer Science and Engineering at Sir M. Visvesvaraya Institute of Technology, Bengaluru with teaching experience of 29 years. She has around 29 publications in various reputed and refereed journals with 72 citations and h index of 4. She has published a book titled "Practical Applications of Speech Signals". She has filed a Patent titled "Cost Effective Ingestible Battery less Electronic Health Pill to Predict Heart Attacks and sudden Cardiac Arrests". She is an Editorial Board member of Science Publishing Group, USA.