Sentiment Analysis using Random Forest Ensemble for Mobile Product Reviews in Kannada

Yashaswini Hegde
SJCE, Mysore
yhegde@gmail.com

S.K. Padma
Department of Information Science
SJCE, Mysore
padmask@gmail.com

***Abstract*---Sentiment Analysis(SA) for Kannada documents has been explored recently. In the recent study [8], the sentiment analysis for Kannada text is explored using Naive Bayes classifier. The objective of this work is to improve the performance of the previous study on the sentiment analyzer for Kannada language explored in the paper [8]. In this work, we propose the ensemble of classifier with random forest technique to identify the polarity of the sentiment and test the performance of the same. Also in this work, some of the limitations of [8] such as handling multi class labels, identification of sentiment polarity of comparative and conditional statements have been addressed. The over all accuracy is improved from 65% to 72 % , indicating our approach based on Random Forest technique is more efficient for SA for Kannada.**

***Keywords*-Product Reviews, Sentiment Analysis, NLP, Ensemble of classifiers, Random Forest, Kannada, ಭಾವನೆಗಳ ವಿಶ್ಲೇಷಣ ತಂತ್ರ,ಕನ್ನಡ**

I. Introduction

A. Sentiment Analysis

A computational methodology of identifying, extracting the sentiment contents found in the text, speech or databases is - the Sentiment Analysis (SA) or Opinion mining. SA also characterizes the emotions, attitudes, subjective impressions, opinions. The polarity or the tendency of the data under consideration can be determined by SA using Natural Language Processing (NLP), Statistics or Machine learning techniques [1] [4] [3]. Generally the accuracy and precision of SA depends on the best preprocessing and sentiment extraction methods,then the classifiers which identifies the polarity of the text. Again the precision and the accuracy of a classifier depends on many factors like choosing

- features selected in the training set
- an efficient classification algorithm or ensembles of same or different base classifiers with lesser average error
- reusing and selecting the multiple versions of data set etc.

Among all these factors ensemble learning technique is considered in our experiment,for Kannada Sentiment analysis.

B. Ensemble learning or Committee based methods

Ensemble learning or committee based method is a consensus approach where instead of single classifier, ensemble of classifiers used to combine the predictions,C_1(X),.....C_m(X). This technique is used to improve the performance of the classifier model. It is designed to correct the errors of its individual members. And hence as a strategy each classifier in the ensemble is allowed to make different errors on different instances of training data samples. Then the combination of these diverse classifiers can reduce the total error. The popular method used to combine the predictions is Majority vote method. Generally majority vote classifier consisting of votes from classifiers $h_1, h_2, h_3...$ is defined as

$$C(X) = argmax_i \sum_{j=1}^{B} w_j I(h_j(X) = i)$$

where w_i are weights and can be set 1/B to get mode of $h_1, h_2, h_3...$ and I is an indicator function. Predictive classifiers like Random forest, Adaptive Boosting are some such ensemble learning techniques.

In section 2, we observe a brief survey of sentiment analysis with ensemble techniques and ensemble classifiers for Kannada, in Section 3, issues addressed in Kannada SA, in section 4, we provide our approach of Kannada Product reviews with Random Forest. And in 5th section, evaluation of the algorithm and Results are discussed.

II. Related Background

The SA is a problem of categorization of sentiment polarity[10]. The problem is categorizing the text by finding a specific polarity, positive or negative (or neutral). There are three levels in categorization of sentiment polarity, namely the entity and aspect level, the sentence level, the document level[12]. In the document level, it is to determine that the whole document expression is negative or positive sentiment. The sentiment of each sentence is categorized in sentence level, where in the entity and aspect level, it is categorizing the opinion of people from their likes and dislikes.

In this section, only a few previous works are observed which are related to Sentiment analysis of product reviews. Hu and Liu [17] has given the list of 2006 positive words and 4783 negative words including some misspelled words which would to help to categorize the customer reviews.

978-1-5090-1560-3/17 $31.00 © 2017 IEEE
DOI 10.1109/IACC.2017.151

Pang and Lee [10] extracted subjective ones, then removed objective sentences, to select the features. They proposed a minimum cut algorithm as a text-categorization technique, to identify the subjective contents. Gann et al. [15] selected 6,799 tokens from the Twitter data.They assigned each token a sentiment score, called TSI(Total Sentiment Index), which tags each token as a positive token or a negative token. Other related works on sentiment analysis can be referred in [12],

A. *Kannada Sentiment Analysis*

There are very few works related to Kannada sentiment analysis that can be listed. Among them Anil Kumar et al. [13] worked on general web documents of various topics . They have worked on pattern based approach and POS tagging and later Turney's and Negator algorithms; Deepamala et al. [14]worked on Kannada sentence boundary detection using rule based and maximum entropy methods.

III. Kannada Sentiment Analysis - Issues

Some of the issues, that are not addressed in [8], and our approach to address those issues are:

1 "samsung avara app gaLa mElE beraLu aaDisidare vishESha pratikriye neeDuttade"ಸಾಂಸ್ಯಂಗ್ ಅವರ ಆಪ್ ಗಳ ಮೇಲೆ ಬೆರಳು ಆಡಿಸಿದರೆ ವಿಶೇಷ ಪ್ರತಿಕ್ರಿಯೆ ನೀಡುತ್ತದೆ.) "ee phone nalli ellaa namUneya aaTagalannu aDetaDeyillade aaDabahudu",(ಈ ಫೋನ್ ನಲ್ಲಿ ಎಲ್ಲಾ ನಮೂನೆಯ ಆಟಗಳನ್ನು ಅಡೆತಡೆಯಿಲ್ಲದೆ ಆಡಬಹುದು.) - to address this kind of sentences which gives positive opinion yet not using any sentiment lexicon an aspect "heccugaarike" is added and they are also rated.

2 "olleya business phone aadare idu neeDuva savalattige bele jaasti" (ಒಳ್ಳೆಯ ಬಿಸಿನೆಸ್ ಫೋನ್ ಆದರೆ ಇದು ನೀಡುವ ಸವಲತ್ತಿಗೆ ಬೆಲೆ ಜಾಸ್ತಿ) an additional class2 label(multi class label) is added which rates positive/native opinions for this kind of sentences which gives conditional opinions.

3 To address presence of comparative statements like "samsung iphone ge paipOTi neeDaballudu" (ಸಾಂಸ್ಯಂಗ್ ಐಫೋನ್ ಗೆ ಪೈಪೋಟಿ ನೀಡಬಲ್ಲದು), "idE shReNiya nokia lumiya gaLige hOlisidare idara kaaryaacharaNa vyavasthe uttama ennabahudu" (ಇದೇ ಶ್ರೇಣಿಯ ನೋಕಿಯ ಲ್ಯುಮಿಯಗಳಿಗೆ ಹೋಲಿಸಿದರೆ ಇದರ ಕಾರ್ಯಾಚರಣ ವ್ಯವಸ್ಥೆ ಉತ್ತಮ ಎನ್ನಬಹುದು) a preprocessing is done programmatically while extracting the aspects which checks previous listing of such aspects and numerically rates compared to the previous ratings.

IV. Experimental Scenarios

A lexicon based SA is chosen, in our approach of Kannada SA, where we have used lexicon entity models for extracting the aspects as in [8]. Further to improve the results we have experimented with Random Forest (RF) ensemble techniques.

A. *DataSet and Feature Selection*

For the data set, weekly mobile product reviews - 'GadgetLoka' [5] is considered, by U.B Pavanaja (from famous Kannada daily 'Prajavani'). We have extended the corpus in [8] with some more aspects around 32, in order to over come some of the limitations such as multi class,comparative and conditional statements in sentiment analysis in [8]. The Kannada unicode text representation is used as the training samples, for the base classifier. The feature set as given in the table IV.1 contains both continuous and categorical aspect values.

A sample entity model is shown in table IV.1. A similar entity model with 28 records(phones) and 32 features serves as the training set for the RF ensemble. It predicts the polarity of the sentiment of the new entity model extracted from the mobile product review. The RF algorithms are used to classify the sentiments of the mobile product reviews in Kannada.

B. *Random Forest*

Random forest (RF) is an ensemble of decision tree classifiers which will output a combined prediction value of each tree in the ensemble. Each decision tree is constructed by using a random subset of the training data with a fixed probability distribution. The deeply grown decision trees has low bias and high variance. Hence they can learn irregular patterns and overfit their training sets. Random forests give improvement over just bagged trees because they decorrelates the trees in the Random forest. The decorrelation is achieved While building the RF. While building the decision trees of RF, a random sample of some M training samples are chosen as split candidate from the original N training set, during a split in a tree is happens each time. This strategy is good because, if at all any strong training sample in the training set bagged trees use this strong sample to split and subsequently all trees look similar and they become correlated and any reduction in the variance is only average of many correlated predictions.But in RF each split will consider subset of the training samples and hence on an average (N-M)/M splits not even consider that strong training sample and so other training samples also have more chance to be the splitting candidates. And due to this procedure the generated trees will become decorrelated and average of the results of the classifier trees will be less variance and hence gives more reliable solution. If RF is built with M=N samples then it is as good as Bagging tree.But generally RF is built with $M = \sqrt{N}$ where N is size of training sample and reduces both test error and out-of-bag error.[7] [6]

The algorithm 1 is executed and tested in R studio environment using RF library.This RF library has the features of estimating important variables in the classification, generating unbiased estimate of the error of generalization as the forest build progresses and has an effective method for estimating missing data. It maintains accuracy when a

Table IV.1
Aspects and their values of different phones: A sample input

Aspects/Features	Phones	
	One Plus	**Jioni elife e3**
ವೇಗ(ಗಿಗಾಹರ್ಟ್ಸ್)	2.5	1.2
ಹೃದಯ	ನಾಲ್ಕು	ನಾಲ್ಕು
ಪ್ರೊಸೆಸರ್	-	Cortex A7
ಗ್ರಾಫಿಕ್ಸ್ ಪ್ರೊಸೆಸರ್	-	
ಮೆಮೊರಿ(ಗಿಗಾಬೈಟ್)	೧೬	೧ + ೧೬
ಅಧಿಕ ಮೆಮೊರಿ(ಗಿಗಾಬೈಟ್)	-	32
ಕಾರ್ಯಾಚರಣ ವ್ಯವಸ್ಥೆ	ಆಂಡ್ರಾಯಿಡ್ 4.4.4 ಸಯನೋಜನ್	ಆಂಡ್ರಾಯಿಡ್ 4.2
ಪಿಕ್ಸೆಲ್ ರೆಸೊಲೂಶನ್	1080 x 1920	1280 X 720
ಪರದೆ/ಟಚ್ ರೆಸ್ಪಾನ್ಸ್	ಸ್ಪರ್ಶಸಂವೇದಿ	ಸ್ಪರ್ಶಸಂವೇದಿ
ನೆಟ್‌ವರ್ಕ್ ಬೆಂಬಲ	-	2/3ಜಿ
ಕ್ಯಾಮೆರಾ(ಮೆಗಾಪಿಕ್ಸೆಲ್)	13	8
ಸ್ವಂತೀ	-	2
ವಿಡಿಯೊ	ಹೈಡೆಫಿನಿಶನ್	೭೨೦p/೩೦fps
ವೈಫೈ	ಇದೆ	ಇದೆ
ಬ್ಲೂಟೂತ್ ಸಂಪರ್ಕ	ಇದೆ	ಇದೆ
ಯುಎಸ್‌ಬಿ	ಆನ್ ದ ಗೋ ಇದೆ	ಇದೆ
ಜಿಪಿಎಸ್	ಇದೆ	ಇದೆ
ಅವಕೆಂಪು(ಇನ್‌ಫ್ರಾರೆಡ್) ದೂರನಿಯಂತ್ರಕ	ಇಲ್ಲ	
ಎನ್‌ಎಫ್‌ಸಿ	ಇಲ್ಲ	
ಬ್ಯಾಟರಿ(mAh)	3100	1800
ಗಾತ್ರ(ಮಿ.ಮೀ.)	152.9 x 75.9 x 8.9	137.5 X 68.4 X 7.9
ತೂಕ (ಗ್ರಾಂ)	162	-
ವಿನ್ಯಾಸ/ರಚನೆ	ತುಂಬ ಚೆನ್ನಾಗಿದೆ	ಚೆನ್ನಾಗಿದೆ
ಎಫ್‌ಎಂ ರೇಡಿಯೊ	ಇಲ್ಲ	ಇದೆ
ಸಂಗೀತ/ಧ್ವನಿಯ ಗುಣಮಟ್ಟ/ಆಡಿಯೊ	ಚೆನ್ನಾಗಿಲ್ಲ	ತುಂಬ ಚೆನ್ನಾಗಿದೆ
ಇಯರ್‌ಫೋನ್	ಇಲ್ಲ	ಇಲ್ಲ
ಆಟ ಆಡುವ ಅನುಭವ	ಚೆನ್ನಾಗಿದೆ	ಚೆನ್ನಾಗಿದೆ
ಕ್ಯಾಮೆರಾ ಗುಣಮಟ್ಟ	ಚೆನ್ನಾಗಿದೆ	ಚೆನ್ನಾಗಿದೆ
ವಿಡಿಯೊ ವೀಕ್ಷಣೆ/ಗುಣಮಟ್ಟ	ಅತ್ಯುತ್ತಮ	ಉತ್ತಮವಾಗಿದೆ
ಕನ್ನಡ ಪಠ್ಯದ ತೋರುವಿಕೆ	-	ಇದೆ
ಹೆಚ್ಚುಗಾರಿಕೆ		ಪ್ರಮುಖ ಕೊರತೆ ಎಂದರೆ ಇದರ ಬ್ಯಾಟರಿ.
ಬೆಲೆ	21,373	15,000
ಕ್ಲಾಸ್	**ಕೊಳ್ಳಬಹುದು**	**ಕೊಳ್ಳಬಹುದು**
ಕ್ಲಾಸ್ 2	**3**	**4**

large proportion of the data are missing. In RF library the OOB is computed at the end of the run. And it is computed assuming j as the class that got most of the votes every time and case n was left out samples then the proportion of times that j is not equal to the true class of n averaged over all cases is the OOB error estimate. This error gives unbiased estimation of error in classification. And also an estimation of which aspects are important(variables of importance).

Algorithm 1: **SA with Random Forest for Kannada**

1: **procedure** RF-KAN
2: Let N be the size of the training set.
3: Let the training set is the sample of the same size N, drawn with replacement from the original data.
4: If there are M input variables, a number $m << M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
5: Each tree is grown to the largest extent possible. There is no pruning.
6: Predict the class of the test set using Majority vote method.
7: **end procedure**

end

This algorithm has been executed with 20 to 5000 trees in the forest. The missing data is chosen as 'maximum numbered value' since our aspect values are categorical. We have chosen the number of variables randomly sampled as candidates at each split as $\sqrt{(no - of - variables)}$ with replacement.

While training cross-validation is considered though it is estimated internally. The set of training samples are drawn with replacement and generally one third of the samples are left out of the original training samples. This is OOB data (out-of-bag) and in our case it varies from 28% to 65%.

Generally with committee learning methods (ensemble of classifiers) the trade off is, computation time and classifier efficiency.

The figures 1, 2, 3, 4, 5, 6 pictorial representations of various aspects of the experiments conducted.

The Mean Decrease in Gini as in fig 1 represents the node impurity. In other words when the tree is split with node on some variable m the gini impurity criterion for the two descendant nodes is less than the parent node. Adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance.

From the fig 1 the important features are identified. The figs 2 and 3 shows the growth of single Decision tree of the Random Forest. The three class la-

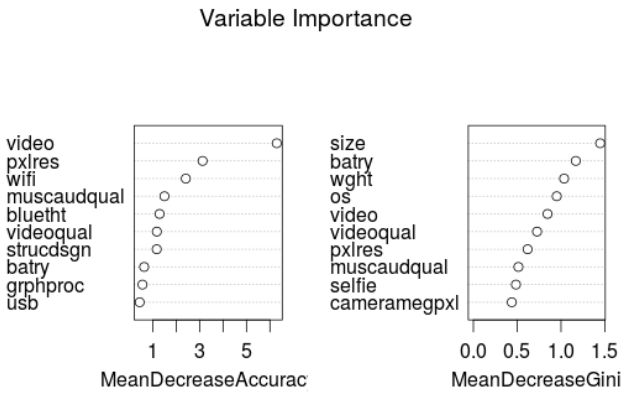

Figure 1. Importance of variables of the training samples with

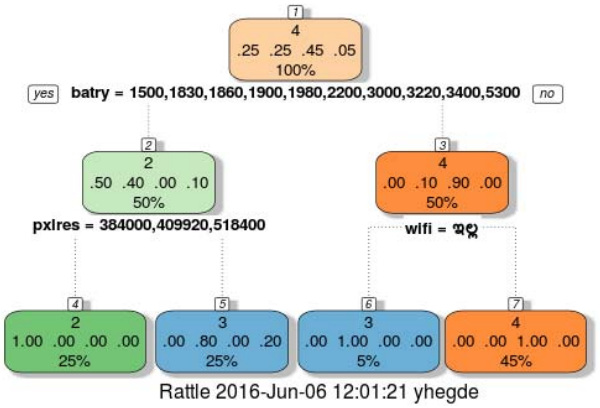

Figure 3. Growth of single Decision Tree with 4 class labels

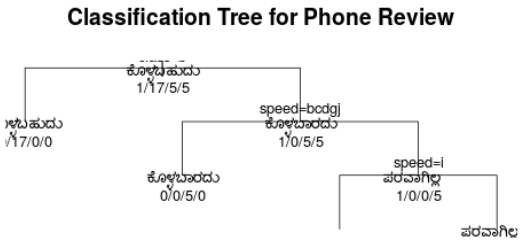

Figure 2. Growth of single Decision Tree with with 3 class lables

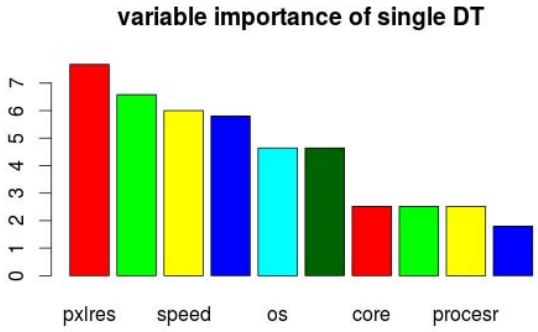

Figure 4. Important variables of classifier with single tree

bels (ಕೊಳ್ಳಬಹುದು,ಪರವಾಗಿಲ್ಲ,ಕೊಳ್ಳಬಾರದು) indicating the sentiments of reviewer (towards the buy-able quality of the phone) are further rated to four class labels (2,3,4,5 for Bad Buy,Okok buy,Good for price and Best for the price respectively) as in figs 2 and 3. And these class labels and ratings are introduced to address the issues discussed earlier.

C. *Evaluation of Algorithm with RF ensemble*

The efficiency of our algorithm is evaluated using the data set created out of Kannada product reviews. In this experiment the data is divided into random subsets and treated as training samples to build the model. And each time the number of trees are varied from 20 to 5000. The evaluation measures used are standard ones and are explained in the next section.

D. *Results*

The output of our RF ensemble classifier are: Confusion matrix of the predicted label, the important aspects/features which effect the efficiency of the classifier model, the OOB percentage, statistics for each class label,the overall accuracy etc. In our experiment the OOB % varies from 25% to 65%. The class error varies from 40% to 100%. The higher error rate is due to class imbalance problem where we have lesser number of training samples for specific type of class labels. The overall results are tabulated in the tables IV.2 IV.3 and IV.4.

The evaluation measures are calculated using the confusion matrix tabulated in IV.2. The diagonals contains the true positives (TP) for respective labels. For a specific label the remaining diagonal values are true negatives (TN). For example for the label 2(bad buy) 3 is TP and 5,6,1 are TNs with respect to label 2. For label 3, TP=5 and 3,6,1 are TNs and so on. Thus TP+TN=15 for all the labels. Hence

Accuracy: $= \frac{(TP+TN)}{(TP+FN+FP+TN)} = 15/16 = 93.75\%$ (for label 2)

Precision: $= \frac{TP}{(TP+FP)} = 3/4 = 75\%$ (for label 2)

Recall: $= \frac{TP}{(TP+FN)} = 3/3 = 100\%$ (for label 2)

Figure 5. Prediction with 4 class labels

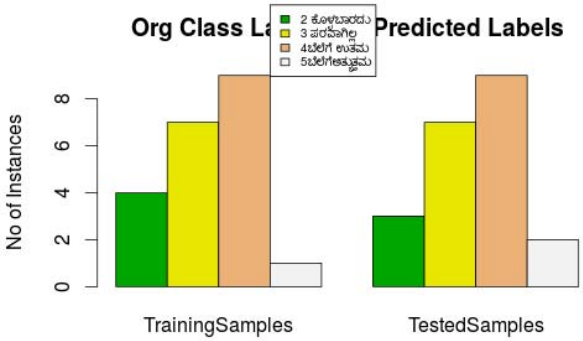

Figure 6. Training v/s Test Prediction

F-Measure: $= \frac{(2TP)}{(2TP+FN+FP)} = 6/7 = 85.71\%$ (for label 2)

Similarly these evaluation measures are computed for the rest of the labels and tabulated in IV.3

Table IV.2
CONFUSION MATRIX OF THE PREDICTIONS

ClassLabels	2	3	4	5	TotalPred(TP+FP)
2(Bad buy)	3	0	1	0	Lbl2=4
3(Okok buy)	0	5	2	0	Lbl3=7
4(Good buy)	0	2	6	1	Lbl4=9
5(Best buy)	0	0	0	1	Lbl5=1
Total(TP+FN)	Lbl2=3	Lbl3=7	Lbl4=9	Lbl5=2	

Statistical Measures used to evaluate the performance of RF ensemble are:

1. **Sensitivity/true positive rate(TPR)/recall =** $\frac{TP}{TP+FN}$

Table IV.3
EVALUATION MEASURES OF ENSEMBLE OF CLASSIFIERS

RF Ensemble Labels	Accuracy	Precision	Recall	F-measure
2(Bad buy)	93.75	75	100	85.71
3(Okok buy)	79	72	72	72
4(Good buy)	71.42	67	63	64.93
5(Best buy)	93.75	100	50	66.66

Table IV.4
STATISTICS BY CLASS

ClassLabels	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	1.0000	0.7143	0.6667	0.50000
Specificity	0.9444	0.8571	0.7500	1.00000
Pos Pred Value	0.7500	0.7143	0.6667	1.00000
Neg Pred Value	1.0000	0.8571	0.7500	0.95000
Prevalence	0.1429	0.3333	0.4286	0.09524
Detection Rate	0.1429	0.2381	0.2857	0.04762
Detection Prevalence	0.1905	0.3333	0.4286	0.04762
Balanced Accuracy	0.9722	0.7857	0.7083	0.75000

2. **Specificity(SPC)or true negative rate =** $\frac{TN}{TN+FP}$
3. **Precision or positive predictive value (PPV) =** $\frac{TP}{TP+FP}$
4. **Negative predictive value (NPV) =** $\frac{TN}{TN+FN}$
5. **Prevalence =** $\frac{TP+FN}{TP+TN+FP+FN}$
6. **Detection Rate =** $\frac{TP}{TP+TN+FP+FN}$
7. **Detection Prevalence =** $\frac{TP+FP}{TP+TN+FP+FN}$
8. **Balanced Accuracy =** $\frac{\frac{TP}{TP+FP}+\frac{TN}{TN+FN}}{2}$

Balanced accuracy is required when the test set are not balanced.

These statistical Measures are also computed (like accuracy , precision etc.) for all the labels and tabulated in percent, in the table IV.4.

The working environment RStudio is shown in the figure 7 with the results.

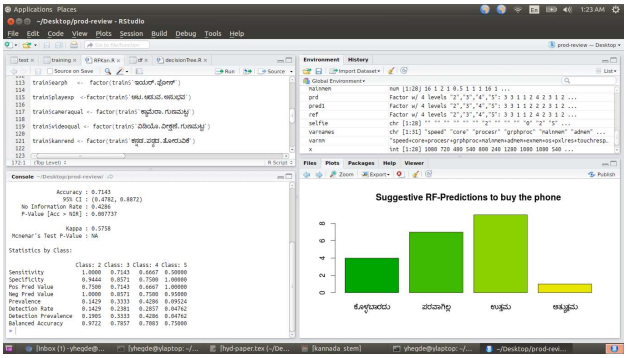

Figure 7. Result in RStudio

V. Conclusion

Our results indicate that Random forests is a good classifier for classifying the multi class Kannada sentiments with over all accuracy of 72%. In this our experiments we have successfully addressed multi class sentiments and sentiments with conditional statements by introducing few more aspects in to the data set. But working on a large carpus is still not possible because of limited number of product review articles in Kannada. We propose to apply, in the future, sophisticated, different aspect extraction methods, with larger data set and natural language processing techniques.

References

[1] Vohra,Teriaya "A COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES "

[2] Bing Liu, Sentiment Analysis and Opinion Mining, 2012.

[3] Mullen, T., Collier, N. (2004). “Sentiment analysis using support vector machines with diverse information sources”. In Proceedings EMNLP’04,pp.412-418.

[4] Medhat W et al., Sentiment analysis algorithms and applications: A survey, Ain Shams Eng J (2014), http://dx.doi.org/10.1016/j.asej.2014.04.011

[5] http://www.prajavani.net/columns/ಗ್ಯಾಜೆಟ್ ಲೋಕ ಯುಬಿ ಪವನಜ

[6] Gareth James, Daniel Witten , T.Hastie, R.Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer New York, 2014. ISBN 1461471370,9781461471370.

[7] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. ©2006,Pearson

[8] Yashaswini Hegde, S.K.Padma. "Sentiment Analysis for Kannada using mobile product reviews: A case study", Advance Computing Conference (IACC), 2015 IEEE International.

[9] Liu B, Hu M, Cheng J (2005) Opinion observer: Analyzing and comparing opinions on the web In: Proceedings of the 14th International Conference on World Wide Web, WWW ’05, 342–351.. ACM, New York, NY, USA.

[10] Pang B,Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL ’04.. Association for Computational Linguistics, Stroudsburg,PA,USA.

[11] Kim S-M, Hovy E (2004) Determining the sentiment of opinions In: Proceedings of the 20th international conference on Computational Linguistics, page 1367.. Association for Computational Linguistics, Stroudsburg, PA, USA.

[12] Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification In: Proceedings of the 49th, Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 151–160.. Association for Computational Linguistics, Stroudsburg, PA, USA.

[13] Anil Kumar KM, Asmita Poojari, Mohana kumari M. Pattern based Approach for Mining Users Opinion from Kannada Web Documents. Discovery, 2015, 45(209), 138-143

[14] Deepamala.N a, Dr. Ramakanth Kumar.P b, Surendra.H c "Kannada Sentence Boundary Detection using Rule based and Maximum Entropy Methods" , ELSEVIER

[15] Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system In: Proceedings of the sencond ASE international conference on Big Data.. ASE.

[16] Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining In: Proceedings of the Seventh conference on International Language Resources and Evaluation.. European Languages Resources Association, Valletta, Malta.

[17] Hu M, Liu B (2004) Mining and summarizing customer reviews In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177.. ACM, New York, NY, USA.