Text Mining Based Approach to Customer Sentiment Analysis Using Machine Learning

Gurjeet Kaur[1]* Richa Dutta[2]

[1] Research Scholar, M.Tech (CSE) Yamuna Group of Institutes, Yamuna Nagar

[2] Assistant Professor, Yamuna Group of Institutes, Yamuna Nagar

Abstract – In the present competitive business scenario vast amount of consumer reviews are written on Web about any product or service whether available online or offline. Web stores a huge amount of customer reviews on any service or product popular amongst masses. The advent of social media and ecommerce has brought the era of a new age business and its customer base is growing exponentially every year. In today's world, the online market is increasingly getting popular and it becomes more and more important to help the customer get the best product by all parameters. The quality of a product is best confirmed by taking the customer reviews from those who are already using that. All popular shopping websites like Amazon, flipkart, ebay etc allow customer reviews once the product has been purchased. These reviews are such huge in numbers on these websites that it is not possible for a customer to consider them all. The proposed work uses text mining techniques like Stanford parser, Sentiword Net and Wordnet 2.1 to parse and extract the sentiment from the reviews in the dataset. This research uses dataset from Amazon.com for musical instruments. The dataset is in JSON parser. The results received from the implementation of proposed technique ascertain the effectiveness of methodology. The computed results when compared to results obtained from Amazon using SVM and Naïve Bayes classifiers confirm that the proposed technique has better performance than the base research by Rushleen et al., who could best achieve 80% accuracy with the dataset adopted against 94.005% achieved by proposed technique. The results from Naive Bayes were found to be better and more explanatory for the inputted data.

----------X----------

1. INTRODUCTION

The changing lifestyles and increasing trend for online transactions and activities like chatting, conferencing, e-commerce, social media communications and online transactions, make internet a very huge store of structured and unstructured data. We can apply Data Mining, Web Mining and Text Mining techniques on this huge amount of information, related to customer opinions/reviews to extract and analyze generalized opinion summary for any product or service.

Opinion Mining or Sentiment Analysis is the technique to analyze public opinion or sentiment from online reviews and increase the credibility of valid products and services. The end user's opinion has always been one missing aspect from offline markets where a good promotion can boost inferior products. The popularity of online review web-sites and blogs/forums bring to us reviews on everything that's available in online or offline market. Sentiment analysis can be classified as semantic orientation-based approaches or knowledge-based or machine-learning algorithms.

Sentiment Mining:

Today, numerous customers and users share their experiences using various social media sites such as Twitter, Facebook and blogs. It has become a challenge for organizations to monitor and understand what people post on social media sites. The need and availability of text mining, sentiment analysis and social network analysis extracts meaningful knowledge and insights. These techniques are continuously evolving and still in very preliminary stage, still needing lot of innovation to automate various aspects of sentiment mining that can be very helpful to any consumer of products or services. Morinaga et al. presented a framework for mining public opinions related to product reputation on the Internet. The researchers find that customer sentiment mining offers increased knowledge discovery from public opinion, as compared to the conventional survey approach.

2. BASE RESEARCH

In the base research, Rushleen Kaur et al in [9] aimed to undertake a stepwise methodology to determine the effects of an average person's tweets over fluctuation

of stock prices of Samsung electronics ltd. It involved extracting tweets from tweeter, data cleaning and application of suitable algorithm to extract the correct sentiment from these. The authors studied the vast impact by twitter feeds. The algorithm accurately analysis the positive, negative and moderate tweets. The algorithm accuracy is measured in terms of accuracy percentage and time complexity. These values are found to be 80% and O(m*n) respectively.

3. PROPOSED METHODOLOGY

The customer reviews of different products for any particular enterprise are considered to extract entity level sentiments. Sentiment mining starts with data acquisition and then data pre-processing. This removes most of the irrelevant content from this huge text getting more refined results. The analysis of reviews is done by (a) listing important features of product (b) assigning an optimum weightage for each of them. This allows us to structure information from reviews by summarizing them in a comprehensive and concise form.

The methodology can be described as below: The review is available in form of sentence S_i and we need to compute the sentiment scores $SS_{a,i}$ of relevant features.

ALGORITHM

Step1: Input all reviews in form of Text

Step2: Organize the reviews into array of sentences and the array elements represent individual review.

Step3: Repeat steps for i=1 to length (array)

Step4: We denote each review by s[i]

Step5: Split s[i] into different factors and the sentiment of each factor is evaluated by sentiment = evaluate (factor)

Step6: Final total Sentiment of all elements

Step7: Evaluate the sentiment to order of 5

Step 9: Find comparison of evaluated result from proposed technique and from the base technique. The comparison will be done using Naïve Bayes and Decision tree J48 algorithm. [end loop].

S -> sentence

Check Value(S)

Take initial value = 0

Step: Convert the sentence to lowercase

Step 2: Perform Filtering to remove extra symbols and unwanted words from the sentences.

Step 3: Each word is taken through stemming

Step 4: Finally extract the sentiment carrying words and compare them to lists of positive words, negative words, domain specific positive and domain specific negative words.

Step 5: Each match modifies the result value.

Match (Word1, Word2)

Step 1: Return true if word1 = word2

Step 2: x = removeVowels (word1)

y = removeVowels (word2)

Step 3: if x=y

Return true

Step 4: Now the synonyms of the words are found by Wordnet and results true when successfully matched, otherwise

return false;

The process flow diagram of our work is shown in the figure below.

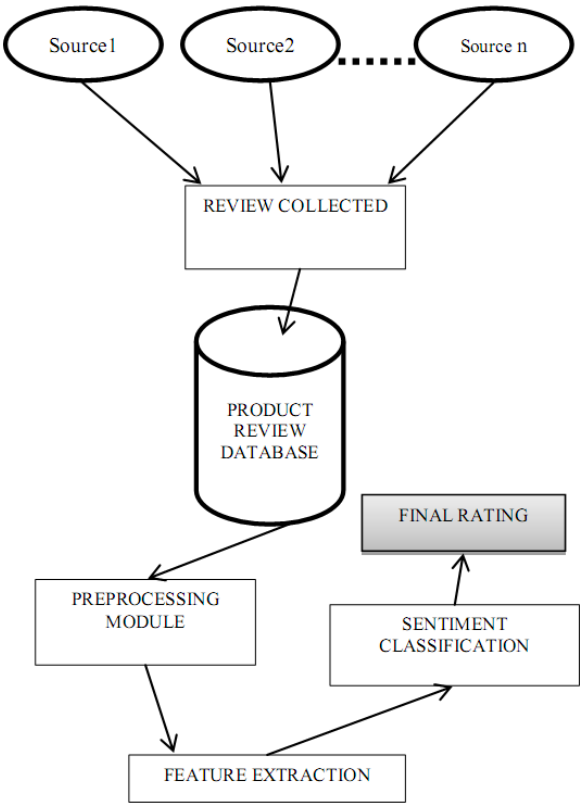

Fig. 1 Demonstration of Research process

A. DATASET USED

The JSON review format is as below:

Data Set format:

Field Name	Datatype
ReviewerId	string
Asin	numeric
reviewerName	string
helpful	array
reviewText	string
Overall	double
Summary	string
unixReviewTime	long number
reviewTime	timestamp

The three most useful fields for our implementation are reviewText, Overall and summary. A part of the json data available to us from AMAZON.COM is listed below:

{"reviewer ID": "AGX9PFO7K90DJ", "asin": "B0053CUHMG", "reviewer Name": "Scotty Boy", "helpful": [0, 0], "review Text": "Beautifully finished. Everything about the construction of instrument is unbelievable. Nice sound in the acoustic too. If this is an indicator of Indonesian made instruments, then all string instruments should be made there. Epiphone has outdone itself. I absolutely adore this instrument. I would love to have a Baritone Uku of this same design.", "overall": 5.0, "summary": "Precision, Accuracy and a work of art", "unix Review Time": 1380153600, "review Time": "09 26, 2013"}, {"reviewer ID": "A1ROUMJOGO4QMB", "asin": "B0053CUHMG", "reviewer Name": "Steve", "helpful": [1, 1], "review Text": "Dunno how, but mine came with no cable and the pickguard wasn't installed. The pickguard is no biggie, I actually didn't want it on there (it seems oversized to me). If I wanted it on their tho, it's got a peel off sticky back. Not having a cable was kinda a big deal. I sent an email to Epiphone (Gibson), got a response within a day and after sending in proof of purchase, they promised to send me a cable. Nice customer service. As far as the Uke, the finish on mine was perfect, no blemishes I can see. I did change out the strings, and it does sound a 'little' better I suppose, but not that much. Probably would have messed around on the stock strings a bit longer, but I listened to the reviews and did a change right away. For $100, if you want a uke, this one looks awesome and plays well.", "overall": 4.0, "summary": "Nice Uke. Had a problem that was taken care of quickly.", "unix Review Time": 1391126400, "review Time": "01 31, 2014"}, {"reviewer ID": "AHUCLL02HS7M5", "asin": "B0055V7UR0", "reviewer Name": "Alex Bartlett", "helpful": [0, 0], "review Text": "I received a selection of pics for a reasonable price. The variety of materials and thicknesses offered good opportunity to explore my options.", "overall": 5.0, "summary": "Handy starter pics", "unix Review Time": 1387497600, "review Time": "12 20, 2013"}, {"reviewer ID": "A13IKQCJKFAP5S", "asin": "B0055V7UR0", "reviewer Name": "applegd07", "helpful": [0, 0], "review Text": "great picks. i like the feel and the control it provides. i like its texture and the size. just perfect.", "overall": 5.0, "summary": "perfect picks", "unix Review Time": 1380931200, "review Time": "10 5, 2013"}, {"reviewer ID": "A3BMI7VGJT60Y7", "asin": "B0055V7UR0", "reviewer Name": "Autumn", "helpful": [1, 1], "review Text": "I don't know how much you can say about picks, but this is a versatile pack. You get a multitude of types and there are 2 of each thickness. If you use a variety of picks or are wanting to test out other types of picks, this is for you.", "overall": 5.0, "summary": "Nice variety", "unix Review Time": 1384387200, "review Time": "11 14, 2013"}]

4. RESULTS

The proposed methodology was implemented in java programming language and taking JSON data from Amazon.com as dataset for the customer reviews on the musical instruments sold via Amazon. The methodology has been implemented in such a way that it can be worked with any dataset no matter what format it is specified in. The only change needed is a change in connection details. Also, we can implement dataset from any other source and from any other domain, but a change in domain requires updating of domain specific positive and negative words.

The data shown above clearly demonstrates that the accuracy achieved via proposed methodology is 94.005% which is very perfect by any standards. The results show that calculated value varies at few places in comparison to user defined value mainly because users tend to give ratings numerically in discrete values but the words explain their experience better. The proposed method evaluates the sentiment on basis of text mining and then calculates the rating based on that. The two values when compared reveal that they are almost 94.005% similar.

The output CSV file produced which becomes the input file for WEKA tool is shown below. This csv contains three fields viz. calculated value, actual value and error.

Table 1: Results for calculated review rating from proposed methodology and User defined value

2.428571429,3,0.19047619
4.19047619,4,0.047619048
4.666666667,4,0.166666667
1.916666667,2,0.041666667
3.333333333,5,0.333333333
4.761904762,5,0.047619048
4.5,4,0.125
3.333333333,5,0.333333333
4.333333333,5,0.133333333
4.333333333,5,0.133333333
4.166666667,5,0.166666667
4.333333333,5,0.133333333
5,5,0
4.333333333,3,0.444444444
4.666666667,4,0.166666667
5,5,0
5,5,0
4.583333333,5,0.083333333
3.333333333,5,0.333333333
3.888888889,5,0.222222222
5,5,0
4.095238095,5,0.180952381
4.428571429,4,0.107142857
3.333333333,5,0.333333333
3.666666667,3,0.222222222
3.333333333,5,0.333333333
3.333333333,5,0.333333333
2.666666667,3,0.111111111
4.666666667,4,0.166666667
4.18627451,3,0.395424837

The above output is converted to ARFF format to be fed into the WEKA tool for SVM and Naïve Byes classification.

SCREENSHOTS:

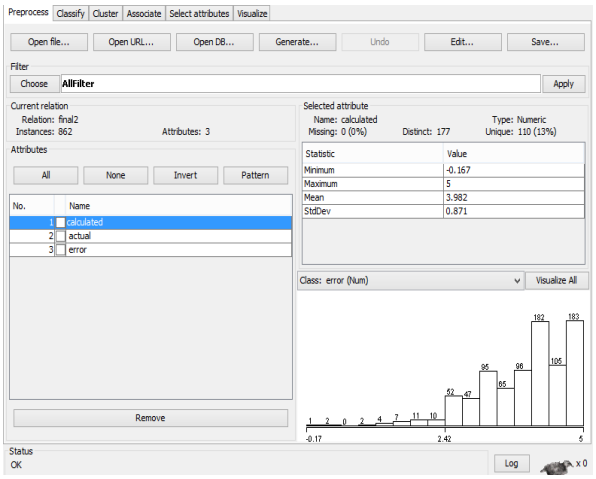

Figure 2: WEKA output for input CSV file for calculated values

The generated CSV file as a result of program execution over available dataset, when fed into the WEKA classifier loads the data and displays the above analysis of input data before classification. The above graph shows that in sample data, the no. of positive reviews outweighs the no. of negative reviews.

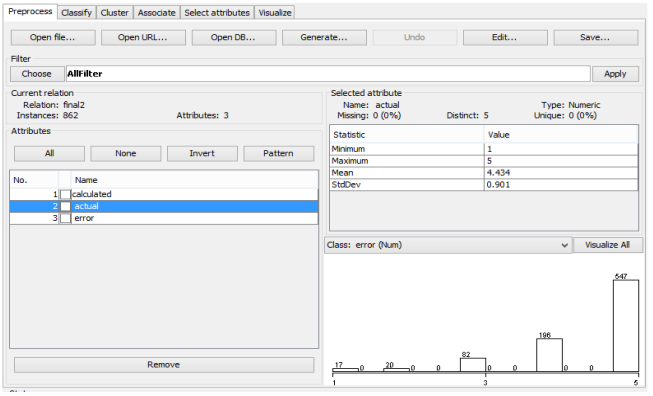

Figure 3: WEKA output for input CSV file for actual values

The results above as shown in graph confirm similarity in pattern if not in numbers. The majority of reviews are positive for all these reviews in consideration.

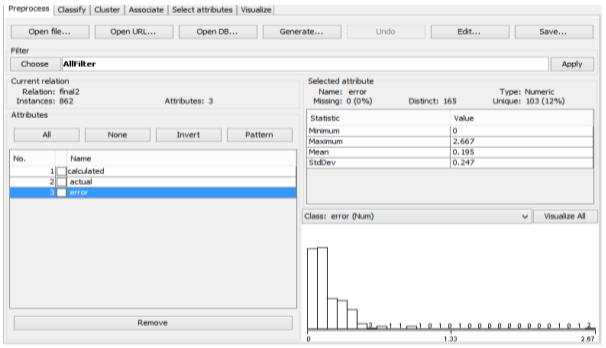

Figure 4: WEKA output for input CSV file for calculated values

The above results for the error value between calculated and actual review score clearly indicate the effectiveness of proposed technique. The majority of reviews when checked for error, reveal that the calculated and actual scores are similar if not same. This clearly speaks about the accuracy of proposed technique.

TABLE 4.2: Result for WEKA classifier for SVM technique using SMOreg filter function

=== Run information ===
Scheme:weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegS

```
MOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
Relation:     final2
Instances:    862
Attributes:   3
              calculated
              actual
              error
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
SMOreg
weights (not support vectors):
 -       0.3789 * (normalized) calculated
 +       0.0347 * (normalized) actual
 +       0.3453
Number of kernel evaluations: 371953 (90.598% cached)
Time taken to build model: 0.41 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient                  0.5257
Mean absolute error                      0.0774
Root mean squared error                  0.2125
Total Number of Instances                862
```

The results above are an indicator for difference between the actual and calculated values. The correlation coefficient of 0.5257 depicts similarity between opinions. The differences can be due to the interpretations. The words understanding by mind and machine cant be similar. The RMS value of 0.2125 indicates strong similarity between results.

TABLE 4.3: Result for WEKA classifier for Naïve Bayes technique

```
=== Run information ===
Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     final2
Instances:    862
Attributes:   4
              calculated
              actual
              error
              result
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
                Class
Attribute            B       A       D       C       E
                (0.33)  (0.32)  (0.13)  (0.11)  (0.11)
=================================================================
calculated
  mean              4.1892  4.6134  3.2132  3.5457  2.8618
  std. dev.         0.6109  0.5609  0.3088  0.5032  1.0185
  weight sum           283     277     110      97      95
  precision         0.0294  0.0294  0.0294  0.0294  0.0294
0.0294
 actual
  mean              4.4558  4.6787  4.8182  4.2062  3.4421
  std. dev.         0.7524  0.5585  0.4086  0.9299  1.5198
  weight sum           283     277     110      97      95
  precision            1       1       1       1       1

 error
  mean              0.1413  0.0167  0.3342  0.2347  0.6562
  std. dev.         0.0217  0.0282  0.0271  0.0214  0.4532
  weight sum           283     277     110      97      95
  precision         0.0163  0.0163  0.0163  0.0163  0.0163
Time taken to build model: 0.02 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         790               91.6473 %
Incorrectly Classified Instances        72                8.3527 %
Kappa statistic                          0.8895
Mean absolute error                      0.0349
Root mean squared error                  0.1476
Relative absolute error                 11.6522 %
Root relative squared error             38.1634 %
Total Number of Instances              862
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.912     0.021       0.956     0.912     0.933     0.994    B
                 0.928     0           1         0.928     0.963     0.997    A
                 0.936     0.02        0.873     0.936     0.904     0.997    D
                 0.928     0           1         0.928     0.963     0.995    C
                 0.863     0.059       0.646     0.863     0.739     0.989    E
Weighted Avg.    0.916     0.016       0.93      0.916     0.921     0.995
=== Confusion Matrix ===
   a   b   c   d   e   <-- classified as
 258   0   0   0  25 |   a = B
  11 257   0   0   9 |   b = A
   0   0 103   0   7 |   c = D
   1   0   2  90   4 |   d = C
   0   0  13   0  82 |   e = E
```

The results above obtained by running the Naïve Bayes algorithm on the output CSV file generated as a result of proposed methodology are a clear evidence for effectiveness of proposed method. The classifier states that the classification of results by proposed method, are accurate by almost 92% which differs slightly from our programmatic evaluation but still perfect byb any standards. The RMS value of 0.1476 again exemplifies the same fact. The last

confusion matrix generated again shows that majority of results fall under a or b classification which means the error rate is very low almost zero.

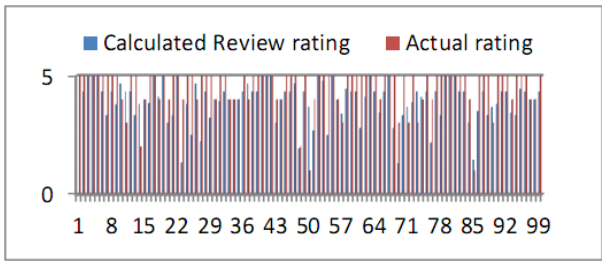

Chart 1: Chart between calculate review rating and actual rating for 1st 100 reviews

The above chart between calculated review rating by proposed methodology and actual rating given by the customer indicates the similarity in both values. The values differ at few places where the customer rating doesn't match the feedback in words. The comparison reveals 94% accuracy in proposed methodology results.

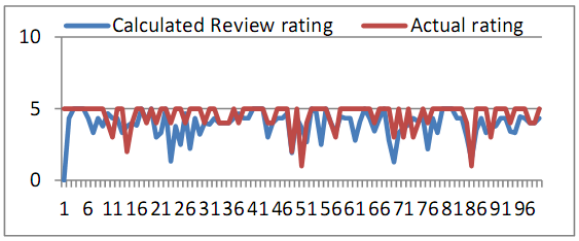

Chart. 2: Line graph between proposed technique results and customer given ratings.

The line graph shown above displays the variation between ratings calculated via proposed technique and the ratings provided by the customer. The values entered by the customer are discrete values which most of the time don't show the exact sentiment. So, the proposed technique calculates the rating based on the sentiment expressed by the user in his reviews. The variation at many a places is mainly due to the concerns expressed by user in words but still given ratings differently.

5. CONCLUSION

In today's world, the online market is increasingly getting popular and it becomes more and more important to help the customer get the best product by all parameters. The quality of a product is best confirmed by taking the customer reviews from those who are already using that. All popular shopping websites like Amazon, flipkart, ebay etc allow customer reviews once the product has been purchased. These reviews are such huge in numbers on these websites that it is not possible for a customer to consider them all. In this research we use machine learning classification techniques like Naïve Bayes and SVM(Support vector machines) to create a sentiment classification system with high degree of accuracy. The proposed work uses text mining techniques like Stanford parser, Sentiword Net and Wordnet 2.1 to parse and extract the sentiment from the reviews in the dataset. This research uses dataset from Amazon.com for musical instruments. The dataset is in JSON parser. The results received from the implementation of proposed technique ascertain the effectiveness of methodology. The computed results when compared to results obtained from Amazon using SVM and Naïve Bayes classifiers confirm that the proposed technique has better performance than the base research by Rushleen et al., who could best achieve 80% accuracy with the dataset adopted against 94.005% achieved by proposed technique. The results from Naive Bayes were found to be better and more explanatory for the inputted data.

The proposed methodology can be used in other domains like social network data to classify the intent and sentiment of an end user. This can be helpful to put come accounts under scrutiny who consistently post some objectionable data. This can be a huge support to identify criminals and terrorists who use social networks for spreading hatred or their messages.

ACKNOWLEDGEMENT

I owe my special thanks to the almighty and then to my guide Er. Ms. Richa Dutta who helped me to accomplish this research work successfully. I am also grateful to my family and friends for all their kind and unconditional support.

REFERENCES

Arindam Chaudhuri and Soumya K. Ghosh (2016). "Sentiment Analysis of Customer Reviews Using Robust Hierarchical Bidirectional Recurrent Neural Network", Advances in Intelligent Systems and Computing 464, Springer 2016.

Ayu Purvarianti (2011). "A Non Deterministic Indonesian Stemmer", ICEEI, IEEE 2011.

Chandhana Surbhi (2013) in "Natural language processing future", International conference on optical imaging sensor, 2013.

Deepshikha Chaturvedi and Shalu Chopra (2014). "Customers Sentiment on Banks", IJCA Vol. 98 No. 13, July 2014. Pg 8-13.

Deepshikha Chaturvedi and Shalu Chopra (2014). "Customers Sentiment on Banks", IJCA Vol. 98 No. 13, July 2014. pp. 8-13.

Fisnik Kastrati, Xiang Li, Christoph Quix and Mohammadreza Khelghati (2011). "Enabling Structured Queries over

Unstructured Documents", IEEE National conference on Mobile Data management, 2011.

https://nlp.stanford.edu/software/lex-parser.html

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Huang Zou, Xiunhuai Tiang, Bini Xie and Bin Liu (2015). "Sentiment analysis with machine learning techniques with syntax features", International Conference on Computational Science and Computational Intelligence IEEE 2015.

Ibrahim Eldesoky Fattoh, Amal Elsayed Aboutabl and Mohamed Hassan (2014). "Tapping into the Power of Automatic Question Generation", IJCA Vol. 103, No. 1, Oct. 2014. pp. 1-6.

James Mountstephens (2013). "Mnemonic phrase generation using genetic algorithms and Natural language processing", IEEE 2013, pp. 527-530.

Kumar Ravi and Vadla Mani Ravi (2015). "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", Knowledge-Based Systems 89-91, Elsevier 2015.

Lucas Povoda, Radim Burget and Malay Kishore Dutta (2016). "Sentiment analysis based on machine learning and received data."IEEE 2016.

M. Kasthuri and S. Britto Ramesh Kumar (2014). "An Improved Rule based Iterative Affix Stripping Stemmer for Tamil Language using K-Mean Clustering", IJCA Vol. 94, No. 13, May 2014. pp. 36-41.

Maryam, Seema Koulkur (2014) in "Feature ranking in sentiment analysis", International Journal of Computer Applications, 2014.

Mukta Takalikar, Manali Kshirsagar and Gauri Dhopvakar (2013). "Intuitive Approaches for Named Entity Recognition and Classification: A survey", IJCA 2013, pp. 35-38.

Neha Raghuvanshi and J. M. Patil (2016). "A brief review on Sentiment Analysis", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – IEEE 2016.

Nie, Liu, Wang, Song (2013) in "The Opinion Mining Based on Fuzzy Domain Sentiment Ontology Tree for Product Reviews" Journal of SW Vol. 8 No.11, 2013.

Oliver Keszocze, Mathias Soeken, Eugen Kuksa and Rolf Drechsler (2013). "lips: An IDE for Model Driven Engineering Based on Natural Language Processing", IEEE 2013. pp. 31-38.

Oscar Romero Llombart (2017). "Using Machine learning techniques for Sentiment Analysis", School of engineering, UAB, Barcelona, IEEE 2017.

Roul, Devanand, Sahay (2014) in "Web Document Clustering and Ranking using Tf-Idf based Apriori Approach", International Conference on Advances in Computer Engineering & Applications, 2014.

Rui Xia, Chengqing, Zong and Shoushan Li (2011). "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences Elsevier 2011. pp. 138-1152.

Rushlene, Navneet, Ravneet and Gurpreet (2016). "Opinion mining and sentiment analysis", IEEE 2016.

S. Samudaria and S. Sasirekha (2011). "Improving the Precision Ratio Using Semantic Based Search", ICSCCN, IEEE 2011, pp. 465-470.

Saani H. and Reghu Raj P. C. (2013). "Structured Information Extraction from On-line Advertisements- A Bayesian Approach", IJARCSSE Sep. 2013. pp. 581-586.

Sheetal Pereira, Uday Joshi (2014) in "Implementation of SVM technique in feedback analysis system", International journal of computer applications, 2014.

Simran Fitzgerald, George Mathews, Colin Morris and Oles Zhulyn (2012). "Using NLP techniques for file fragment classification", Digital Investigation Elsevier 2012. pp. S44-S49.

Tian Xia (2011) in "Improved VSM text classification by title vector based document representation method", ICCSE Pg 210-213, 2011.

Xueying Zhang, Chunju Zhang, Chaoli Du and Shaonan Zhu (2011). "SVM based Extraction of Spatial Relations in Text", IEEE 2011, pp. 529-533.

Corresponding Author

Gurjeet Kaur*

Research Scholar, M.Tech (CSE) Yamuna Group of Institutes, Yamuna Nagar