Disquisition of Sentiment Inquiry with Hashing and Counting Vectorizer using Machine Learning Classification

Kota Venkateswara Rao, M. Shyamala Devi

Abstract*: With the rapid growth in technology, analysis of feedback and reviews by the customers in companies and industries becomes a major challenge. The profit of the company mainly depends on the customer satisfaction. The view of the customer can be analyzed only through feedback. The review analysis can be utilized for the prediction of current sales and future sales of the company. With this overview, the paper aims in performing the sentiment analysis of the movie review. The Type of comment given by the customer is predicted and categorized into classes. The sentiment Analysis on movie Review dataset taken from the KAGGLE leading Dataset repository is used for implementation. The categorization of sentiment classes is achieved in five categories. Firstly, the target count for each sentiment is portrayed. The Resampling is done for equalizing the target sentiment count. Secondly, the extraction of sentiment feature words for each target is displayed and the data cleaning is done with Term Frequency Inverse document Frequency method. Thirdly, the resampled dataset is then fitted with the various classifiers like Multinomial Naives Bayes Classifier, Logistic Regression Classifier, KNearest Neighbors Classifier, Bernoulli Naives Bayes Classifier, Complement Naives Bayes Classifier, Nearest Centroid Classifer, Passive Aggressive Classifier, SGD Classifier, Ridge Classifier, Perceptron Classifier. Fourth, the feature extraction is done with Hashing Vectorizer and Counting Vectorizer. The vocabulary features are also displayed from the dataset. Fifth, the Performance analysis of clasifier is done with metrics like Accuracy, Recall, FScore and Precision. The implementation is carried out using python code in Spyder Anaconda Navigator IP Console. Experimental results shows that the sentiment prediction and classification done by Ridge classifier is found to be effective with Precision of 0.89, Recall of 0.88, FScore of 0.87 and Accuracy of 89%.

***Index Terms*: Accuracy, Recall, FScore, Sentiment and Precision**

I. INTRODUCTION

Sentiment processing and analysis is a major task in the field of Prediction of language processors [1]. Nowadays companies are moving forward to spend for feedback analysis of their customers and employees for forecasting the company revenue and turnover. with the technological growth, the customers are giving their feedback in the social media which is viewed and influenced by the people all over the country. So analyzing the online reviews of their products becomes a challenging issue for the manufacturing companies. The purpose of the sentiment analysis is to analyze and determine the type of sentiment of the text review messages given by the customers. Each text given by the customers are highly sensitive with non standard grammatical text structures and with low or high integrity. So determining the polarity and the sensitiveness of the feedback review is a challenging task. This makes the application of machine learning and natural language processing to come into picture for predicting the sentiment analysis on movie reviews.

The paper is prepared with the Section 2 exploring the literature survey and related works. Section 3 deals with the proposed work continued with execution details and performance comparison in Section 4. Finally the paper is concluded in Section 5.

II. RELATED WORK

A. Literature Review

The sentiment classification method mainly focuses on machine learning methods along with the natural language processing. The large scale sentiment dictionary is designed to enhance the sentiment classification method and performance. The machine learning methods were used based on the parsing and sentiment information to predict the sentiment type [2].

The phrase level emotion detection model is designed to predict the implicit emotion detection and sentiment classification [3]. The Support vector machine is used along with the probability output text weighting in order to enhance the sentiment classification accuracy metric [4].

The LSTM (Long Short-Term Memory) is a kind of neural network which can detain long term text dependencies in a sentence sequence. This is done by designing the block of storage units and it is used to update the block information in the storage and to make the permanent memory thereby enhancing the depth calculation of the sentiment tree analysis [5]. The Long Short-Term Memory combined with the target dependent variable is used to improve the prediction accuracy [6]. The concept of forward tree-structured long short-term memory networks is designed to enhance the sentiment classification [7].

Revised Manuscript Received on November 06, 2019.

Kota Venkateswara Rao, Research Scholar**,** Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

M. Shyamala Devi, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

With the enormous applications of machine learning techniques, the deep neural network combined with attention model to improvise the sentiment prediction analysis by focusing on classification relationships and text evaluation [8]. A neural architecture is proposed to exploit the instant available sentimental lexicon resources like lexicon-driven contextual attention and contrastive co-attention to improve the sentimental classification accuracy [9]. A capsule network is designed with activity vector denoting the sentiment instantiation parameters. Single lower level capsule tends to send its corresponding output to higher level capsules having a large scalar product and sentiment prediction from the lower level capsule [10]. Three methods were proposed to enhance the dynamic routing process to improve the interruption of some error capsule that may contain redundant information or it would have not been trained properly. This makes the sentiment prediction in capsule network with high precision and accuracy [11]. An effective routing technique that effectively reduces the computational complexity in case of multiple sentiment analysis data sets [12]. The attention mechanism is incorporated with the capsule sentiment network for the extraction of relation in a multi-label learning framework [13]. The Capsule network for sentiment type analysis is done in domain adaptation scenario with adaptation of semantic rules to improve the comprehensive sentence representation learning [14]. The concept of classification and its methodology is learnt [15]-[33].

III. PROPOSED WORK

A. Preliminaries

B. Term Frequency Inverse Document Frequency

There are various methods to find the TFDF and it has two metrics namely Term Frequency and Inverse document frequency. The Term frequency represents the amount of time a particular term appears on a page by dividing it with number word in a text document and it is given below.

$$Term\ Frequency = \frac{1 + \log(Keyword\ Count)}{log(total\ word\ count\ in\ the\ document)}$$

The formula for finding the inverse document frequency is given below.

$$IDF = \log of \frac{1 + \text{Total Documents}}{Documents\ with\ keywords}$$

C. Proposed System Architecture

The overall framework of this paper is shown in Fig. 1

IV. PROPOSED WORK

In this work, the counting vectorizer and hashing vectorizer are used to predict the sentiment type of the people using their movie reviews. Our contribution of this work is pointed out here.

(i) Firstly, the target count for each sentiment is portrayed. The Resampling is done for equalizing the target sentiment count.

(ii) Secondly, the extraction of sentiment feature words for each target is displayed and the data cleaning is done with Term Frequency Inverse document Frequency method.

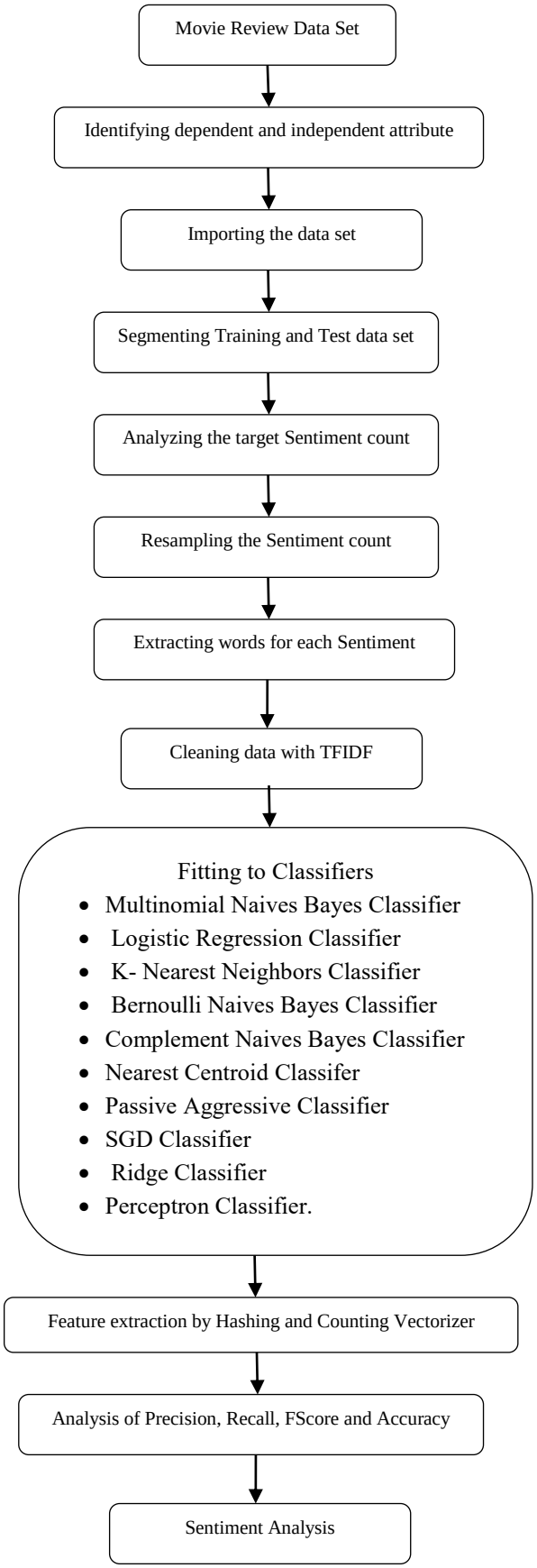

Fig. 1 System Architecture of Sentiment Analysis

(iii) Thirdly, the resampled dataset is then fitted with the various classifiers like Multinomial Naives Bayes Classifier, Logistic Regression Classifier, KNearest Neighbors Classifier, Bernoulli Naives Bayes Classifier, Complement Naives Bayes Classifier, Nearest Centroid Classifer, Passive Aggressive Classifier, SGD Classifier, Ridge Classifier, Perceptron Classifier.

(iv) Fourth, the feature extraction is done with Hashing Vectorizer and Counting Vectorizer. The vocabulary features are also displayed from the dataset.

(v) Fifth, the execution assessment of clasifier is done with metrics like Accuracy, FScore, Recall and Precision

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Sentiment Analysis and Prediction

The Sentiment Analysis for Movie Reviews Dataset extracted from Kaggle ML dataset warehouse is used for implementation with 2 independent variable and 1 Sentiment dependent variable. The dataset consists of 1,56,060 individual's data. The attribute are shown below.

1. Sentence Id
2. Phrase
3. Sentiment - Dependent Attribute

The type of sentiment of the dependent variable is shown below.

B. Performance Analysis

The target count for each sentiment dependent variable of the Sentiment Analysis of the Movie Review dataset is shown in the fig 2.

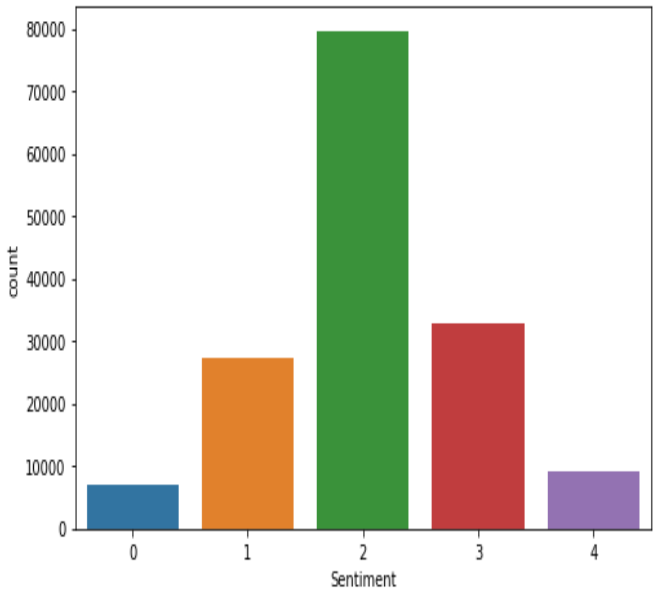

Fig. 2. Target count of sentiment dependent variable

The resampling of the target count for each sentiment dependent variable of the Sentiment Analysis of the Movie Review dataset is done and is shown in fig 3.

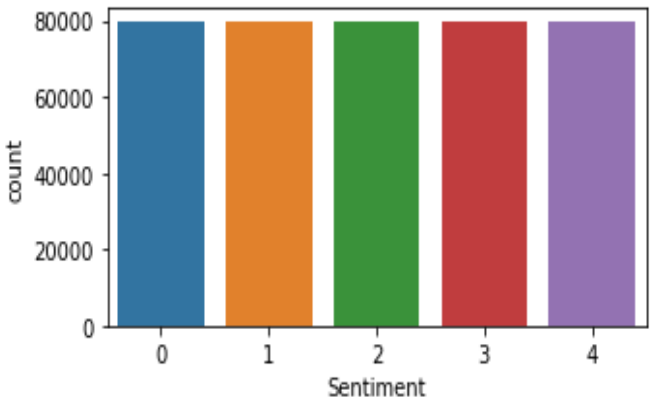

Fig. 3. Resampling of Target count of sentiment dependent variable

The extraction of sentiment feature words for negative target is displayed and is shown in fig 4.

Fig. 4. Sentiment feature words for negative target

The extraction of sentiment feature words for somewhat negative target is displayed and is shown in fig 5.

Fig. 5. Sentiment feature words for somewhat negative target

The extraction of sentiment feature words for neutral target is displayed and is shown in fig 6.

Fig. 6. Sentiment feature words for neutral target

The extraction of sentiment feature words for somewhat positive target is displayed and is shown in fig 7.

Fig. 7. Sentiment feature words for somewhat positive target

The extraction of sentiment feature words for positive target is displayed and is shown in fig 8.

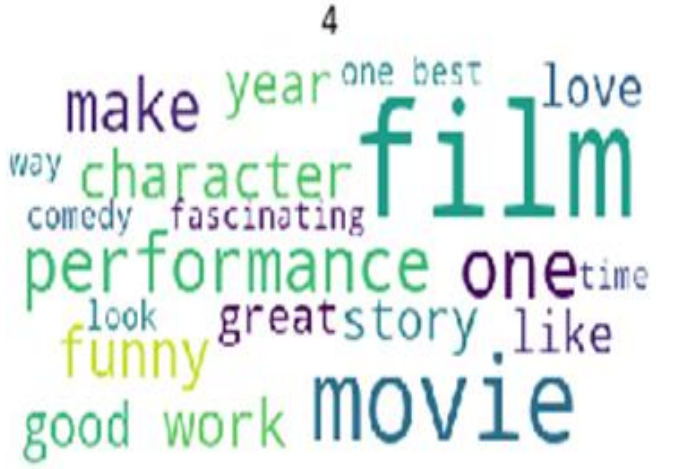

Fig. 8. Sentiment feature words for positive target

The starting and ending time of Term Frequency and inverse document frequency is shown in Fig. 9.

end	float	1	1563088823.119122
expected	Series	(39015,)	Series object of pandas.core.series module
predicted	int64	(39015,)	[1 3 2 ... 3 3 2]
start	float	1	1563088812.6347218

Fig. 9. Starting and ending time of TFIDF

The resampled dataset is then fitted with the various classifiers like Multinomial Naives Bayes Classifier, Logistic Regression Classifier, KNearest Neighbors Classifier, Bernoulli Naives Bayes Classifier, Complement Naives Bayes Classifier, Nearest Centroid Classifer, Passive Aggressive Classifier, SGD Classifier, Ridge Classifier, Perceptron Classifier. The obtained confusion matrix for each of the classifiers is shown from Fig. 10 – Fig. 19.

cm_MultinomialNB_Feature - NumPy array

	0	1	2	3	4
0	99	865	810	19	1
1	37	2152	4560	171	0
2	5	821	17858	1072	5
3	0	73	4450	3615	85
4	0	6	662	1434	215

Fig. 10. Confusion Matrix of Multinomial NBayes Classifier

cm_logreg_Feature - NumPy array

	0	1	2	3	4
0	293	815	656	29	1
1	125	2128	4460	198	9
2	22	851	17850	1003	35
3	3	103	4339	3525	253
4	0	12	520	1274	511

Fig. 11. Logistic Regression Confusion Matrix

cm_KNN_Feature - NumPy array

	0	1	2	3	4
0	403	500	877	14	0
1	273	1712	4821	106	8
2	75	914	18028	706	38
3	14	149	5931	1851	278
4	1	19	1278	643	376

Fig. 12. KNN Confusion Matrix

cm_BernoulliNB_Feature - NumPy array

	0	1	2	3	4
0	416	743	610	24	1
1	348	2437	3879	235	21
2	118	1239	16929	1353	122
3	15	227	4301	3163	517
4	3	14	735	1038	527

Fig. 13. Bernoulli N Bayes Classifier Confusion Matrix

cm_ComplementNB_Feature - NumPy array

	0	1	2	3	4
0	745	817	195	28	9
1	821	3797	1867	374	61
2	1161	3282	11240	3462	616
3	132	435	1977	4645	1034
4	12	29	168	1117	991

Fig. 14. Confusion Matrix of Complement NBayes Classifier

cm_NearestCentroid_Feature - NumPy array

	0	1	2	3	4
0	803	585	273	89	44
1	844	2871	2482	476	247
2	712	2779	13551	1979	740
3	120	618	3103	3010	1372
4	13	94	350	694	1166

Fig. 15. Confusion Matrix of Nearest Centroid Classifier

cm_PassiveAggressiveClassifier_Feature - NumPy array

	0	1	2	3	4
0	726	821	232	13	2
1	630	3270	2762	222	36
2	226	1860	15627	1876	172
3	40	256	3001	3880	1046
4	5	22	220	1006	1064

Fig. 16. Passive Aggressive Confusion Matrix

cm_SGDClassifier_Feature - NumPy array

	0	1	2	3	4
0	125	156	1495	18	0
1	62	367	6359	123	9
2	19	178	19213	315	36
3	3	32	6908	1154	126
4	0	6	1502	523	286

Fig. 17. SGD Confusion Matrix

cm_RidgeClassifier_Feature - NumPy array

	0	1	2	3	4
0	595	902	272	24	1
1	404	3139	3139	222	16
2	113	1478	16553	1554	63
3	10	195	3276	4152	590
4	3	14	259	1231	810

Fig. 18. Ridge Confusion Matrix

cm_PerceptronClassifier_Feature - NumPy array

	0	1	2	3	4
0	716	791	237	49	1
1	735	3153	2570	438	24
2	308	2160	14131	3010	152
3	34	295	2724	4338	832
4	7	42	225	1198	845

Fig. 19. Perceptron Confusion Matrix

The performance metrics and its prediction analysis is shown in the Table 1 - Table. 2. The parameters that are used to measure the performance of the sentiment type prediction is shown in Fig. 22 – Fig. 25.

Table. 1 Analysis of Precision and Recall Score parameters

Classifier	Precision	Recall
Multinomial NBayes Classifier	0.74	0.73
Logistic Regression Classifier	0.81	0.80
K Nearest Neighbors Classifier	0.84	0.82
Bernoulli Naives Bayes Classifier	0.79	0.78
Complement N Bayes Classifier	0.76	0.75
Nearest Centroid Classifer	0.77	0.76
Passive Aggressive Classifier	0.87	0.86
SGD Classifier	0.86	0.84
Ridge Classifier	0.89	0.88
Perceptron Classifier	0.79	0.77

The Starting and ending time for the hashing vectorizer is shown in Fig. 20.

```
In [22]: print("Starting Processing time = ",start)
    ...: print("End Processing time = ",end)
Starting Processing time =  1563124772.7563655
End Processing time =  1563124783.7343657

In [23]: print("HashingVectorizer finished in: ", end - start)
HashingVectorizer finished in:  10.978000164031982
```

Fig. 20. Processing time of hashing vectorizer

The Starting and ending time for the Counting vectorizer is shown in Fig. 21.

```
In [88]: end = time.time()
    ...: print("Starting Processing time = ",start)
    ...: print("End Processing time = ",end)
    ...: print("CountVectorizer finished in: ", end - start)
Starting Processing time =  1563126651.1747735
End Processing time =  1563126655.4137735
CountVectorizer finished in:  4.239000082015991
```

Fig. 21. Processing time of Counting vectorizer

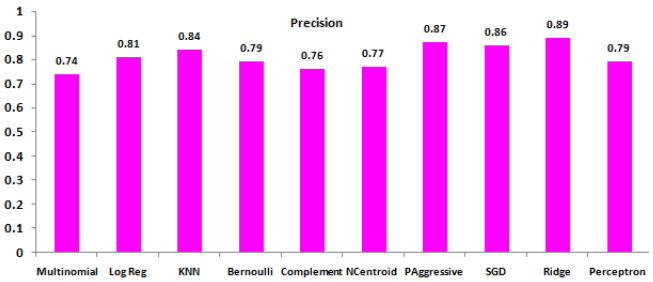

Fig. 22. Precision of Classifiers

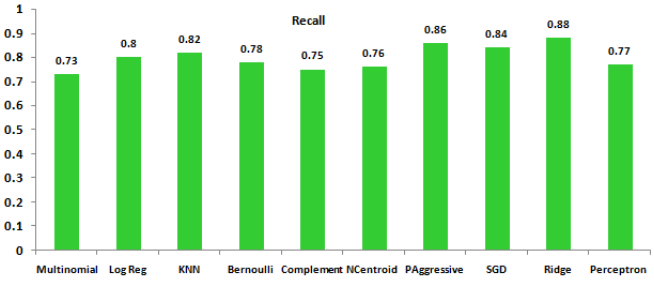

Fig. 23. Recall of Classifiers

Table. 1 Analysis of FScore parameters

Classifier	FScore	Accuracy (%)
Multinomial Naives Bayes Classifier	0.72	74
Logistic Regression Classifier	0.80	80
K Nearest Neighbors Classifier	0.83	84
Bernoulli Naives Bayes Classifier	0.77	78
Complement Naives Bayes Classifier	0.76	77
Nearest Centroid Classifer	0.75	76
Passive Aggressive Classifier	0.84	83
SGD Classifier	0.85	84
Ridge Classifier	0.87	89
Perceptron Classifier	0.76	73

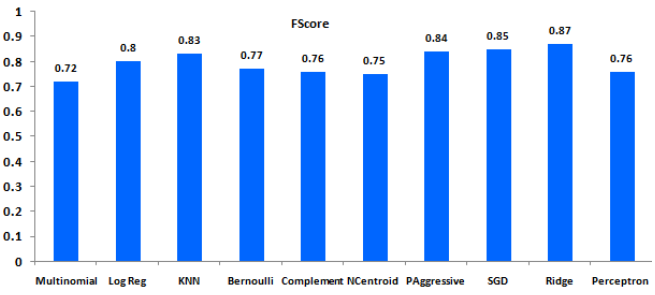

Fig. 24. FScore of Classifiers

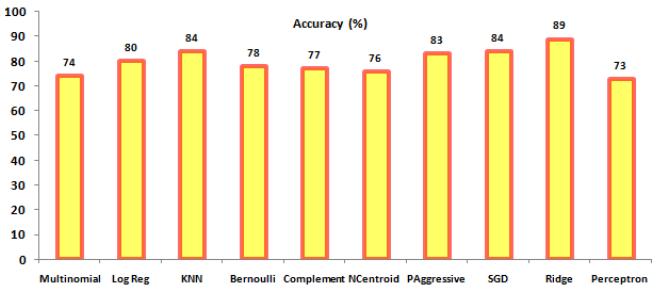

Fig. 25. Accuracy of Classifiers

VI. CONCLUSION

This paper analyzes the sentiment type prediction by using various classification algorithms. The extraction of feature words for each of the sentiment target variable is found and is displayed. The Resampling of the dataset is also done based on the target variable. Experimental results shows that the sentiment prediction and classification done by Ridge classifier is found to be effective with Precision of 0.89, Recall of 0.88, FScore of 0.87 and Accuracy of 89%.

REFERENCES

1. B. Liu, ``Sentiment analysis and opinion mining," Synth. Lectures Hum.Lang. Technol., vol. 5, no. 1, pp. 1_167, May 2012.
2. Z. Y. Yan, Q. Bin, S. Q. Hui, and L. Ting, ``Large-scale sentiment lexicon collection and its application in sentiment classification," J. Chin. Inf. Process., vol. 31, no. 2, pp. 187-193, 2017.
3. Odbal and Z. F. Wang, ``Emotion analysis model using Compositional
4. Semantics," Acta Automatica Sinica, vol. 41, no. 12, pp. 2125_2137, 2015.
5. P. Li,W. Xu, C. Ma, J. Sun, and Y. Yan, ``IOA: Improving SVM based sentiment classification through post processing," in Proc. 9th Int. Workshop Semantic Eval., Jun. 2015, pp. 545_550.
6. J. Xu, D. Chen, X. Qiu, and X. Huang, ``Cached long short-term memory neural networks for document-level sentiment classification," in Proc. Conf. Empirical Methods Natural Lang. Process., Nov. 2016, pp. 1660_1669.
7. D. Tang, B. Qin, X. Feng, and T. Liu, ``Effective LSTMs for target dependent sentiment classification," in Proc. COLING 26th Int.l Conf.
8. Comput. Linguistics, Tech. Papers, Dec. 2015, pp. 3298_3307.
9. K. S. Tai, R. Socher, and C. D. Manning, ``Improved semantic representations from tree-structured long short-term memory networks," in Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Nat-ural Lang. Process., Jul. 2015, pp. 1556_1566.
10. P. Zhou et al., ``Attention-based bidirectional long short-term memory
11. networks for relation classification," in Proc. 54th Annu. Meeting Assoc.
12. Comput. Linguistics, Aug. 2016, pp. 207_212.
13. A. Galassi, M. Lippi, and P. Torroni. (Feb. 2019). ``Attention, please!Acritical review of neural attention models in natural language processing." [Online]. Available: https://arxiv.org/abs/1902.02181
14. S. Sabour, N. Frosst, and G. E. Hinton. (2017).``Dynamic routing between capsules." [Online]. Available: https://arxiv.org/abs/1710.09829
15. W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, and Z. Zhao. (2018). ``Investigating capsule networks with dynamic routing for text classification." [Online]. Available: https://arxiv.org/abs/1804.00538
16. J. Kim, S. Jang, S. Choi, and E. Park. (2018). ``Text classification using
17. capsules." [Online]. Available: https://arxiv.org/abs/1808.03976
18. N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, and H. Chen, ``Attention based capsule networks with dynamic routing for relation extraction," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Nov. 2018, pp. 986_992.
19. B. Zhang, X. Xu, M. Yang, X. Chen, and Y. Ye, ``Cross-domain sentiment classification by capsule network with semantic rules," IEEE Access, vol. 6, pp. 58284_58294, Oct. 2018
20. M. Shyamala Devi, Shakila Basheer, Rincy Merlin Mathew, "Exploration of Multiple Linear Regression with Ensembling Schemes for Roof Fall Assessment using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019.
21. Shakila Basheer, Rincy Merlin Mathew, M. Shyamala Devi, "Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 127-133.
22. Rincy Merlin Mathew, M. Shyamala Devi, Shakila Basheer," Exploration of Neighbor Kernels and Feature Estimators for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 597-605.

23. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Nariboyena Vijaya Sai Ram, "Backward Eliminated Formulation of Fire Area Coverage using Machine Learning Regression", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp.1565-1569
24. M. Shyamala Devi, Ankita Shil, Prakhar Katyayan, Tanmay Surana, "Constituent Depletion and Divination of Hypothyroid Prevalance using Machine Learning Classification", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 1607-1612
25. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Sairam Kondapalli, "Recognition of Forest Fire Spruce Type Tagging using Machine Learning Classification", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, pp. 4309 – 4313, 16 September 2019.
26. M. Shyamala Devi, Usha Vudatha, Sukriti Mukherjee, Bhavya Reddy Donthiri, S B Adhiyan, Nallareddy Jishnu, " Linear Attribute Projection and Performance Assessment for Signifying the Absenteeism at Work using Machine Learning", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, pp. 1262 – 1267, 16 September 2019.
27. M. Shyamala Devi, Mothe Sunil Goud, G. Sai Teja, MallyPally Sai Bharath, "Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.11, pp. 604 – 609, 30 September 2019.
28. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, " Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 952 – 956, 30 August 2019.
29. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 957 – 961, 30 August 2019.
30. Rincy Merlin Mathew, R. Suguna, M. Shyamala Devi, "Composite Model Fabrication of Classification with Transformed Target Regressor for Customer Segmentation using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 962 – 966, 30 August 2019.
31. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 4800-4807.
32. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 6198-6203.
33. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, , vol. 22, no. 4, 25 June 2019, pp. 729-739. DOI: 10.1080/09720510.2019.1609729ISSN: 0972-0510 (Print), 2169-0014 (Online).
34. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739. DOI: 10.1080/09720510.2019.1609729ISSN: 0972-0510 (Print), 2169-0014 (Online).
35. Shyamala Devi Munisamy, Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification", Learning and Analytics in Intelligent Systems, LAIS, Springer, vol. 3, pp. 604-612, June 2019.
36. Suguna Ramadass, Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers", Springer's book series entitled "Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 613-620, June 2019.
37. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
38. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.