Penguin Search Optimization Based Feature Selection for Automated Opinion Mining

T. Anuprathibha, C. S. KanimozhiSelvi

Abstract: ***Twitter sentiment analysis is a vital concept in determining the public opinions about products, services, events or personality. Analyzing the medical tweets on a specific topic can provide immense benefits in medical industry. However, the medical tweets require efficient feature selection approach to produce significantly accurate results. Penguin search optimization algorithm (PeSOA) has the ability to resolve NP-hard problems. This paper aims at developing an automated opinion mining framework by modeling the feature selection problem as NP-hard optimization problem and using PeSOA based feature selection approach to solve it. Initially, the medical tweets based on cancer and drugs keywords are extracted and pre-processed to filter the relevant informative tweets. Then the features are extracted based on the Natural Language Processing (NLP) concepts and the optimal features are selected using PeSOA whose results are fed as input to three baseline classifiers to achieve optimal and accurate sentiment classification. The experimental results obtained through MATLAB simulations on cancer and drug tweets using k-Nearest Neighbor (KNN), Naïve Bayes (NB) and Support Vector Machine (SVM) indicate that the proposed PeSOA feature selection based tweet opinion mining has improved the classification performance significantly. It shows that the PeSOA feature selection with the SVM classifier provides superior sentiment classification than the other classifiers.***

Keywords: ***Natural Language Processing, Opinion mining, Penguin Search Optimization algorithm, Sentiment analysis, Twitter.***

I. INTRODUCTION

Opinion mining is the detection and investigation of the sentiments of the contents in text data [1]. The content mostly utilized is the raw online text data that are exchanged by social media users. Twitter is a primary social medium platform with significant source of short texts that can be utilized in analyzing the public opinions. Social media like twitter and Facebook has vital part in assessing the performance of personality, products, services and events [2]. The opinions of public users through Twitter and other social media platforms are considered as the golden mines of business enterprises. The data mining application is utilized in extracting the reviews of movies, products, politics, etc. and analyzes the dynamics towards the resulting opinions. This provides an insight about their product or any type of interlinking process with the users. Using these insights, the organizations or personalities can alter their products and services to create major interest of users towards them [3]. These benefits drive the research community to develop efficient opinion mining approaches to generate accurate and faster opinion analysis.

Opinion mining is an automation methodology of extracting the tweet opinions with the help of NLP techniques [4]. The feature selection process is performed using many criteria. Some works utilized statistical measures and similarity measures to select the features while some used the sentiment scores based on lexicons [5]. Recent works employed optimization and other machine learning algorithms to select the optimal features. The other major process is the classification stage. Many works have utilized the Naïve Bayes, SVM and k-NN algorithms for classification [6], [7]. However these classifiers have weaknesses on selection of features and hence the feature selection techniques are utilized into the classifiers to provide high accuracy of sentiment classification [8].

Many optimization algorithms have been successfully employed for feature selection in tweet data namely genetic algorithm, Shuffled frog leaping algorithm (SFLA) and modified SFLA [9]. This paper aims at developing an efficient opinion mining framework using PeSOA [10] based feature selection and SVM, NB and KNN algorithms for classification. The utilization of the PeSOA into the classifiers improves the overall accuracy with less convergence rate. This article is ordered with section 2 presenting discussion of related works. Section 3 presents the PeSOA based opinion mining system whose evaluation results are illustrated in section 4. Section presents a conclusion of this paper.

II. RELATED WORKS

Many researchers developed advanced techniques in the recent past for feature selection in opinion mining processes. Onan & Korukoğlu, [11] introduced genetic based feature selection for improved text sentiment classification. This approach aggregates the several individual feature lists and selects the optimal features. Gupta et al, [12] used PSO for feature selection for aspect based sentiment analysis. Li et al, [13] proposed a global optimization approach using PSO to provide multi-polarity sentiment analysis.

Manuscript published on 30 September 2019
* Correspondence Author

T. Anuprathibha*, Research scholar, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode, Tamil Nadu- 638060, India. Email: jb.anuprathibha@gmail.com.

Dr. C. S. KanimozhiSelvi, Associate Professor, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Tamil Nadu- 638060, India. E-mail: kec.kanimozhi@gmail.com.

It uses the information gain with the SVM for improved classification accuracy. However, the time complexity is high when the data size increases. Shang et al, [14] utilized PSO for feature selection in sentiment classification that resolves the limitations of binary PSO such as irrational update formula of velocity.

Liu et al, [15] proposed a feature selection approach using the multi-swarm PSO for sentiment analysis to reduce redundancy of text features and improve the classification accuracy. Akhtar et al, [16] developed multi-objective optimization for aspect based sentiment analysis. This approach also uses SVM and CRF for aspect based sentiment classification. However, this approach has limitations in convergence rate. Similarly, Akhtar et al, [17] also utilized single objective PSO for feature selection in a two-step method for aspect based sentiment analysis. Kumar & Khorwal, [18] employed the firefly algorithm for feature selection for decreasing the computational complexity and feature set size.

Tubishat et al, [19] developed an improved whale optimization algorithm for feature selection in Arabic sentiment analysis. This approach utilized information gain for feature reduction and SVM for classification with high accuracy and less execution time. Iqbal et al, [20] developed a hybrid sentiment analysis framework using genetic algorithm and increased the accuracy. Ahmad et al, [21] proposed ant colony optimization (ACO) for feature selection using a wrapper approach with integrated ACO for feature selection and KNN for classification. Kumar et al, [22] proposed binary cuckoo search for feature selection and employed the TF-IDF weighting schemes and SVM classifier to utilize the optimal features for enhancing the sentiment analysis accuracy.

Though the methods in literature provide accurate sentiment analysis, there are limitations in the suggested optimization algorithms. The major limitations are the slow convergence and pre-mature convergence due to one-dimensional search process. This paper aims at overcoming these limitations by developing an opinion mining framework for medical tweets using PeSOA based feature selection. This approach has better convergence rate and avoids the pre-mature convergence.

III. MEDICAL TWEET OPINION MINING FRAMEWORK USING PESOA FEATURE SELECTION

The recommended opinion mining framework has been aimed for the enhancement of the sentiment analysis from medical tweets. Fig.1 shows the architecture flow of the recommended opinion mining framework with the PeSOA based feature selection technique. The approach utilizes the Twitter API for data collection. The input data are collected from Twitter API using the keywords related to cancer and drugs. 6400 tweets are collected using different types of cancer keywords and 500 tweets based on drugs. In training phase, around 2500 tweets are used and the remaining tweets are used for testing. These data are pre-processed and then the features are extracted using feature descriptors. Then the features are selected using the PeSOA algorithm. Lastly three classifiers are utilized to appraise the classification accuracy in opinion mining based on these selected features.

The pre-processing of the input data collected from Twitter API is achieved by performing many tasks including stemming and stop word removal steps. The extraction of features namely the content words, function words, POS tags and POS n-grams features is done to develop the classifier [23]. The content words are the nouns and the words that define independent meaning in a sentence while the function words express grammar relationships between words but no meaning when considered separately. POS tags utilize the Noun, verb, pronoun, adverb and adjectives to formulate features. The Part of speech n-grams are generated upto trigrams have been utilized. The features are also utilized in combinations of content words + function words, function words + POS n-grams, and content words + function words + POS n-grams. The next step is the feature selection. The procedure of feature selection using PeSOA is explained in this section.

A. Feature selection using PeSOA

Feature selection is the practice of choosing one or a set of aspects that provide the best sentiment classification results. The best option currently is to model the feature selection problem into NP-hard optimization problem and resolve it using advanced optimization algorithms. The PeSOA algorithm selects the optimal features in this work which are ranked using the information gain matric. PeSOA is based on the fish hunting behavior of penguin groups in ice holes. The penguins forms themselves into certain number of groups and each group randomly searches for fishes until their oxygen reserves are exhausted. Then they restore the oxygen and search again till they find sufficient fish. Then they share the food location with other groups and compare to select the best location for hunting. This process is adapted for PeSOA and utilized for optimal selection of features.

First, the population of penguins is initialized and the initial oxygen reserves and other parameters are set. Then the penguins are assembled into smaller groups and each group travels themselves towards one of the available food locations. The mapping of PeSOA to the feature selection problem must be performed in adequate manner. The random population of penguin solutions are set as features and the groups are deliberated as feature subsets. For each feature, the fitness is computed and the optimal values decide the best solution. The other solutions move towards the best solutions. This measure is expressed as

$$X_{new} = X_{old} + rand \times (X_{l\ best} - X_{l\ old}) \tag{1}$$

Where X_{new} is the new solution, X_{old} is the old solution, $X_{l\ best}$ is the local best solution, $X_{l\ old}$ is the former local best solution and $rand$ is a arbitrary number between [0,1].

After each plunge of the penguins, the oxygen backup of the penguin is updated using Eq. (2).

$$O_j^i(t+1) = O_j^i(t) + (f(X_{new}) - f(X_{old})) \times |X_{new} + X_{old}| \tag{2}$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

where $O_j^i(t+1)$ is the new oxygen reserve, $O_j^i(t)$ is the last oxygen reserve and *f* is the objective function formulated based on error rate function.

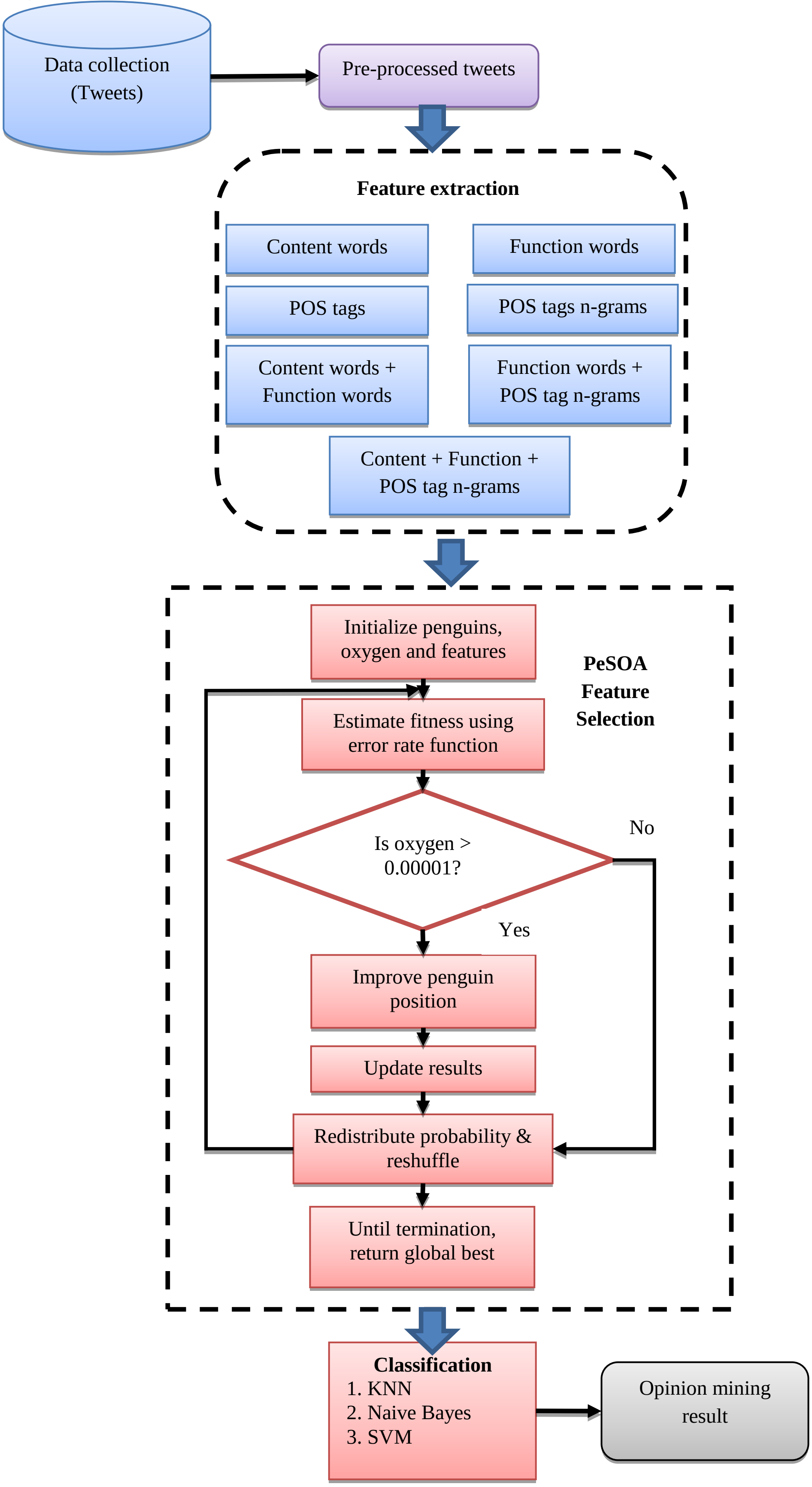

Fig.1. Architecture of PeSOA feature selection based opinion mining framework

Similarly, the Quantity of Eaten Fish (QEF) and the group membership of penguins are also updated. The QEF is expressed as the energy content of the food.

$$QEF^{i}(t+1) = QEF^{i}(t) + \textstyle\sum_{j=1}^{n}\left(O_j^{i}(t+1) + O_j^{i}(t)\right) \quad (3)$$

The affiliation update of penguins is done by estimating a likelihood P_i of linking to the group i.

$$P_i(t+1) = \frac{QEF^{i}(t)}{\sum QEF^{i}(t)} \quad (4)$$

Once the update processes are completed, the global best features are returned. The complete procedures in PeSOA for feature selection are précised as follows.

Pseudo code of the PeSOA based feature selection:

```
Begin
Read the pre-processed tweet data
Set amount of features (penguin population size)
Create arbitrary population of P results (penguins) in clusters;
Compute the error rate (fitness function) for each penguin;
For i=1 to number of generations;
For each penguin i $\in$ P do
Initialize the oxygen reserve for each penguin
While oxygen backups are not exhausted do (until 0.00001)
Take an arbitrary step.
Advance the penguin location using Eq. (1)
Update the local best solution;
Update oxygen backup using Eq. (2);
End while
End for
Update amounts of eaten fish in the holes using Eq. (3).
Update group membership using Eq. (4).
Reorganizes the likelihoods of penguins in holes and levels
Eliminate the groups with no members;
Update best-solution;
End
```

The feature subsets nominated in this process are deliberated as the best solutions. The information gain metric ranks the selected features and decrease the number of feature subsets. After the selection of optimal feature subsets, the classification is done using k-NN, NB and SVM. The classification process of these classifiers is upgraded using the PeSOA. The investigational results are conducted to validate the performance of the classifiers.

IV. PERFORMANCE EVALUATION

The evaluation of the recommended opinion mining structure is assessed in MATLAB tool using the collected cancer and drug datasets of varying data sizes. The comparison of the PeSOA feature selection based classifiers is performed to determine the efficient classifier for this medical tweet opinion mining work. The performance metrics used are processing time, accuracy, precision, recall, and f-measure for the two datasets- cancer and drugs.

Table I shows the accuracy assessment of the PeSOA feature selection based classifiers. The results show that the PeSOA-SVM classifier improves the sentiment analysis of both cancer and drug datasets with high accuracy than PeSOA-KNN and PeSOA-NB classifiers. In view of 5000 tweets in cancer dataset, the accuracy of PeSOA-SVM is 78.8% which is superior to the other compared classifiers. The main reason for this improvement is due to the better convergence and improved feature selection by PeSOA.

Table II illustrates the precision assessment of the PeSOA feature selection based classifiers. The results show that the PeSOA-SVM classifier has high precision than PeSOA-KNN and PeSOA-NB classifiers. For 5000 tweets in cancer dataset, the precision of PeSOA-SVM is 79% which is larger than the other compared classifiers.

Table III illustrates the recall comparison of the PeSOA feature selection based classifiers. The results show that the PeSOA-SVM classifier has high recall than PeSOA-KNN and PeSOA-NB classifiers. For 5000 tweets in cancer dataset, the recall of PeSOA-SVM is 69% which is better than the other compared classifiers. The superior performance of PeSOA-SVM depends on the optimal feature subset selection with less error.

Table IV displays the F-measure assessment of the PeSOA feature selection based classifiers. The results show that the PeSOA-SVM classifier has high F-measure than PeSOA-KNN and PeSOA-NB classifiers. Considering 5000 tweets in cancer dataset, the F-measure of PeSOA-SVM is 81.9% which is superior to the other compared classifiers.

Table V displays the processing time (measured in seconds) evaluation of the PeSOA feature selection based classifiers. The convergence speed of PeSOA is higher and hence the processing time is much lesser. The results indicate that the PeSOA-SVM classifier has less processing time than the other classifiers. This performance is mainly due to the adaptability of SVM with PeSOA for larger data sizes due to the multi-dimensional search process.

From the comparisons, it can be established that the recommended opinion mining structure using PeSOA feature selection and SVM classification has better performance which is demonstrated through better values of performance metrics. This confirms that the PeSOA algorithm offers significantly improved sentiment analysis performance.

Table-I: Accuracy (%) comparison

Methods	Cancer					Drugs				
	1000	*2000*	*3000*	*4000*	*5000*	*100*	*200*	*300*	*400*	*500*
PeSOA-KNN	78.2	78.1	78.5	78.1	77.8	79.2	79.1	79.1	78.9	78.8
PeSOA-NB	78.6	78.6	78.8	78.3	78.1	79.4	79.4	79.4	79.1	79.1
PeSOA-SVM	**79.4**	**79.5**	**79.4**	**79.0**	**78.8**	**79.7**	**79.6**	**79.6**	**79.3**	**79.3**

Table-II: Precision (%) comparison

Methods	Cancer					Drugs				
	1000	*2000*	*3000*	*4000*	*5000*	*100*	*200*	*300*	*400*	*500*
PeSOA-KNN	79.1	78.9	78.8	78.4	78.1	78.9	78.8	78.8	78.8	78.6
PeSOA-NB	79.4	79.1	79.1	78.8	78.8	79.1	79.1	79.1	79.1	79.0
PeSOA-SVM	**79.6**	**79.3**	**79.3**	**79.0**	**79.0**	**79.3**	**79.3**	**79.3**	**79.3**	**79.1**

Table-III: Recall (%) comparison

Methods	Cancer					Drugs				
	1000	*2000*	*3000*	*4000*	*5000*	*100*	*200*	*300*	*400*	*500*
PeSOA-KNN	69.1	68.9	68.8	68.4	68.1	77.8	77.8	76.4	78.1	78.0
PeSOA-NB	69.4	69.1	69.1	68.8	68.8	78.1	78.1	76.8	78.8	78.5
PeSOA-SVM	**69.6**	**69.3**	**69.3**	**69.0**	**69.0**	**78.8**	**78.8**	**77.0**	**79.0**	**79.0**

Table-IV: F-measure (%) comparison

Methods	Cancer					Drugs				
	1000	*2000*	*3000*	*4000*	*5000*	*100*	*200*	*300*	*400*	*500*
PeSOA-KNN	80.8	80.8	80.7	80.5	80.5	81.3	81.4	81.5	81.4	81.3
PeSOA-NB	81.4	81.2	81.1	81.0	81.1	82.5	82.5	82.5	82.5	82.4
PeSOA-SVM	**82.5**	**82.3**	**82.2**	**81.9**	**81.9**	**82.9**	**82.9**	**82.7**	**82.6**	**82.7**

Table-V: Processing time (seconds) comparison

Methods	Cancer					Drugs				
	1000	*2000*	*3000*	*4000*	*5000*	*100*	*200*	*300*	*400*	*500*
PeSOA-KNN	10.453	14.54	17.86	20.66	24.21	1.10	2.89	3.42	4.98	5.83
PeSOA-NB	10.321	14.11	17.45	20.41	24.05	0.98	2.76	3.32	4.72	5.65
PeSOA-SVM	**10.119**	**13.99**	**17.22**	**20.07**	**23.89**	**0.85**	**2.62**	**3.26**	**4.61**	**5.32**

V. CONCLUSION

Twitter opinion mining is highly undertaken for the business ventures to identify the customer mindset about their products and reviews. Providing highly accurate sentiment classification is mainly depended on the feature selection and classification processes. The feature selection technique is focused in this article to improve the performance of medical tweet opinion mining. PeSOA based selection of optimal features and feature reduction using the information gain metric is performed while classification using three classifiers. The evaluation results proved that the PeSOA has improved the classification performance and PeSOA-SVM classifier has superior sentiment classification results. In future, the convergence rate of PeSOA will be further improved to avoid stagnation of the algorithm in local optimum solutions.

REFERENCES

1. X. Zhou, X. Tao, J. Yong and Z. Yang, “Sentiment analysis on tweets for social events,” In Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 557-562, 2013.
2. R. Gaspar, C. Pedro, P. Panagiotopoulos and B. Seibt, “Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events,” Computers in Human Behavior, vol. 56, pp. 179-191, 2016.
3. G. Beigi, X. Hu, R. Maciejewski and H. Liu, “An overview of sentiment analysis in social media and its applications in disaster relief,” In Sentiment analysis and ontology engineering, Springer, Cham, pp. 313-340, 2016.

4. P. H. Shahana and B. Omman, “Evaluation of features on sentimental analysis,” Procedia Computer Science, vol. 46, pp. 1585-1592, 2015.
5. A. Jurek, M. D. Mulvenna and Y. Bi, “Improved lexicon-based sentiment analysis for social media analytics,” Security Informatics, vol. 4, no. 1, p. 9, 2015.
6. S. V. Wawre and S. N. Deshmukh, “Sentiment classification using machine learning techniques,” International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 819-821, 2016.
7. D. N. Devi, C. K. Kumar and S. Prasad, “A feature based approach for sentiment analysis by using support vector machine,” In 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 3-8, 2016.
8. B. M. Jadav and V. B. Vaghela, “Sentiment analysis using support vector machine based on feature selection and semantic analysis,” International Journal of Computer Applications, vol. 146, no. 13, 2016.
9. T. Anuprathibha and C. S. KanimozhiSelvi, “Medical Opinion from Twitter: Automating Social Media Opinions for Health Informatics,” Journal of Medical Imaging and Health Informatics, vol. 6, no. 8, pp. 2005-2011, 2016.
10. Y. Gheraibia and A. Moussaoui, “Penguins search optimization algorithm (PeSOA),” In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, Berlin, Heidelberg, pp. 222-231, 2013.
11. A. Onan and S. Korukoğlu, “A feature selection model based on genetic rank aggregation for text sentiment classification,” Journal of Information Science, vol. 43, no. 1, pp. 25-38, 2017.
12. D. K. Gupta, K. S. Reddy and A. Ekbal, “PSO-ASENT: Feature selection using particle swarm optimization for aspect based sentiment analysis,” In International conference on applications of natural language to information systems, Springer, Cham, pp. 220-233, 2015.
13. X. Li, J. Li and Y. Wu, “A global optimization approach to multi-polarity sentiment analysis,” PloS one, vol. 10, no. 4, p. e0124672, 2015.
14. L. Shang, Z. Zhou and X. Liu, “Particle swarm optimization-based feature selection in sentiment classification,” Soft Computing, vol. 20, no. 10, pp. 3821-3834, 2016.
15. Z. Liu, S. Liu, L. Liu, J. Sun, X. Peng and T. Wang, “Sentiment recognition of online course reviews using multi-swarm optimization-based selected features,” Neurocomputing, vol. 185, pp. 11-20, 2016.
16. M. S. Akhtar, S. Kohail, A. Kumar, A. Ekbal and C. Biemann, “Feature selection using multi-objective optimization for aspect based sentiment analysis,” In International Conference on Applications of Natural Language to Information Systems, Springer, Cham, pp. 15-27, 2017.
17. M. S. Akhtar, D. Gupta, A. Ekbal and P. Bhattacharyya, “Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis,” Knowledge-Based Systems, vol. 125, pp. 116–135, 2017.
18. A. Kumar and R. Khorwal, “Firefly algorithm for feature selection in sentiment analysis”. In Computational Intelligence in Data Mining, Springer, Singapore, pp. 693-703, 2017.
19. M. Tubishat, M. A. Abushariah, N. Idris and I. Aljarah, “Improved whale optimization algorithm for feature selection in Arabic sentiment analysis,” Applied Intelligence, pp. 1-20, 2018.
20. F. Iqbal, , J. M. Hashmi, B. C. Fung, R. Batool, A. M. Khattak, S. Aleem and P. C. Hung, “A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction,” IEEE Access, vol. 7, pp. 14637-14652, 2019.
21. S. R. Ahmad, A. A. Bakar and M. R. Yaakub, “Ant colony optimization for text feature selection in sentiment analysis,” Intelligent Data Analysis, vol. 23, no. 1, pp. 133-158, 2019.
22. A. Kumar, A. Jaiswal, S. Garg, S. Verma and S. Kumar, “Sentiment analysis using cuckoo search for optimized feature selection on Kaggle tweets,” International Journal of Information Retrieval Research (IJIRR), vol. 9, no. 1, pp. 1-15, 2019.
23. A. Bell, J. M. Brenier, M. Gregory, C. Girand and D. Jurafsky, “Predictability effects on durations of content and function words in conversational English,” Journal of Memory and Language, vol. 60, no. 1, pp. 92-111, 2009.

AUTHORS PROFILE

T. Anuprathibha has obtained her Bachelor of Computer Science at Sri Saradha Niketan College of Science for Women from Bharathidasan University in 2000.She obtained her Master of Computer Applications at Periyar Maniammai College of Technology For Women from Bharathidasan University in 2003. She has obtained her Master of Computer Science & Engineering at M. Kumarasamy College of Engineering from Anna University Coimbatore in 2011. Currently she is a research scholar at Kongu Engineering College, pursuing her Ph.D. in the area of opinion mining under the guidance of Dr. C. S. KanimozhiSelvi.

Dr. C. S. Kanimozhi Selvi is a faculty member in the Department of Computer Science and Engineering of Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India. She received her Bachelor’s degree in Computer Science in 1994 from Bharathiar University and a Master’s degree in Computer Applications from Bharathiar University in 1998. Then, she received her Master’s degree M.E., in Computer Science and Engineering from Anna University, Chennai in 2004. She has completed her Ph.D. in 2011. She has been in the teaching profession for the past 19 years. Her areas of academic interest include data mining, database management systems and cloud computing. She has published 20 articles in international journals and more than 30 papers in international and national conferences.

Retrieval Number: B2629078219/19©BEIESP
DOI:10.35940/ijrte.B2629.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication