International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Enhanced Medical Tweet Opinion Mining using Improved Dolphin Echolocation Algorithm Based Feature Selection

T. Anuprathibha, C. S. KanimozhiSelvi

Abstract*: Extraction and analysis of public opinions from social network data can provide interesting outcomes and inferences about product, service, event or personality. Twitter is one of the most popular medium for analyzing the public sentiment through user tweets. Feature specific opinion analysis provides highly accurate and effective classification and categorization of public opinions. This paper focuses on developing an opinion mining framework for automated analysis of tweet opinions using efficient feature selection and classification algorithms. For this purpose, an Improved Dolphin Echolocation Algorithm (IDEA) is developed by enhancing the optimization performance of the Dolphin Echolocation Algorithm (DEA). The limitations of DEA are the insufficient exploration and exploitation properties in local optimum solutions and also impact the convergence rate. These shortcomings are overcome by the proposed IDEA algorithm. In this work, first the tweets are collected and pre-processed to extract the features using Part-of-Speech (POS) tagging and n-grams aided by a dictionary. Using IDEA, the feature subset candidates are selected and the outcomes are fed as input to the baseline classifiers to obtain highly accurate opinion classification. The evaluation of the k-Nearest Neighbor (KNN), Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers using the two feature selection approaches of DEA and IDEA are performed over cancer and drug tweets datasets and the results illustrate that the classification accuracy of opinions is enhanced significantly through the IDEA based feature selection than the traditional DEA algorithm. These results justify the utilization of the proposed IDEA algorithm for improving the opinion mining applications in different fields.

***Index Terms*: Dolphin Echolocation Algorithm, Feature specific opinion analysis, Improved Dolphin Echolocation Algorithm, Opinion mining, sentiment analysis, Support Vector Machine.**

I. INTRODUCTION

Sentiment analysis from the social networking data is an interesting research topic as it improves the accuracy and effectiveness of public opinion mining [1]. In recent years, the trend of extracting public opinions from popular social media like Facebook and Twitter has been increasing rapidly. The social media platforms are utilized by the people for communicating their opinions and experiences from which the organizations and business ventures can gather vital information that help in improving their business markets [2]. The business profiles disseminate the knowledge about their products and services from the social media, rather than traditional survey models, for better business development. The digital marketing concept has revolutionized the promotion of contents to larger audiences through specialized recommendations of the products and services. Multiple organizations compete to extract knowledge from various sources and utilize them to gain the customer attention [3]. The sentiment analysis of user opinions from social media takes center stage in these functions.

Twitter is the most popular micro-blogging site and the opinion mining from the user tweets is one of the most interesting approaches of obtaining public opinions [4]. The main reason for the suggestion of utilizing Twitter is due to its higher usage for business services increases the knowledge extent compared to Facebook and LinkedIn. However, the higher usage of Twitter also increases the risks in extraction of genuine tweets since the spam tweets are most common [5]. A statistic report has shown that only around 1% of the total Twitter user accounts are verified genuine accounts for major public figures and organizations. This shows that the majority number of accounts is not verified and may be forged or spam created for the purpose of content sharing and marketing [6]. Hence the important challenge in tweet opinion mining has to be dealt with the spams. Therefore the sentiment analysis approach has to be highly advanced to tackle such challenges and provide better public opinion mining.

The three important categories of opinion mining that are utilized to extract public opinions from tweets are lexicon-based methods, machine learning-based methods, and hybrid methods [7]. Lexicon-based methods utilize predefined dictionaries to determine the sentiments of data [8] while the machine learning-based methods are most commonly applied with better results [9]. However, the lexicon-based methods provide less accurate results for short hand texts and the machine learning-based methods restrain from better opinion mining due to data size and unlabeled data. Recent researches utilized the hybrid methods to provide better opinion mining than individual lexicon-based and machine learning-based methods [10]. In this paper, an automated opinion mining framework is presented based on the hybrid opinion mining methods.

Manuscript published on 30 August 2019.
*****Correspondence Author(s)
T. Anuprathibha, Research scholar, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode, Tamil Nadu-638060, India.
Dr. C. S. KanimozhiSelvi, Associate Professor, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Tamil Nadu- 638060, India.

The proposed framework uses dictionary for sentiment determination, IDEA for feature selection and machine learning classifiers for sentiment classification. The performance of the proposed framework is evaluated and compared with state-of-the-art methods to illustrate its efficiency. This article is structured as: Section 2 offers a literature survey of recent related works. Section 3 demonstrates the proposed opinion mining approach using IDEA feature selection whose evaluation results are offered in section 4. Finally, the conclusion of this article along with future research scope is discussed in section 5.

II. RELATED WORKS

Latest research works highlighted the use of lexicon and machine learning techniques for twitter sentiment analysis. Stojanovski et al, [11] developed deep neural network architecture based sentiment analysis from tweets using automatically annotated dataset. This approach detects 7 emotions from the annotated dataset using unsupervised learning of deep neural networks. However this approach lacks the clear definition for detecting sarcasm tweets. Mostafa, [12] proposed a geo-located Twitter opinion polarity analysis to map halal food consumers. This analysis used expert-predefined lexicon of seed adjectives for determining the sentiments of the tweets. However, this method has limitations in extracting the geo-located tweets which are very minimal (around 2-5%) of the overall tweets. Zainuddin et al, [13] developed a hybrid sentiment classification approach of aspect-based twitter sentiment analysis and improved the accuracy of sentiment determination. However, the approach is limited to sub-par performance due to the complexity. Al-Thubaity et al, [14] utilized a sentiment lexicon in the analysis of Saudi dialect tweet opinion mining. This approach utilized modern standard Arabic text phrases along with the Arabic lexicon to improve the classification accuracy. However, the lexicon employed is very small and hence the performance is still questionable for larger lexicons. Rodrigues & Chiplunkar, [15] proposed a new big data approach for tweet sentiment analysis using a hybrid approach. The proposed Hybrid Lexicon-Naive Bayesian Classifier (HL-NBC) method is the combination of NBC with a popular lexicon to improve the classification accuracy of tweet opinions.

Alahmadi & Zeng, [16] proposed a Twitter-based recommender system to address cold-start using genetic algorithm based trust modeling for accurate sentiment analysis. This approach also utilizes the Support Vector Regression algorithm to predict the sentiments. However, this approach fails to effectively detect the sarcasm tweets. Yuvaraj & Sabari, [17] proposed a Twitter sentiment classification approach using binary shuffled frog algorithm. This approach utilized the feature selection technique with ensemble classifier to improve the classification accuracy. Kumar & Jaiswal, [18] presented an optimal feature selection approach based on swarm intelligence algorithms for improving the accuracy of sentiment prediction of tweets. This approach utilized binary grey-wolf and binary moth flame optimization algorithms for selecting optimal features and improved the sentiment classification accuracy. However, this approach has limitations in handling the larger data size. Tubishat et al, [19] developed a feature selection approach for sentiment analysis of Arabic text using an improved whale optimization algorithm (WOA). The authors improved the WOA using Elite Opposition-Based Learning (EOBL) at initialization phase and incorporated evolutionary operators at end of each iteration to improve the convergence rate. The information gain is used in SVM classifiers to reduce the search space and improve the sentiment prediction accuracy.

Goel & Garg, [20] proposed a sentiment analysis model using Gravitational Search Optimization (GSO) algorithm for social networking data. This model utilized GSO for feature extraction that improves the overall sentiment classification accuracy. Budhi et al, [21] proposed the use of multi-PSO based classifier for sentiment polarity prediction of user tweets. This approach employed the parallel processing technique by utilizing multiple PSO based classifiers for improved sentiment prediction. Packiam & Prakash, [22] developed a novel integrated framework using modular optimization for sentiment prediction from big tweet data. The authors employed modular optimization-based feature selection with multi-class SVM for achieving high accuracy of sentiment analysis with less execution time. However, this approach is most suited for only smaller data size.

From the literature, it can be understood that although the methods are aimed at supremacy in sentiment analysis of tweets there are many limitations and room for improvements. This paper focuses on developing a hybrid opinion mining framework to avoid those limitations and provide high performance in tweet sentiment analysis.

III. PROPOSED OPINION MINING FRAMEWORK

The proposed automated opinion mining framework is developed to improve the tweet opinion mining. The architecture of the proposed framework is given in Fig. 1. The opinion mining model utilizes general modeling steps of pre-processing, feature extraction and selection and classification on the tweets collected from Twitter through Twitter API based on specified keywords. First the tweets are preprocessed and the features are extracted by feature descriptors with the help of medilexicon dictionary. The features are selected using the DEA and IDEA algorithms based on which the sentiments of the tweets are classified.

Retrieval Number: J93170881019/19©BEIESP
DOI: 10.35940/ijitee.J9317.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

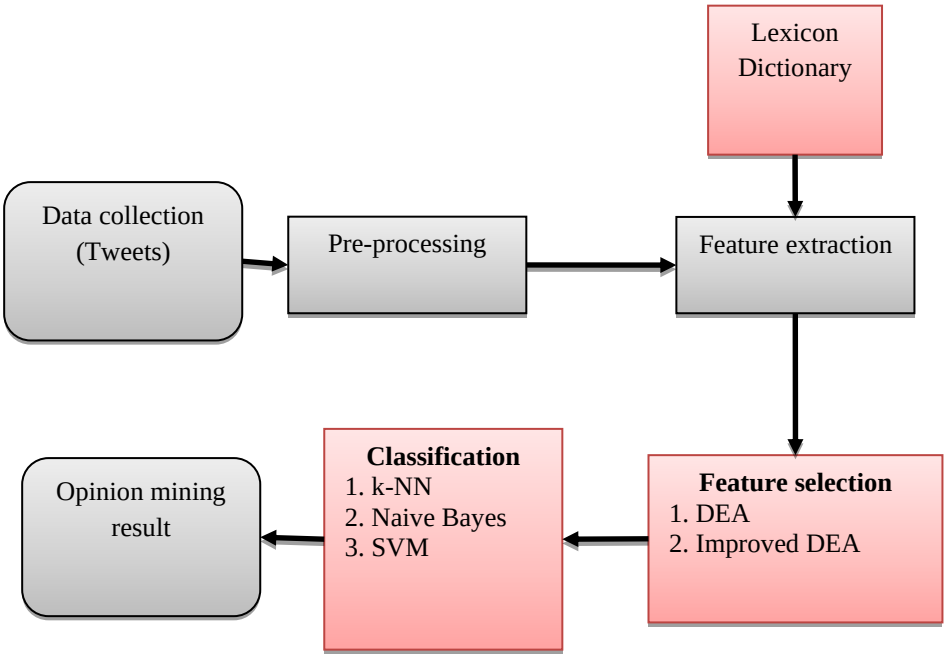

Fig.1. Automated opinion mining model

A. Data Collection

The tweets are gathered from Twitter API using the keywords related to cancer and drugs to form two medical datasets. The datasets contain 6400 tweets of cancer and 500 tweets on drugs respectively, from which 2500 data samples are employed in training the system before testing the remaining tweets. The proposed model has the ability to be updated with real-time inclusion of additional tweets.

B. Pre-processing

Pre-processing in the proposed opinion mining framework includes the steps of data cleaning, URL removal, duplicate tweet removal, spell checking, punctuation removal, tokenization, case normalization, stemming and stop word removal. These processes are performed to remove the irrelevant tweet data to filter and provide highly effective data as input to the system.

C. Lexicon dictionary and feature extraction

The lexicon dictionary enables the system to extract the sentiment words meanings from the tweets. The sentiment features from the tweets are mainly the nouns, pronouns and adjectives while the verb and adverbs provide additional information. As the tweets are medical related phrases, the online medilexicon dictionary from www.medilexicon.com is utilized to extract the phrase meanings to formulate the tweets. The content words, function words, POS tags and POS n-grams features are extracted to improve the classifier performance [23] which are also used in combinations of (content words & function words), (function words & POS n-grams), and (content, function words & POS n-grams).

D. Feature Selection

For the purpose of selecting the optimal features from the set of features extracted in the previous step, efficient optimization algorithms are used. The Dolphin Echolocation Algorithm and Improved Dolphin Echolocation Algorithm are used in this work. The DEA algorithm is one of the recent optimization algorithms which provide very effective performance. The features are mapped as food sources of dolphins and the best features are selected as the optimal feature subsets. However, the DEA also suffers from the common issue of optimization algorithms. The exploration and exploitation abilities may not be efficiently balanced in some cases causing slow convergence and the algorithm getting trapped in local optimum state. To overcome these limitations, the improved DEA is proposed by changing the bilinear coefficient function of DEA and also utilize probability based next step location determination to improve the convergence rate. The features are then ranked based on information gain metric for feature reduction and then the optimal feature subset is selected.

a) DEA feature selection

DEA is based on the characteristics of dolphins' ability to generate the special type of click sounds during the hunting process [24]. When the dolphins search for food, they generate clicks and once the clicks hit the prey their echoes are reflected back to the dolphin. Then by analysing the echoes, the location and the distance from the prey are estimated by the dolphins this ability is called as the dolphin echolocation based on which the optimization is carried out.

First the population of dolphins is initiated and for each feature, the alternatives of the search space are sorted either in ascending or descending order. Using this sorting method, for feature, vectors are created as the columns of alternatives matrix. Then the NL locations are selected for the dolphins randomly and the change of convergence factor is determined by the calculation of PP of the current loop.

$$PP(Loop_i) = PP_1 + (1 - PP_1)\frac{Loop_i^{power} - 1}{(Loops\ Number)^{power} - 1} \quad (1)$$

Where PP is the predefined probability, PP_1 is the convergence factor (CF) of the first loop, $Loop_i$ is the number of the current loop, power is the degree of the curve and $Loops\ Number$ is number of loops in which the algorithm should converge.

The fitness for each location is computed based on the error rate formula with 0.57 as the threshold value.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Then the accumulative fitness is computed according to dolphin rules for i-th location, j-th variable and $k = -R_e$ to R_e. Here the $Coefficient(k)$ is highlighted separately to make sense of modifying the convergence factor.

$$AF_{(A+k)j} = Coefficient(k) \times Fitness(i) + AF_{(A+k)j} \quad (2)$$

$$Coefficient(k) = (1 - \frac{|k|}{R_e}) \quad (3)$$

Where $AF_{(A+k)j}$ is the accumulative fitness of the $(A+k)$th alternative to be chosen for the j-th variable; R_e is the effective radius in which accumulative fitness of the alternative A's neighbors are affected from its fitness.

To distribute the possibility much evenly in search space, a small value is added to all the arrays as $AF = AF + \varepsilon$. Then the best location of current loop is found and the AF is set to zero. For variable j, calculate the probability $\overline{P_{ij}}$ of choosing alternative i according to the following relationship:

$$\overline{P_{ij}} = \frac{AF_{ij}}{\sum AF_{ij}} \quad (4)$$

In the final stage, the alternatives of variables with best locations are assigned with a probability equal to PP while the remaining locations are assigned with probability given by

$$P_{ij} = (1 - PP)\overline{P_{ij}} \quad (5)$$

This probability will help in determining the next step locations and finally the global best location is selected. As per the mapping of algorithm, this location is the high ranked feature based on the information gain. This feature subset will become the optimal feature. Algorithm 1 shows the steps involved in DEA based feature selection.

Algorithm 1: DEA based feature selection

1. Initiate NL locations for a dolphin randomly and sort alternatives of search space for creating Alternatives matrix.
2. Calculate the PP of the current loop using Eq. (1).
3. Estimate the fitness of each location.
4. Compute the accumulative fitness using Eq. (2).
5. Adjust the value of arrays to distribute the possibility of search space.
6. Determine the best location of current loop and let their $AF = 0$.
7. Calculate the probability $\overline{P_{ij}}$ for each variable.
8. Assign a probability equal to PP to all alternatives chosen for all best location variables and devote rest of the probability to the other alternatives using Eq. (5).
9. Calculate the next step locations according to the probabilities assigned to each alternative.
10. Repeat Steps 2–9 until a termination condition is met.

b) IDEA based feature selection

Although Dolphin echolocation algorithm provides efficient results, the exploration and exploitation abilities are not efficiently balanced and results in local optimum solutions. Also the convergence rate is slow as in other optimization algorithms. To overcome these limitations, two concepts are proposed to develop the improved DEA. First step is to improve the exploitation ability by changing the bilinear coefficient function $Coefficient(k)$ into a non-linear equation. The second step is to modify the estimation of next step location using the chaotic Gauss map concept to improve the convergence rate.

First the population of dolphins is initiated and the NL locations are randomly selected for each dolphin based on feature subsets to sort the alternatives and create the Alternatives matrix. Then the probability PP is computed and the fitness values based on error rate for each location is estimated. Then the bilinear coefficient function $Coefficient(k)$ is modified into a non-linear equation to support all direction movement for the dolphins in the feature search space.

$$Coefficient(k) = 1 - \frac{\sqrt{R_e^2 - (|k| - R_e)^2}}{R_e} \quad (6)$$

The non-linear property of the coefficient function enables the features to be compared in less iteration and improves the exploration. Using this $Coefficient(k)$, the AF estimation equation (2) in dolphin echolocation algorithm is modified into the following equation

$$AF_{(A+k)j} = \left(1 - \frac{\sqrt{R_e^2 - (|k| - R_e)^2}}{R_e}\right) \times Fitness(i) + AF_{(A+k)j} \quad (7)$$

Then the remaining processes are performed as in DEA until the best locations are obtained. In the penultimate step, the estimation of the next step locations is performed by using the chaotic Gauss map concept. It formulates the next step location determination by

$$x^{t+1} = \begin{cases} 0 & if\ x' = 0 \\ \frac{1}{x'} - \left[\frac{1}{x'}\right] & otherwise \end{cases} \quad (8)$$

"Where x' is a random number in (0,1) and x^{t+1} is a chaotic sequence in (0,1). This formulation provides the ability to select the next step locations in almost all possible alternatives and reduces the search time which improves the convergence speed. Algorithm 2 shows the steps involved in IDEA based feature selection.

Algorithm 2: IDEA based feature selection

1. Initiate NL locations for a dolphin randomly and sort alternatives of search space for creating Alternatives matrix.
2. Calculate the PP of the current loop using Eq. (1).
3. Estimate the fitness of each location.
4. Compute the accumulative fitness using Eq. (7).
5. Adjust the value of arrays to distribute the possibility of search space.
6. Determine the best location of current loop and let their $AF = 0$.
7. Calculate the probability $\overline{P_{ij}}$ for each variable using Eq. (4).
8. Assign PP probability to best locations and different probability to other locations using Eq. (5).

9. Calculate the next step locations using Eq. (8).
10. Repeat Steps 2–9 until a termination condition is met.

E. Sentiment Classification

The final stage of the proposed opinion mining framework is the classification of the sentiment polarity of the tweets. The baseline classifiers namely k-NN, NB and SVM are used for classification. The baseline classifiers are selected on the basis on more commonly used methods. The combination of the feature selection technique using DEA and the baseline classifiers is performed first to evaluate the classification performance due to DEA. Similarly, the IDEA based feature selection with the baseline classifiers is also performed.

IV. PERFORMANCE EVALUATION

The performance is evaluated in MATLAB tool using the collected cancer and drug related datasets in varying data sizes. The number of tweets in cancer dataset ranges from 1000 to 5000 while that in drug dataset ranges from 100 to 500. The comparison of the IDEA and DEA based baseline classifiers are performed to determine the efficiency in terms of accuracy, precision, recall, f-measure and processing time. Table 1 displays the accuracy evaluation of the DEA feature selection based classifiers with the IDEA feature selection based classifiers. IDEA-SVM outperforms other models; for instance when considering 5000 tweets in cancer dataset, the accuracy of IDEA-SVM is 96.58% which is greater than the other compared approaches. Similarly, for most data ranges in drug dataset, the IDEA-SVM has higher accuracy. Likewise, the IDEA based classifiers outperform their corresponding DEA based classifiers with better accuracy.

Table 2 demonstrates the precision evaluation of the DEA feature selection based classifiers with the improved DEA feature selection based classifiers. For most data ranges in cancer and drug dataset, the IDEA-SVM has better precision values with the IDEA based classifiers outperforming their corresponding DEA based classifiers. For instance, when considering 5000 tweets in cancer dataset, the precision of IDEA-SVM is 99.12% which is greater than the other compared approaches.

Table 3 illustrates the recall evaluation of the DEA feature selection based classifiers with the improved DEA feature selection based classifiers. For most data ranges in cancer and most data ranges in drug dataset, the IDEA-SVM has better recall values. For instance, considering 5000 tweets in cancer dataset, the recall of IDEA-SVM is 96.54% which is greater than the other compared approaches. The comparison between DEA and IDEA classifiers shows that the IDEA classifiers have higher recall values than their corresponding DEA classifiers.

Table 4 displays the F-measure evaluation of the DEA feature selection based classifiers with the improved DEA feature selection based classifiers. For most data ranges in cancer and drug dataset, the IDEA-SVM has better F-measure. For instance, when considering 4000 tweets in cancer dataset, the F-measure of IDEA-SVM is 97.5% which is greater than the other compared approaches. It is also noted that the IDEA classifiers perform better than their corresponding DEA classifiers with higher values of F-measure.

Table 5 shows the processing time (measured in seconds) evaluation of the DEA feature selection based classifiers with the improved DEA feature selection based classifiers. IDEA-SVM has less processing time for different size of data and it should also be noted that the IDEA based classifiers perform better than their corresponding DEA classifiers.

From the comparison results, it can be found that the proposed opinion mining framework using Improved DEA feature selection and SVM classification has better performance which is proved through with higher values of accuracy, precision, recall and f-measure while less processing time. This verifies that the Improved DEA algorithm is significantly better than the DEA optimization algorithm for opinion mining applications.

V. CONCLUSION

This paper was aimed at developing an automated opinion mining framework for medical tweets with highly efficient feature selection strategy. The DEA optimization algorithm was used to select optimal features but due to the convergence rate limitations, the improved DEA has been proposed. This IDEA algorithm improves the feature selection process and also improves the accuracy of classifiers which is evident from the simulation results. In future, the proposed model will be tested in other domains to evaluate its suitability for diverse applications. The possibility of developing hybrid feature selection algorithms by combining IDEA with other efficient optimization algorithms will also be investigated.

TABLE I
ACCURACY (%) COMPARISON

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
DEA-KNN	86.43	86.88	86.44	86.89	86.85	89.56	89.67	89.46	89.55	89.67
DEA-NB	88.2	88.3	88.26	88.66	88.71	90.86	90.9	90.93	90.79	90.9
DEA-SVM	92.99	93.33	93.54	93.34	93.53	93.21	93.33	93.41	93.3	93.33
IDEA-KNN	89.87	89.76	89.77	89.59	89.48	90.4	90.46	90.48	90.46	90.46
IDEA-NB	91.23	91.36	91.64	91.62	91.79	93.67	**96.92**	93.59	96.55	93.59
IDEA-SVM	**96.46**	**96.53**	**96.58**	**96.58**	**96.58**	**97.81**	96.85	**97.82**	**97.88**	**97.88**

TABLE II
PRECISION (%) COMPARISON

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
DEA-KNN	91.22	91.18	91.15	91.18	91.18	92.98	93.15	93.15	93.21	93.2
DEA-NB	93.8	93.79	93.79	93.83	93.81	95.33	95.38	95.53	95.54	95.66
DEA-SVM	98.85	98.82	98.81	98.85	98.85	98.32	98.32	98.36	98.28	98.34
IDEA-KNN	92.21	92.2	92.19	92.2	92.22	94.5	94.53	94.52	94.54	94.5
IDEA-NB	95.49	97.51	97.55	**98.55**	95.57	96.42	**98.4**	96.4	96.39	96.45
IDEA-SVM	**99.12**	**99.11**	**99.11**	98.16	**99.12**	**98.38**	98.39	**99.4**	**99.43**	**99.43**

TABLE III
RECALL (%) COMPARISON

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
DEA-KNN	89.21	89.89	89.75	89.79	89.76	90.38	90.46	90.28	90.45	90.44
DEA-NB	91.05	91.35	91.34	91.28	91.28	92.45	92.37	92.45	92.39	92.37
DEA-SVM	92.54	92.5	92.51	92.57	92.5	93.13	93.55	93.78	93.68	93.78
IDEA-KNN	90.9	90.98	91.13	90.92	90.89	91.56	91.08	91.76	91.72	91.99
IDEA-NB	**95.11**	93.36	93.28	93.37	93.36	94.87	**96.98**	95.23	94.83	94.87
IDEA-SVM	94.89	**96.31**	**96.56**	**96.55**	**96.54**	**96.48**	96.48	**97.48**	**97.39**	**97.48**

TABLE IV
F-MEASURE (%) COMPARISON

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
DEA-KNN	88.75	88.5	88.47	88.45	88.55	91.22	90.28	90.33	90.24	90.06
DEA-NB	91.77	91.36	91.49	91.4	91.43	92.35	92.07	92.98	92.34	92.76
DEA-SVM	94.26	94.32	94.32	94.51	94.3	95.18	95.1	95.24	95.41	95.37
IDEA-KNN	90.6	90.64	90.68	90.59	90.65	91.8	91.86	91.75	91.78	91.82
IDEA-NB	95.6	95.64	**97.65**	95.63	95.6	96.33	**97.54**	**97.9**	96.49	96.54
IDEA-SVM	**98.1**	**98.34**	97.52	**97.5**	**97.65**	**97.39**	97.48	97.87	**98.2**	**98.25**

TABLE V
PROCESSING TIME (SECONDS) COMPARISON

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
DEA-KNN	38.97	51.89	67.66	92.78	131.57	12.76	19.98	29.17	37.5	45.49
DEA-NB	35.12	49.08	66.18	90.8	128.29	11.89	19.1	28.54	35.96	44.32
DEA-SVM	33.76	47.88	64.32	90.16	128.74	11.03	18.67	27.65	35.03	44.18
IDEA-KNN	31.90	46.25	62.8	88.66	103.90	10.98	16.72	26.67	33.65	33.31
IDEA-NB	30.89	44.9	61.78	**87.54**	103.48	9.34	15.8	26.02	32.76	32.86
IDEA-SVM	**29.15**	**42.85**	**61.3**	87.68	**102.61**	**8.95**	**14.83**	**25.47**	**31.15**	**32.22**

Retrieval Number: J93170881019/19©BEIESP
DOI: 10.35940/ijitee.J9317.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

REFERENCES

1. E. Kouloumpis, T. Wilson and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," In *Fifth International AAAI conference on weblogs and social media*, pp. 538-54, July 2011.
2. A. Sabari, "A High End Sentimental Analysis in Social Media Using Hash Tags," *Journal of Applied Science and Engineering Methodologies*, vol. 1, no. 1, pp. 137-143, 2015.
3. H. Saif, Y. He and H. Alani, "Semantic sentiment analysis of twitter," In *International semantic web conference*, Springer, Berlin, Heidelberg, pp. 508-524, Nov 2012.
4. S. Kumar, F. Morstatter and H. Liu, *Twitter data analytics*. New York: Springer, pp. 1041-4347, 2014.
5. K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and society*. Peter Lang, vol. 89, 2014.
6. K. Thomas, C. Grier, D. Song and V. Paxson, "Suspended accounts in retrospect: an analysis of twitter spam," In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 243-258, Nov 2011.
7. B. Liu, "Sentiment Analysis and Subjectivity," *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627-666, 2010.
8. M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267-307, 2011.
9. G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez and J. Gordon, "Empirical study of machine learning based approach for opinion mining in tweets," In *Mexican international conference on Artificial intelligence*, Springer, Berlin, Heidelberg, pp. 1-14, Oct 2012.
10. A. Mudinas, D. Zhang and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, ACM, p. 5, Aug 2012.
11. D. Stojanovski, G. Strezoski, G. Madjarov, I. Dimitrovski and I. Chorbev, "Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages," *Multimedia Tools and Applications*, vol. 77, no. 24, pp. 32213-32242, 2018.
12. M. M. Mostafa, "Mining and mapping halal food consumers: A geo-located Twitter opinion polarity analysis," *Journal of Food Products Marketing*, vol. 24, no. 7, pp. 858-879, 2018.
13. N. Zainuddin, A. Selamat and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," *Applied Intelligence*, vol. 48, no. 5, pp. 1218-1232, 2018.
14. A. Al-Thubaity, Q. Alqahtani and A. Aljandal, "Sentiment lexicon for sentiment analysis of Saudi dialect tweets," *Procedia computer science*, vol. 142, pp. 301-307, 2018.
15. A. P. Rodrigues and N. N. Chiplunkar, "A new big data approach for topic classification and sentiment analysis of Twitter data," *Evolutionary Intelligence*, pp. 1-11, 2019.
16. D. H. Alahmadi and X. J. Zeng, "Twitter-based recommender system to address cold-start: A genetic algorithm based trust modelling and probabilistic sentiment analysis," In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1045-1052, Nov 2015.
17. N. Yuvaraj and A. Sabari, "Twitter sentiment classification using binary shuffled frog algorithm," *Intelligent Automation & Soft Computing*, vol. 23, no. 2, pp. 373-381, 2017.
18. A. Kumar and A. Jaiswal, "Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter," *Multimedia Tools and Applications*, pp. 1-25, 2019.
19. M. Tubishat, M. A. Abushariah, N. Idris and I. Aljarah, "Improved whale optimization algorithm for feature selection in Arabic sentiment analysis," *Applied Intelligence*, vol. 49, no. 5, pp. 1688-1707, 2019.
20. L. Goel and A. Garg, "Sentiment Analysis of Social Networking Websites using Gravitational Search Optimization Algorithm," *International Journal of Applied Evolutionary Computation (IJAEC)*, vol. 9, no. 1, pp. 76-85, 2018.
21. G. S. Budhi, R. Chiong, Z. Hu, I. Pranata and S. Dhakal, "Multi-PSO based Classifier Selection and Parameter Optimisation for Sentiment Polarity Prediction," In *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pp. 68-73, Nov 2018.
22. R. M. Packiam and V. S. J. Prakash, "A Novel Integrated Framework Based on Modular Optimization for Efficient Analytics on Twitter Big Data," In *Information and Communication Technology for Intelligent Systems*, Springer, Singapore, pp. 213-224, 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.
24. A. Kaveh and N. Farhoudi, "A new optimization method: Dolphin echolocation," *Advances in Engineering Software*, vol. 59, pp. 53-70, 2013.

AUTHORS BIOGRAPHIES

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.