An Optimal Clustering with Hybrid Metaheuristic Algorithm for Sentiment Analysis and Classification

Mohammed K. Hassan[1], Dina K. Hassan[2], Ahmed K. Metawee[3], Bassem Hassan[4]

[1]Mechatronics Department, Faculty of Engineering, Horus University in Egypt (HUE), Egypt

[2]Accounting Department, Faculty of Commerce, Kafr El Sheikh University, Egypt

[3]Accounting Department, Faculty of Commerce, Mansoura University, Egypt

[4]Dassault Systemes Deutschland GmbH, Meitnerstraße 8, 70563 Stuttgart, Germany

Emails: mkhassan@horus.edu.eg, dina.abdelsalam@com.kfs.edu.eg, metawee68@mans.edu.eg, bassem.hassan@3ds.com

Abstract

Sentimental Analysis (SA) becomes a familiar topic among business people, which is commonly applied for the classification of sentiments from online reviews. It is generally treated as a sentiment classification (SC) problem where the online reviews are categorized into positive or negative polarities using the words that exist in the online reviews. With this motivation, this paper presents a new K-means clustering with hybrid metaheuristic algorithm (KMC-HMA) for SA and classification. The proposed KMC-HMA technique initially performs data preprocessing to remove the unwanted words from the product reviews. In addition, K-means clustering technique is used for the clustering of the massive quantity of the applied product reviews. Moreover, the clustered data are fed into the classification model based on hybrid ant colony optimization (ACO) with dragonfly algorithm (DFA). The ACO algorithm is used for the classification of product reviews and the performance of the ACO algorithm can be optimally tuned by the use of DFA. The performance validation of the KMC-HMA technique is validated using two datasets such as Canon and ipod. The experimental values pointed out the superior performance of the KMC-HMA technique over the recent state of art techniques.

Keywords: Sentiment analysis, Data classification, Metaheuristics, Clustering algorithm, Hybrid algorithms, Rule based classifier.

1. Introduction

Internet has modified the mode in which how individuals portray their convictions and sentiments about items or administrations, occasions, themes, communications, and so on. It is predominantly done by means of informal organizations, audit sites, web gatherings, online journals, and Internet remarks [1-3]. These writings are highly sentimental and are hence valuable for organizations or people willing to further develop their item advertising methodologies and react with them. Strategies programmed for sentiment analysis (SA) are regularly picked for this reason. SA has been utilized for investigating sentiments of tweets on healthcare area [4], assess educators capacity and foresee understudy execution distinguish relevant extremity of medications, film, and eatery audits, inspect popular assessment on items and administrations and cultural issues, characterize e-students and their subjects of revenue in their informal organization's communications, perform assessment mining in tweets, break down sentiment direction of microblogs, and conjecture stock costs from sentiments of information signals in the monetary business sectors [5, 6]. These days, SA just as data mining is comprehensively explored research regions [7]. It

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

utilizes natural language processing (NLP) and text mining to separate individuals' feelings or feelings towards an occasion, item, or others [3, 4]. As a rule, SA forces distinguishing four components containing element, its angle, assessment holder, and his sentiment [5]. The extricated suppositions can be ordered to one or the other level headed or emotional content. The abstract content additionally can be arranged to positive or negative sentiments [8]. Fig. 1 shows the generic process involved in SA.

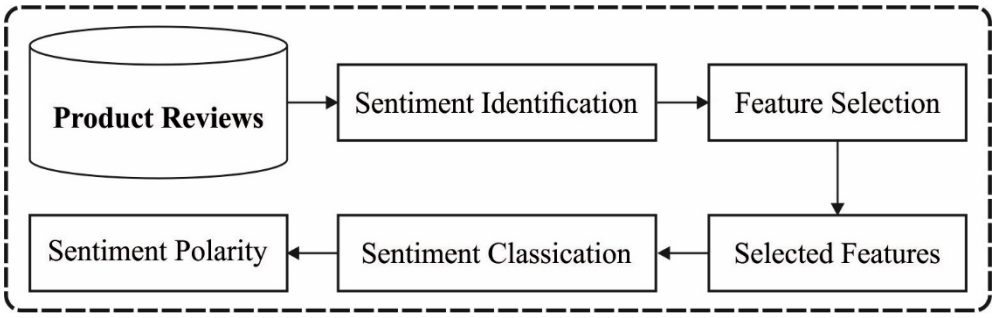

Fig. 1. Generic Process involved in SA

The issue of SA is a hot research subject. Despite the fact that SA is a significant region and right now has a wide scope of uses, it unmistakably is anything but a clear errand and has many difficulties identified with regular language handling (NLP). Late examinations on SA keep on confronting hypothetical and specialized issues that block their general exactness in extremity location [9, 10]. Hussein et al. considered the connection between those issues and the sentiment structure, just as their effect on the exactness of the outcomes. This work confirms that precision involves high worry among the most recent examinations on SA and demonstrates that it is influenced by certain difficulties, like tending to invalidation or area reliance. Web-based media are significant wellsprings of information for SA. Interpersonal organizations are ceaselessly growing, producing considerably more unpredictable and interrelated data. A few examinations center around building incredible models to settle the persistently expanding intricacy of huge information, just as to extend SA to a wide scope of utilizations, from monetary anticipating [11] and showcasing techniques [7] to medication analysis and different regions [12-14].

van Atteveldt et al. [15] gives a comprehensive evaluation of SA approaches utilizing a validation set of Dutch economic presents for comparing efficiency of manual annotation, crowd coding, several dictionaries and ML utilized combine classic and DL techniques. The 3 important assumptions are: (1) An optimum efficiency has been still reached to trained human/crowd coding; (2) All of the above utilized dictionaries come closer to suitable levels of validities; and (3) ML, particularly DL, significantly exhibits dictionary based techniques however decreases short of human efficiency.

Wang et al. [16] present a new technique to aspect level SA dependent upon the newly projected paradigm of Gradual Machine Learning (GML) that allows perfect machine labeling with no condition to manual labeling work. It starts with any simple examples from the task that is automatically labeled with machine thorough higher accuracy, afterward, slowly labels the further issues samples with iterative factor graph inferences. Mendon et al. [17] introduce a structure for analyzing user sentiment on Twitter over natural disasters utilizing the data pre-processed approaches and hybrid of ML, statistical modeling, and lexicon based method. It can be elect TF-IDF as well as K-means to sentiment classifier amongst affinitive and hierarchical clusters. The Latent Dirichlet Allocation, a pipeline of Doc2Vec as well as K-means utilized for capturing themes, next execute multi-level polarity indices classifier and their time series exploration.

In [18], an optimized based ML technique was presented for classifying the Twitter data. The procedure has been completed in 3 phases. The primary phase data was gathered and pre-processing, the second phase data was optimizing by removal required features, and the third phase the upgraded trained set was classified as to distinct classes with executing varying ML approaches. El-Affendi et al. [19] present a new DL based multilevel parallel attention neural (MPAN) method which utilized an easy positioning binary embedding scheme (PBES) for concurrently calculating contextualized embedded at the character,

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

word, and sentence level. The MPAN technique then calculates multi-level attention vector as well as concatenate them at the output levels for producing reasonable accurateness.

This paper presents a new K-means clustering with hybrid metaheuristic algorithm (KMC-HMA) for SA and classification. The proposed KMC-HMA technique initially performs data preprocessing to remove the unwanted words from the product reviews. In addition, K-means clustering technique is used for the clustering of the massive quantity of the applied product reviews. Moreover, the clustered data are fed into the classification model based on hybrid ant colony optimization (ACO) with dragonfly algorithm (DFA). The ACO algorithm is used for the classification of product reviews and the performance of the ACO algorithm can be optimally tuned by the use of DFA. The performance validation of the KMC-HMA technique is validated using two datasets such as Canon and ipod.

2. The Proposed Model

In this study, an effective KMC-HMA technique is derived for SA and classification. The proposed KMC-HMA technique initially performs data preprocessing to remove the unwanted words from the product reviews. Next, clustering and classification processes are carried out as elaborated in the following areas.

2.1 Preprocessing

The unwanted noise elements like URLs, stop words, hash tags, multiple spaces, etc. needs to be removed prior to extracting features. The URLs are removed using regular expression matching. The hash tag (#), punctuation marks like /, _, \is eliminated and more number of white spaces is replaced by white space. Next, all the words in the online review are transformed to lowercases. The stop word (an, a, is, the, an) also the word that doesn't begin with the alphabet are eliminated. Stop word dictionaries and acronym dictionaries are also used to enhance the accuracy of the data set.

2.2 K-Means Clustering

K-means clustering [20] is a popular information clustering technique which groups n data point to K cluster with the minimization of distance of data point from K CHs in an iterative manner. The distance is computed by cosine measure/Euclidean distance. In K -means method, the process of selecting the values of K for a dataset with unidentified numbers of classes is a difficult task. In those situations, elbow technique, information criterion method, silhouette method, etc. are employed to choose K value. The normalized feature vector is applied to KMC-HMA algorithm, which incorporates K-means and ACO techniques for clustering the information. Since K-means are the simplest and efficient clustering technique, it is widely used in various fields. But, it suffers from the drawbacks of trapping to initial clusters. The created clusters can be employed for additional investigation. So, in such cases, the created cluster from K-means are employed in ACO technique to optimize the cluster-heads. The number of online product reviews is increasing day by day and the dataset size becomes very large. Since the larger dataset may increase the burden of ACO and stuck to less classification accuracy. Hence, the presented approach modifies the initial procedure of ACO method that leads to fast convergences and improved classification accuracy. For KMC-HMA method, the solution attained from K-means is applied for initialization of KMC-HMA method.

Let n represent the number of online product reviews which are clustered to N classes. Every online product review is denoted as a featured vector which holds S numbers of features as well as all the features were measured in [0, T]. The likelihood sharing of all features are computed by

$$p_i = \frac{O_i}{n} \tag{1}$$

where I indicate ith feature value $(0 \leq i \leq T)$ and O_i represents the overall amount of online product reviews has ith feature value.

$$\mu = \sum_{i=1}^{T} i\, p_i \tag{21}$$

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

Any online product review is categorized into classes D_j where it contains minimal Euclidean distance. So, the likelihood of existence w_j of classes are equated as,

$$w_j = \sum_{i \in D_j} p_i \tag{3}$$

The mean of class D_j is computed as

$$\mu_j = \sum_{i \in D_j} \frac{ip_i}{w_j} \tag{4}$$

The interclass variances are normally determined by:

$$\sigma^2 = \sum_{j=1}^{N} w_j(\mu_j - \mu)^2 \tag{5}$$

For the clustering process of various online product reviews, the interclass variances displayed in Eq. (5) must be increased.

2.3 Data Classification

ACO algorithm is developed by [21], based on the foraging nature of real nature of real ants. In ACO based data classification process, the ant finds the shortest path between two points. The ants select the feasible path using a probability function, which is obtained by the amount of pheromones present in the heuristic and path function. Once the ant passed all the available routes, the route with maximum quantity of pheromones and the heuristic values through high likelihood would be preferred. Once ant enters a route, the pheromones begin to increase. When enough number of ants use the same paths, it becomes a candidate rule. The candidate rules will become discover rules only if the qualities are better. The ACO algorithm is used for the classification of product reviews and the performance of the ACO algorithm can be optimally tuned by the use of DFA. DFA is a recently developed bio-inspired optimization method deployed by [22]. It is mainly evolved from static and dynamic performances of dragonflies swarm behavior.

Because of this behavior, 5 important attributes that affect the individuals' upgrading location are Alignment, Separation, Attraction towards food source, Distraction of enemy, and Cohesion. Such attributes are explained numerically given below.

Separation: This parameter is measured by the given formula:

$$S_i = -\sum_{k=1}^{M} Y - Y_k, \tag{6}$$

where Ymeans the individual's recent place, Y_k denotes the location of *kth* neighbouring individuals and Mdefines the overall count of neighbouring separations.

Alignment: It shows that mean of velocities which are determined by given expression:

$$A_i = \frac{\sum_{k=1}^{M} V_k}{M}, \tag{7}$$

where $V_k$implies the velocity of *kth* neighbouring individuals.

Cohesion: the attributes are evaluated by:

$$C_i = \frac{\sum_{k=1}^{M} Y_k}{M} - Y \tag{8}$$

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

Attraction towards a food source: It signifies a distance among location of present individual and location of food source (Y^+) and it can be determined by the provided notion:

$$F_i = Y^+ - Y \tag{9}$$

Distraction outwards an enemy: It means a distance among position of recent individual and location of an enemy (Y^-) which is measured by:

$$E_i = Y^- - Y \tag{10}$$

In dragonfly, nature is a unification of 5 variables. Then, 2 vectors are employed for upgrading dragonflies' place in a search space, such as step vectors ($\Delta\, Y$) and location vectors (Y). Step vectors are illustrated by:

$$\Delta\, Y_{t+1} = (aA_i + sS_i + cC_i + eE_i + fF_i) + w\, \Delta\, Y_t\,, \tag{11}$$

where a means the alignment weight, $A_i$refers to the position of *ith* individuals, $s$denotes the separation weights, $S_i$implies differentiation of *ith* individuals, $c$defines cohesion weights, $C_i$represents cohesion of *ith* individuals, e depicts enemies weights, $E_i$defines location of enemies in *ith* individuals, $f$signifies the food weight, $F_i$showcases food sources of *ith* individuals, w shows the inertia weightiness, also t exhibits the described as iteration number.

3. Performance Validation

In this section, the performance validation of the KMC-HMA technique on the two datasets namely Canon and ipod. Table 1 and Fig. 2 reports the detailed comparative study of the KMC-HMA technique with compared models on Canon dataset. On examining the results interms of sensitivity, it is obvious that the CSKA and SVMA techniques have resulted in poor outcomes with the lower sensitivity of 0.842 and 0.854. In line with, the PSOA and NNA techniques have reached a slightly increased sensitivity of 0.867 and 0.867. Then, the ACOA and ACO-K techniques have accomplished a moderately reasonable outcome with the sensitivity of 0.982 and 0.986. But the proposed KMC-HMA technique has offered a higher sensitivity of 0.989.

In addition, on investigating the outcomes with respect to specificity, it can be clear that the CSKA and SVMA techniques have resulted in poor results with the lesser specificity of 0.515 and 0.544. Also, the NNA and PSOA methods have attained a somewhat improved specificity of 0.587 and 0.641. Afterward, the ACOA and ACO-K methods have accomplished a moderately reasonable outcome with the specificity of 0.901 and 0.935. However, the presented KMC-HMA methodology has obtainable a superior specificity of 0.945.

Table 1 Comparative analysis of proposed method with existing methods for Canon dataset

Methods	Sensitivity	Specificity	Accuracy	F-score
Ours	0.989	0.945	0.980	0.994
ACO-K	0.986	0.935	0.975	0.984
ACOA	0.982	0.901	0.964	0.977
PSOA	0.867	0.641	0.825	0.890
CSKA	0.842	0.515	0.775	0.856
SVMA	0.854	0.544	0.803	0.879
NNA	0.867	0.587	0.819	0.888

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

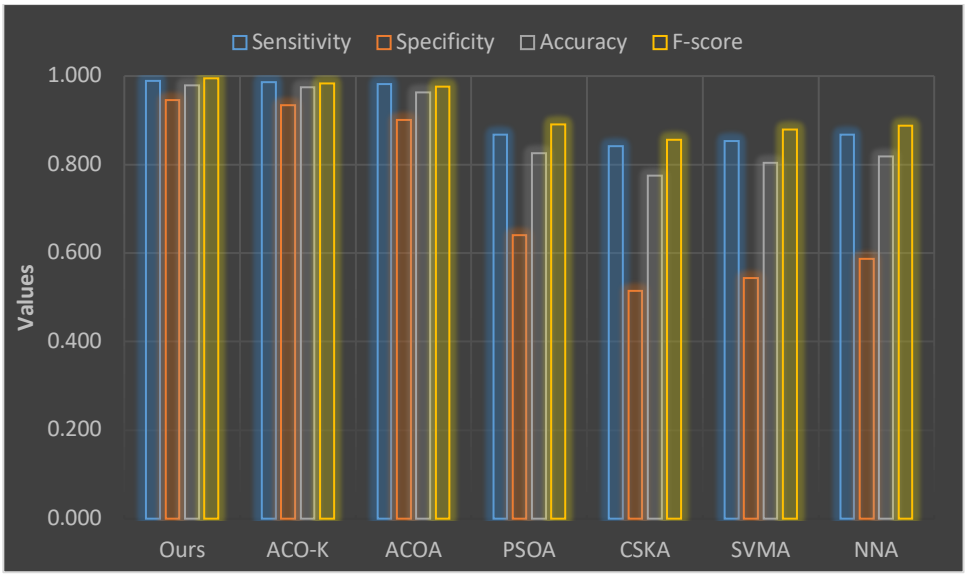

Fig. 2. Comparative study of KMC-HMA technique on Canon dataset

Moreover, on examining the results interms of accuracy, it is apparent that the CSKA and SVMA methods have resulted in the least result with the lower accuracy of 0.775 and 0.803. Similarly, the NNA and PSOA techniques have achieved somewhat higher accuracy of 0.819 and 0.825. Afterward, the ACOA and ACO-K techniques have accomplished a moderately reasonable outcome with the accuracy of 0.964 and 0.975. Finally, the projected KMC-HMA technique has existing an improved accuracy of 0.980.

Finally, on scrutinizing the results with respect to F-score, it can be stated that the CSKA and SVMA methods have resulted in worse outcomes with the minimum F-score of 0.856 and 0.879. Similarly, the NNA and PSOA approaches have obtained a slightly superior F-score of 0.888 and 0.890. In addition, the ACOA and ACO-K techniques have accomplished a moderately reasonable outcome with the F-score of 0.977 and 0.984. Lastly, the proposed KMC-HMA technique has accessible a maximum F-score of 0.994.

Table 2 and Fig. 3 offer the detailed comparative study of the KMC-HMA manner with compared techniques on iPod dataset. On investigative the results with respect to sensitivity, it can be evident that the PSOA and SVMA manners have resulted in to least outcome with the lesser sensitivity of 0.807 and 0.820. Also, the CSKA and NNA techniques have reached a somewhat higher sensitivity of 0.822 and 0.833. Then, the ACOA and ACO-K methods have accomplished a moderately reasonable outcome with the sensitivity of 0.912 and 0.949. Lastly, the projected KMC-HMA method has obtainable a maximal sensitivity of 0.963.

Table 2 Comparative analysis of proposed method with existing methods for iPod dataset

Methods	Sensitivity	Specificity	Accuracy	F-score
Ours	0.963	0.998	0.999	0.968
ACO-K	0.949	0.995	0.985	0.965
ACOA	0.912	0.994	0.975	0.943
PSOA	0.807	0.939	0.914	0.778
CSKA	0.822	0.926	0.908	0.750
SVMA	0.820	0.854	0.848	0.649
NNA	0.833	0.857	0.853	0.659

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

Next, on scrutinizing the outcomes interms of specificity, it is noticeable that the SVMA and NNA techniques have resulted in worse results with the lower specificity of 0.854 and 0.857. In line with, the CSKA and PSOA techniques have gained a slightly superior specificity of 0.926 and 0.939. Then, the ACOA and ACO-K methods have accomplished a moderately reasonable outcome with the specificity of 0.994 and 0.995. However, the proposed KMC-HMA methodology has offered a higher specificity of 0.998.

Then, on examining the outcomes with respect to accuracy, it can be clear that the SVMA and NNA techniques have resulted in minimal outcomes with the least accuracy of 0.848 and 0.853. Along with that, the CSKA and PSOA techniques have reached a somewhat maximum accuracy of 0.908 and 0.914. Afterward, the ACOA and ACO-K techniques have accomplished a moderately reasonable outcome with the accuracy of 0.975 and 0.985. But the presented KMC-HMA technique has existed a maximum accuracy of 0.999.

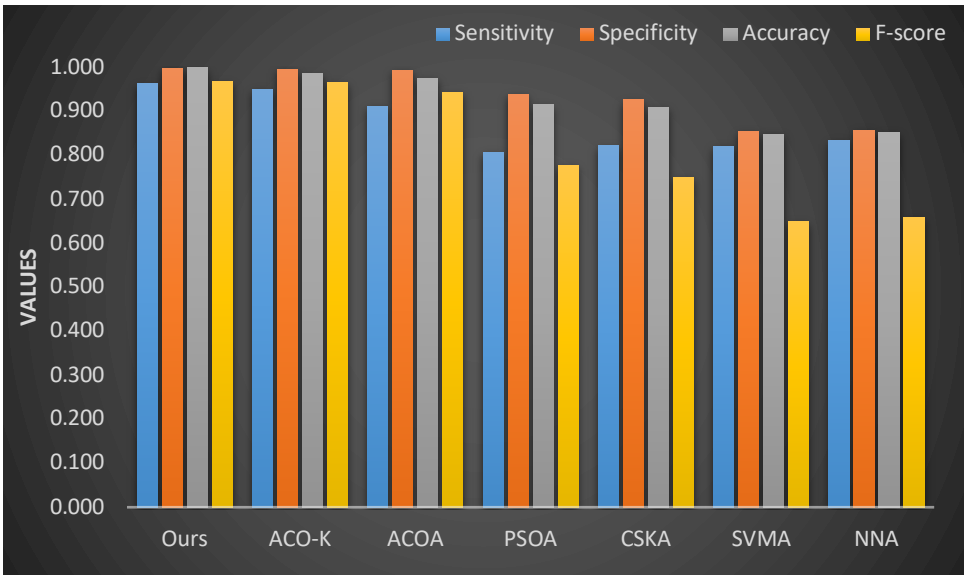

Fig. 3. Comparative study of KMC-HMA technique on ipod dataset

Finally, on inspecting the results interms of F-score, it can be evident that the SVMA and NNA approaches have resulted in poor outcomes with the minimum F-score of 0.649 and 0.659. Besides, the CSKA and PSOA manners have gained to a slightly increased F-score of 0.750 and 0.778. At the same time, the ACOA and ACO-K techniques have accomplished a moderately reasonable outcome with the F-score of 0.943 and 0.965. Eventually, the proposed KMC-HMA methodology has accessible a higher F-score of 0.968.

4. Conclusion

In this paper, an effective KMC-HMA technique is derived for SA and classification. The proposed KMC-HMA technique initially performs data preprocessing to remove the unwanted words from the product reviews. In addition, K-means clustering technique is used for the clustering of the massive quantity of the applied product reviews. Moreover, the clustered data are fed into the classification model based on hybrid ACO with DFA. The ACO algorithm is used for the classification of product reviews and the performance of the ACO algorithm can be optimally tuned by the use of DFA. The performance validation of the KMC-HMA technique is validated using two datasets such as Canon and ipod. The experimental values pointed out the superior performance of the KMC-HMA technique over the recent state of art techniques. The efficacy of the KMC-HMA technique can be boosted by the use of outlier detection approaches.

References

[1] Hasan, A., Moin, S., Karim, A. and Shamshirband, S., 2018. Machine learning-based sentiment analysis for twitter accounts. Mathematical and Computational Applications, 23(1), p.11.

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021

[2] Baid, P., Gupta, A. and Chaplot, N., 2017. Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7), pp.45-49.
[3] Mitra, A., 2020. Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies (UCCT), 2(03), pp.145-152.
[4] Jagdale, R.S., Shirsat, V.S. and Deshmukh, S.N., 2019. Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing (pp. 639-647). Springer, Singapore.
[5] Aziz, A.A. and Starkey, A., 2019. Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. IEEE Access, 8, pp.17722-17733.
[6] Rathi, M., Malik, A., Varshney, D., Sharma, R. and Mendiratta, S., 2018, August. Sentiment analysis of tweets using machine learning approach. In 2018 Eleventh international conference on contemporary computing (IC3) (pp. 1-3). IEEE.
[7] Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B., 2019. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy, 21(6), p.589.
[8] Abd El-Jawad, M.H., Hodhod, R. and Omar, Y.M., 2018, December. Sentiment analysis of social media networks using machine learning. In 2018 14th international computer engineering conference (ICENCO) (pp. 174-176). IEEE.
[9] Ahmad, M., Aftab, S., Muhammad, S.S. and Ahmad, S., 2017. Machine learning techniques for sentiment analysis: A review. Int. J. Multidiscip. Sci. Eng, 8(3), p.27.
[10] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P. and Tech, B., 2017. Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), pp.29-34.
[11] Arulmurugan, R., Sabarmathi, K.R. and Anandakumar, H.J.C.C., 2019. Classification of sentence level sentiment analysis using cloud machine learning techniques. Cluster Computing, 22(1), pp.1199-1209.
[12] Renault, T., 2020. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. Digital Finance, 2(1), pp.1-13.
[13] Mukhtar, N., Khan, M.A. and Chiragh, N., 2018. Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. Telematics and Informatics, 35(8), pp.2173-2183.
[14] Yang, P. and Chen, Y., 2017, December. A survey on sentiment analysis by using machine learning methods. In 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (pp. 117-121). IEEE.
[15] van Atteveldt, W., van der Velden, M.A. and Boukes, M., 2021. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. Communication Methods and Measures, 15(2), pp.121-140.
[16] Wang, Y., Chen, Q., Shen, J., Hou, B., Ahmed, M. and Li, Z., 2021. Aspect-level sentiment analysis based on gradual machine learning. Knowledge-Based Systems, 212, p.106509.
[17] Mendon, S., Dutta, P., Behl, A. and Lessmann, S., 2021. A Hybrid approach of machine learning and lexicons to sentiment analysis: enhanced insights from twitter data of natural disasters. Information Systems Frontiers, pp.1-24.
[18] Naresh, A. and Venkata Krishna, P., 2021. An efficient approach for sentiment analysis using machine learning algorithm. Evolutionary Intelligence, 14, pp.725-731.
[19] El-Affendi, M.A., Alrajhi, K. and Hussain, A., 2021. A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain arabic sentiment analysis. IEEE Access, 9, pp.7508-7518.
[20] S. MuthuKumaran, P.Suresh, A unified framework of sentimental analysis for online product reviews using enhanced ant colony optimization algorithm, International Journal of Pure and Applied Mathematics, Vol. 119, No. 4, 2018, pp. 489-496.
[21] Dorigo, M. and Stützle, T., 2019. Ant colony optimization: overview and recent advances. Handbook of metaheuristics, pp.311-351.
[22] Rahman, C.M. and Rashid, T.A., 2019. Dragonfly algorithm and its applications in applied science survey. Computational Intelligence and Neuroscience, 2019.

DOI: 10.5281/zenodo.5501329
Received: January 22, 2021 Accepted: August 13, 2021