Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
https://doi.org/10.53070/bbd.991715

Journal of Computer Science
ISSN: 2548-1304
Volume: IDAP-2021, Issue: Special pp: 35-40
@Anatolian Science

A new plant intelligence-based method for sentiment analysis: Chaotic sunflower optimization

Suna YILDIRIM*[1], Güngör YILDIRIM [2], Bilal ALATAŞ[3]

[1] IT department of the Secretary General of Special Provincial Administration, Elazig, Turkey

[2]Computer Engineering Dept., Fırat University, Elazig, Turkey

[3]Software Engineering Dept., Fırat University, Elazig, Turkey

(sunayildirim23@gmail.com.tr, gungor.yildirim@firat.edu.tr, balatas@firat.edu.tr)

Received: Sep.6, 2021 Accepted: Sep.16, 2021 Published: Oct.20, 2021

Abstract— Various social networking applications provide people with many opportunities such as expressing, commenting, disseminating and transmitting their opinions within certain limits. The emotions and ideas that people express in their messages make sense of thousands of articles and opinions published instantly. Trying to make sense of emotional data, generating meaningful information from these data, analyzing these data, and making predictions and inferences on these data is a new important study field. In this study, sentiment analysis is considered an optimization problem in order to achieve high performance. For this purpose, sunflower optimization, which is one of the new and successful plant intelligence-based algorithms, has been modelled as a sentiment analyzer for the first time. A chaotic sunflower optimization algorithm was used by combining sunflower optimization and chaos theory in order to make effective sentiment analysis. In order for the proposed method to effectively solve the sentiment analysis problem, a suitable representation form and fitness function have been proposed. The proposed method treats the data as a search space and searches for a solution for analysis by detecting emotion in this search space. An up-to-date data set including customer feedback and satisfaction information was used in the study. Results based on accuracy, precision, and recall metrics show that plant intelligence-based metaheuristic algorithms can provide high performance.

Keywords: Sentiment analysis, plant intelligence, chaotic sunflower optimization, chaos

1.Introduction

Sentiment analysis (SA), called opinion mining (OM) in some sources, can be defined as a process in which people can obtain information about companies, products, policies, events, etc. using natural language processing techniques. Although SA and OM are two concepts that can be used interchangeably, some researchers have recently stated that there are differences between these concepts. In a broader definition, while OM extracts people's views on an entity, SA identifies the emotion expressed in a text and then analyzes it (Medhat et al, 2014). After the 2000s, sentiment analysis gained momentum in every field, especially with the rapid increase in internet use. When people want to know about a product or company, it is very difficult and costly to get objective information from the people around them. Manually making these inferences from information on social media can be very laborious and time-consuming. In addition to social media, user communities can be created to exchange information about products and companies. Comments from other users/customers can be provided through these communities. Since these comments will not be in a standard, it will be difficult to analyze manually. In such a case, if the comments and opinions are modelled as classification problems, the SA can provide effective solutions and perform automated analysis. In order to overcome such problems, machine learning methods have been used frequently. On the other hand, the development of different analysis methods in this field is still a hot area of study.

In recent years, with the spread of metaheuristic optimization methods, different solutions have been proposed to problems in different disciplines. In addition, these methods offer flexible opportunities as they enable different mathematical approaches, give stable results, and are suitable for hybrid working techniques (Akyol et al, 2020). These advantages have enabled them to be used in solving social media analysis problems. This article presents the results of an experimental study based on an optimization-based approach to SA problems. The study focuses on the performances of the plant intelligence-based meta-heuristic in SA analysis. For this, the Chaotic Sunflower Optimization (CSO) algorithm, a version of the Sunflower Optimization algorithm that is one of the successful plant intelligence-based approaches in recent years, was used. In this method, the relevant data set is evaluated as a search space and the solutions produced by the candidates are evaluated according to their positions in this search space. In the study, a data set created with customer feedback was used. The obtained results have proven that plant intelligence-based metaheuristic approaches can provide high performance.

The next sections of the paper are scheduled as follows; the background and literature summary about sentiment analysis will be shared in Section 2. The basic details of the CSO algorithm and the methodology used will be given in Section 3 and Section 4, respectively. The results of the experiments will be shared in Section 5, and Section 6 will include the conclusion.

2. Background and Related Works

Sentiment analysis techniques suggested in the literature are generally grouped under two main headings. These are machine learning-based approaches and lexicon-based methods (Medhat et al, 2014). Lexicon-based methods usually use a predefined set of lexicons. These methods make inferences using statistical or semantic analysis. In machine learning-based approaches, supervised and unsupervised algorithms are used. Here, the structure of the dataset also changes the type of machine learning algorithm used. In supervised techniques, well-known classification algorithms in the literature have succeeded in providing effective solutions. The optimization-based approach used in this report can also be considered as an approach in the supervised class. Unsupervised approaches are mostly used in problems where the data set is complex or there is no classification information.

Abdu et al. (2021) presented a review of the latest updates in multimodal sentiment analysis. The most popular datasets in their field and the most popular feature extraction methods are categorized and discussed. Thirty-five of the articles considered were categorized and summarized according to the architecture used in each model. The efficiency and effectiveness of the thirty-five models used were compared on two datasets commonly used for multimodal sentiment analysis (CMU-MOSI and CMU-MOSEI). Al-Twairesh and Al-Negheimish (2019) proposed a collection of superficial and deep features for the sentiment classification of Arabic tweets. Also, different models have been explored to incorporate emotion into word embedding. A study was conducted to evaluate the combination of general word embedding, emotion-specific word embedding, and manually created features. The models were evaluated on three Arabic Tweets datasets. The results showed that generic word embedding generated from a large dataset of Arabic tweets using the popular word2vec method outperformed sentiment-specific embedding. Aziz and Starkey (2020) explain why it is important to be able to measure the performance of SML (Supervised Machine Learning) models against real-world datasets in real-time. They stated that the purpose of the article is to provide a technique that can be used to explore the predictive capability and anomalies of SML models. Basiri et al. (2021) conducted their study to find the general feelings (opinions) of people in eight countries about Covid-19. They collected tweets from these eight countries and Google Trends users using coronavirus-related keyword searches. For sentiment analysis, they proposed a new hybrid fusion model using five deep classifiers and combined them to improve the final output using a meta-learning method. Birjali et al. (2021) presented an overview of sentiment analysis and approaches related to this field. In the article, the most used classification techniques to perform sentiment analysis were examined and categorized. Various emotion classification techniques are categorized and examined with their advantages/disadvantages.

Carosia et al. (2021) proposed SA based investment strategies using Artificial Neural Networks for the Brazilian financial market for the period from June 2018 to June 2019. Gavilanes et al. (2021) proposed a fully automatic unsupervised methodology to evaluate the quality of the dictionary created from online emoji sources. Jindal and Aron (2021) analyzed the studies in the field of sentiment analysis. In their studies, various techniques for the emotional analysis of social media data were examined. Additionally, in the literature, sensitivity changes (Liang et al, 2020; Mukherjee et al, 2021), commercial efficiency analyses (Smetanin, 2020) and semantic fuzziness analyzes (Fang et al, 2018) was handled as a sentiment analysis problem.

3. Sunflower Optimization Algorithm and CSO

In the sunflower algorithm, the basic factor is the distance from the sun. The inverse square radiation law applies here. The principle in the algorithm is to represent the orientation of sunflowers to the sun. For this, the sun (X^*) in the population is the reference for other sunflowers (X_i). The orientation, which takes into account the distance to the reference, is expressed by Eq. 1.

$$\vec{s}_i = \frac{X^* - X_i}{||X^* - X_i||}, \quad i = 1, 2, \dots, n_p \tag{1}$$

The orientation steps of the population individuals (plants) are among the critical parameters of the sunflower algorithm. At the orientation speed of the population individuals, the ith plant pollinates with the (i-1)th plant and creates a new plant with random positions. This random position is proportional to the distance between individuals. The general expression of the sun orientation steps is given in Eq. 2. In this equation, P_i represents the pollination probability among the individuals concerned, and λ represents the inertia coefficient.

$$d_i = \lambda P_i(||X_i + X_{i-1}||)||X_i + X_{i-1}|| \tag{2}$$

The other basic criteria in determining the step values is the maximum step length (d_{max}). This parameter, shown in Eq. 3, is directly proportional to the Euclidean distance between the default upper limit (X_{max}) and lower limit (X_{min}) in the problem definition, and inversely proportional to the population size (N_{pop}). Based on these basic parameters, the new individual is calculated by Eq. 4.

$$d_{max} = \frac{||X_{max} - X_{min}||}{2N_{pop}} \tag{3}$$

$$\vec{X}_{i+1} = \vec{X}_i + d_i \vec{s}_i \tag{4}$$

Chaotic maps are the randomness of a mathematically simple deterministic dynamic system, and a chaotic system can be considered as a source of randomness (Gomes et al, 2019). In their previous study, the success of the chaotic version of the sunflower algorithm has been proven by the authors (Yıldırım et al., 2021). The Tent chaotic map (Eq. 5), which provided high performance in previous studies, was used as a source of randomness in the AO algorithm and this version was named CSO.

$$X_{n+1} = \left\{ \begin{array}{ll} \mu X_n, & X_n < \frac{1}{2} \\ \mu(1 - X_n), & \frac{1}{2} \leq X_n \end{array} \right\} \tag{5}$$

4. Methodology

In optimization-based approaches, the relevant data set needs pre-processing steps. In these stages, firstly, word roots are obtained. This includes case converting, N-char filtering, punctuation deletion, and stemmer processing. Then, the weight of each word in the data set (*WW*) is calculated by Eq.6. For this, the number of repetitions of each word in the data set and the maximum number of repetitions are taken into account. Words with a weight value above a certain threshold are selected as the attributes of the search space. In the last step, it is found which attribute word is used in which record in the data set. It is evaluated as 1 (true) if a word is used in a record, and 0 (false) otherwise. Thus, a binary search space is created.

$$WW_i = \frac{Repetition_i}{Max\ Repetition} \tag{6}$$

The variables of CSO candidates are in the range of [0, 1]. The similarity and class match of a candidate (x_i) with each record in the search space (d_i) forms the basis of the evaluation criterion. For similarity, Jaccard Similarity whose expression is given in Eq.7 is used. As seen in Eq.8, the fitness is based on weighted accuracy, precision, and recall metrics. The true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values used in the calculation of these metrics are calculated with the updating rules in Table 1.

$$Jaccard\ Value_{\vec{X}} = \frac{\sum_{i=0}^{K} x_i \times d_i}{\sum_{i=1}^{K}(x_i)^2 + \sum_{i=1}^{K}(d_i)^2 - \sum_{i=0}^{K} x_i \times d_i} \tag{7}$$

$$F = w_1 \frac{TP+TN}{TP+TN+FP+FN} + w_2 \frac{TP}{TP+FP} + w_3 \frac{TP}{TP+FN} \tag{8}$$

Table 1. Updating of TP, FP, FN, and TN for each record in the dataset

Condition	*Updating*
If $Jaccard\ Value_{\vec{N}} \geq\ Threshold$ ***and*** *the Class searched == the Record Class*	Increase TP by 1
If $Jaccard\ Value_{\vec{N}} \geq\ Threshold$ ***and*** *the Class searched != the Record Class*	Increase FP by 1
If $Jaccard\ Value_{\vec{N}}\ < Threshold$ ***and*** *the Class searched == the Record Class*	Increase FN by 1
If $Jaccard\ Value_{\vec{N}}\ < Threshold$ ***and*** *the Class searched! = the Record Class*	Increase TN by 1

5. Experiments

The experiments were carried out on the TripAdvisor dataset (Alam et al, 2016). This dataset contains a total of 20491 customer opinions and has 5 classes. These classes are one (1421), two (1793), three (2184), four (6039) and five (9054), respectively. The data set was first converted to binary data set by performing the pre-processing steps described above. 70% of this data set was used for training and 30% of it for testing. The search space contains a total of 177 features (words) and 1 class. Thus, the search space dimensions were 178 x 14344 for training, and 178 x 6147 for testing. In the experiments, the Jaccard similarity threshold was 20% and w_1, w_2, and w_3 were $\frac{1}{3}$.

Table 2. The Results For Class-1

METHOD	Acc	Pre	Rec
KOA	0.669	0.133	0.683
SVM	0.532	0.564	0.457
NB	0.497	0.398	0.444
DT	0.423	0.330	0.295
IBk	0.400	0.235	0.237

Table 3. The Results For Class-2

METHOD	Acc	Pre	Rec
KOA	0.578	0.106	0.513
SVM	0.532	0.367	0.174
NB	0.497	0.275	0.259
DT	0.423	0.209	0.187
IBk	0.400	0.162	0.138

Table 4. The Results For Class-3

METHOD	Acc	Pre	Rec
SVM	0.532	-	-
KOA	0.529	0.114	0.504
NB	0.497	0.270	0.204
DT	0.423	0.166	0.166
IBk	0.400	0.161	0.147

Table 5. The Results For Class-4

METHOD	Acc	Pre	Rec
KOA	0.557	0.332	0.497
SVM	0.532	0.432	0.4
NB	0.497	0.448	0.35
DT	0.423	0.357	0.352
IBk	0.400	0.364	0.392

Table 6. The Results For Class-5

METHOD	Acc	Pre	Rec
SVM	0.532	0.586	0.834
KOA	0.522	0.464	0.534
NB	0.497	0.612	0.701
DT	0.423	0.575	0.601
IBk	0.400	0.545	0.543

Table 7. Statistical results of CSO experiments

	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
Best	0.669	0.133	0.765	0.578	0.106	0.627	0.529	0.119	0.625	0.557	0.332	0.645	0.522	0.464	0.610
Worst	0.466	0.093	0.647	0.478	0.094	0.503	0.448	0.112	0.504	0.504	0.326	0.497	0.497	0.444	0.453
Mean	0.585	0.113	0.705	0.517	0.099	0.559	0.503	0.114	0.543	0.524	0.329	0.589	0.506	0.451	0.538
Median	0.628	0.117	0.694	0.515	0.098	0.574	0.511	0.114	0.537	0.516	0.329	0.611	0.504	0.449	0.542
Std	0.074	0.013	0.038	0.032	0.004	0.040	0.022	0.002	0.032	0.017	0.002	0.048	0.007	0.005	0.040

For comparison, well-known classification algorithms in the literature were used. These algorithms are Naive Bayes-NB (Rish, 2001), Support Vector Machine-SVM (Yue et al, 2003), Decision Tree-DT (Kotsiantis, 2013) and IBk (Moayedi et al, 2019). A total of 20 independent experiments were carried out. The comparative results for the Accuracy (Acc), Precision (Pre), and Recall (Rec) metrics obtained in the experiments are presented in Table 2-6, and the statistical results of all the experiments are presented in Table 7.

In the tables, all metric values of CSO are given separately. Since standard classification algorithms give a single accuracy for all classes, this value is shown in all tables. On the other hand, the precision and recall values of the standard methods are given separately for each class in the tables. Considering the weight values of the classes, CSO obtained the best value with 0.538 in the accuracy metric. The closest accuracy to this is SVM with 0.532. CSO achieved the highest accuracy values in Classes 1, 2, and 4. In addition, the best recall values in four classes were obtained by CSO. In this metric, NB was more successful than other standard classification methods. On the other hand, CSO was not successful in precision in all classes. In this metric, NB and SVM were more successful than other methods. Except for SVM, all methods managed to get results for each class. In the statistical results, mean and median values were close to each other, and it was also seen that the best recall value was quite high in some classes

5. Conclusions

Sentiment analysis is among the important study topics of social media technologies. This paper focused on the plant-intelligence-based metaheuristic sentiment analysis, which is an alternative and competitive analysis method for sentiment analysis. For this, the chaotic sunflower algorithm, one of the most prominent plant-intelligence-based metaheuristic algorithms, was used. As a result of the experiments, it was seen that this approach could produce successful results. The experiments were conducted on the Trip Advisor data set, which includes customer satisfaction information about hotels. In the experiments performed with accuracy, precision, and recall metrics, the best values in accuracy and recall were obtained by CSO. The optimization method used was single-objective and the experience showed that multi-objective optimization approaches would be possible in this field. After this study, the authors will gather their attention on developing multi-objective solution methods in social media analysis.

References

Abdu S, Yousef A, Salem A (2021) Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey. *Information Fusion* 76:204-226.

Akyol S, Alatas B (2020) Sentiment classification within online social media using whale optimization algorithm and social impact theory based optimization. *Physica A: Statistical Mechanics and its Applications* 540:123094.

Alam MH, Ryu WJ, Lee S (2016) Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences* 339:206–223.

Al-Twairesh N, Al-Nagheimish H (2019) Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets. *IEEE Access* 7:84122-84131.

Aziz A, Starkey A (2020) Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches. *IEEE Access* 8:17722-17733.

Basiri ME, Nemati S, Abdar M, Asadi S, Acharrya UR (2021) A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems* 228:107242.

Birjali M, Kasri M, Beni-Hssane A (2021) A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226:107134.

Carosia AE, Coelho GP, Silva AE (2021) Investment strategies applied to the Brazilian stock market: A methodology based on Sentiment Analysis with deep learning. *Expert Systems with Applications* 184:115470.

Fang Y, Tan H, Zhang J (2018) Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness. *IEEE Access* 6:20625-20631.

Gavilanes MF, Montenegro EC, Mendez SG, Castano FG, Martinez JJ (2021) Evaluation of online emoji description resources for sentiment analysis purposes. *Expert Systems with Applications* 184:115279.

Gomes GF, Cunha Jr SS, Ancelotti Jr AC (2019) A sunflower optimization (SFO) algorithm applied to damage identification on laminated composite plates. *Engineering with Computers* 35: 619-626.

Jindal K, Aron R (2021) A systematic study of sentiment analysis for social media data. *Materials Today: Proceedings* Article in Press.

Kotsiantis SB (2013) Decision trees: a recent overview. *Artif Intell Rev* 39:261–283.

Liang H, Ganeshbabu U, Thorne T (2020) A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution. *IEEE Access* 8:54164-54174.

Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications:A survey. *Ain Shams Engineering Journal* 5(4):1093-1113.

Moayedi H, Bui D, Kalantar B, Foong LK (2019) Machine-Learning-Based Classification Approaches toward Recognizing Slope Stability Failure. *Applied Sciences* 9(21):4638.

Mukherjee P, Badr Y, Doppalapui S, Srinivasan SM, Sangwan RS, Sharma R (2021) Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *Procedia Computer Science* 185:370-379.

Rish I (2001) An empirical study of the naive Bayes classifier, IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM New York, pp.41-46.

Smetanin S (2020) The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives. *IEEE Access* 8:110693-110719.

Yıldırım S, Yıldırım G, Alatas B (2021) Salınımlı Kaotik Ayçiçeği Optimizasyon Algoritması. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi* Article in Press

Yue S, Li P, Hao P (2003) SVM classification: Its contents and challenges. *Appl. Math. Chin. Univ*. 18:332–342.