Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Ensemble machine learning technique-based plagiarism detection over opinions in social media

Sethu Vinayaga Vadivu, Palanigurupackiam Nagaraj & Bagavathi Ammai Shanmugam Murugan

To cite this article: Sethu Vinayaga Vadivu, Palanigurupackiam Nagaraj & Bagavathi Ammai Shanmugam Murugan (2024) Ensemble machine learning technique-based plagiarism detection over opinions in social media, Automatika, 65:3, 983-991, DOI: 10.1080/00051144.2024.2326383

To link to this article: https://doi.org/10.1080/00051144.2024.2326383

Published online: 15 Mar 2024.

AUTOMATIKA
2024, VOL. 65, NO. 3, 983–991
https://doi.org/10.1080/00051144.2024.2326383

Ensemble machine learning technique-based plagiarism detection over opinions in social media

Sethu Vinayaga Vadivu[a], Palanigurupackiam Nagaraj[a] and Bagavathi Ammai Shanmugam Murugan[b]

[a]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, India; [b]Department of Computer Science and Engineering, M Kumarasamy College of Engineering and Technology, Karur, India

ABSTRACT

With the progressive enhancement of social media, several people prefer posting their opinions on various social media instead of posting on radios, television or newspapers. The postings differ in dimensions and include various titles and comments. Nowadays, the formation of plagiarism is increasing tremendously which occurs by rewriting or repeating one's work. There are many ways to detect plagiarism by browsing through the internet. The significant intention of this paper involves the detection of plagiarism in social media using four different phases, namely the data pre-processing phase, *n*-gram evaluation, similarity/distance computation analysis and the plagiarism detection phase. The pre-processing includes data cleaning processes, such as the removal of redundant data, upper case letters, noise, irrelevant punctuations and characterizing into a vector form. After pre-processing the data are fed for *n*-gram evaluation to develop a posting attribution system. Then finally, an ensemble support vector machine-based African vulture optimization (ESVM-AVO) approach is employed to detect plagiarism which signifies that the performance based on detection is enhanced and the execution time in obtaining a high rate of detection accuracy is very low. Finally, the performance evaluation and the comparative analysis are carried out to determine the performance of the proposed system.

ARTICLE HISTORY
Received 2 February 2024
Accepted 27 February 2024

KEYWORDS
Plagiarism; *n*-gram; support vector machine; African vulture optimization; opinion mining; social media

1. Introduction

Nowadays, the spreading of gossip and false messages may be half-tale discovering new techniques to deliver the news to the people. The capacity to recognize the trade scam is rising. Using doormat accounts, the messages spread as long as possible and one methodology frequently employed in the website is Twitter which has enhanced signals for messaging. The posting of the same or approximately the same messages of various accounts of the user is governed by a single user. The post was connected mechanically from a distrustful user account and the messages were reviewed. This investigation is done by implementing a similar procedure of plagiarism identification. The objective of this study is to detect the collision of Twitter accounts using plagiarism detection [1]. It is a difficult task to process the data. There are many ways to detect plagiarism by browsing through the internet. The sentence matching identification is the linguistic comparability measurement.

Accordingly, if the outcome of the report is similar between two queries they are said to be semantically equivalent [2]. There are some complications in text processing using the machine. The machine finds it difficult to understand the meaning and words. The sentence with the company name Apple is recognized wrongly by the machine as Apple fruit. This happens because the machine fails to get the meaning of the sentence. Many machine learning methods are employed for the identification of paraphrases. The Recurrent Neural network (RNN) model is popular in machine learning [3]. The extensive accessibility of webcasts modified the education process. The formation of plagiarism is increasing tremendously, it occurs by rewriting or repeating one's work. Another method, the detection of plagiarism in four stages is done in an orderly to find the resemblance of papers. Normally, the four steps are done by individual research. But in this method, the first two steps, collection and analysis, are mechanical processes then the last two steps, confirmation and investigation, are partly mechanical processes. If the analysis section has plagiarism, it is examined in the next process.

The submission of the report is checked by the human investigator [4]. This makes the detection process effective. Social media postings have various complications in using high-level programming language and in managing performance analysis such as articles and summaries of works. The postings differ in dimensions and include various topics. The author's

CONTACT Sethu Vinayaga Vadivu vinayagavadivus123@gmail.com Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil 626126, India

view of writing is afflicted by diverse titles and comments. Social media postings lack of data in comparing with other automatic technology. Many methods are employed for text attribution. To avoid complicated analysis the simple classification is proposed by the researchers. This method is applicable not only for verification but also for profiling purposes [5]. There are physical and behavioural approaches. The physical approach is unique among persons based on their profile, fingerprint and hand geometry. It is used because of its high precision. It has limitations as it gets eliminated from the real framework. The behavioural approach is that it depends on the user. By the extraction of high-level characteristics, style of writing, structure of index and reflecting the author's strategies, their features are formulated. The statistical classifiers and *n*-gram approach are used to determine the resemblance of writings in social media postings. The outcome of these processes shows the effectual authorship of text in social media on different topics [6]. This paper proposes an ensemble support vector machine-based African vulture optimization (ESVM-AVO) approach to detect plagiarism from the original data using the *n*-gram data mining technique. The major contribution of the paper is delineated as follows.

- The *n*-gram analysis is performed in the plain text of the aggregated data for every profile
- To propose an ensemble support vector machine-based African vulture optimization (ESVM-AVO) approach to detect plagiarism
- Evaluating the proposed approach with SVM and NB classifiers to determine the detection accuracy rate.

The rest of the paper is structured as follows. Section 2 depicts the past literature review regarding plagiarism detection. The proposed methodology to detect plagiarism is presented in Section 3. Section 4 illustrates the results and analysis to predict the accuracy rate. Finally, the conclusion of the paper is presented in Section 5.

2. Literature review

Albrektsson et al. [7] suggested that the identification of plagiarism is done by finding the similitude between the text and messages. The growing complexity of copying one's report increased the methods for plagiarism detection and algorithms. The frequently used technique is matching the full subject of files and reviewing the sentence by numerical methods. Measuring the geometer length is the modern method that is applied to documents. From one document the fragment of the message is taken as a base point and is checked in the next document.

Dumitrina et al. [8] developed scholastic plagiarism in the students are collecting data from the internet for their academic works. The rewriting of academic works is mainly done through the Internet. Fetching data from the Internet is not a problem but modifying and storing the information flatter complications. The researcher conducted an online poll among students. The result shows that translating the text/message in one's own words without the author's consideration is also plagiarism.

Pratama et al. [9] submitted a report on the use of plagiarism checkers in the educational field. The plagiarism detector detects the presence of paraphrase, copy and paste and translation tasks. It encourages the students to describe the theory on their own, reduces plagiarism and attains data originality. With a plagiarism detector, the students can learn the reference and correct the errors. Fernando et al. [9] proposed the detection of homogeneous text by correlating the two documents. Filling the information with the additional neural network has improved features and therefore, the results will have more accuracy.

Francisco et al. [10] proposed the programming plagiarism detection technique. In this, the plagiarism assistant tool pk2 is employed to find and dissuade copy writings. The execution trace is compared on programmes in place of source comparison. During the development process, more guidance is provided by analyzing the negatives.

Waqar et al. [11] introduced a new method of detecting plagiarism. Using a distance matrix each divided document's text was verified. The method of Jaccard is utilized when the span of the source document is larger than the apprehensive document. The whole text processing was efficiently done by NLP approaches and the similitude of text was determined by distance algorithm.

3. Proposed methodology

Figure 1 depicts the schematic workflow of the proposed approach. The blocks involved in the proposed methodology to detect plagiarism in social media are the data pre-processing phase, *n*-gram evaluation, similarity/distance computation analysis as well as the plagiarism detection phase. The detailed description of each respective phase is discussed in the following section.

3.1. Data pre-processing

The initial data pre-processing phase is carried out on data collected from the social media platforms. Generally, users need to log in to various social media platforms such as Facebook, Twitter, Instagram, YouTube, etc. before posting pictures or videos and viewing comments. The users are free to post their opinions on

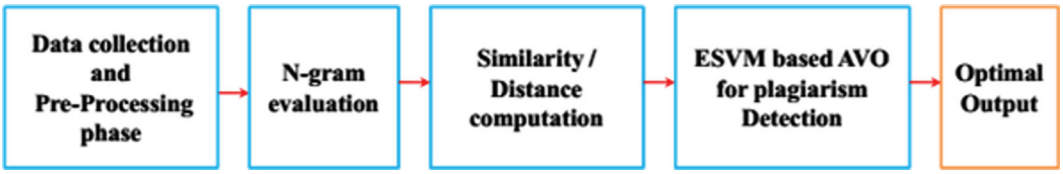

Figure 1. Schematic workflow of the proposed approach.

any subject. It is necessary to collect the comments and user names to analyze the writing style of the author. Usually, the comments will be very short with limited words and it becomes a tedious task to analyze a short single comment. Hence, the posts that are downloaded are collected for every writer following the name of the user. The collected posts are concatenated and provided for pre-processing analysis. The pre-processing includes data cleaning processes such as the removal of redundant data, upper case letters, noise, irrelevant punctuations and characterizing into vector forms. After pre-processing the data are fed for n-gram evaluation to develop a posting attribution system.

3.2. n-gram evaluation

The sequences of n adjacent units are referred to as an n-gram analysis in which the adjacent unit ranges from characters, bytes or bits. The n-gram analysis is performed in the plain text of the aggregated data of every author. The text containing n adjacent units takes a current window and sliding window to extract n-grams. These windows perhaps or perhaps do not overlap with one another. By varying the n value provides diverse sizes of windows (i.e. unigram, bigram, trigrams, etc.). Similarly, for a given post or text, an n-gram evaluation is performed. In addition to this, the total number of occurrences has not become normalized and it can be done by converting the total number of occurrences from the number of tallies mentioned in percentage. Thus the author's profile is normalized using n-gram evaluation for all n values [5].

3.3. Similarity/distance computation analysis

The similarity measures are computed among the n-gram spectra for every n value to analyze each profile by employing an Euclidean distance formula in n-dimensional area. The mathematical expressions for the n-gram spectra in vector (n) form containing 2^N elements are determined in Equation (1) [5].

$$n = (g(0), g(2), \ldots . g(2^{N-1}) \tag{1}$$

Let us assume 2 n-grams $(N = k)$ from two different types of users X and Y. Then

$$n_X = (\alpha(0), \alpha(2), \ldots . \alpha(2^{k-1}) \text{ and}$$
$$n_Y = (\beta(0), \beta(2), \ldots . \beta(2^{k-1}) \tag{2}$$

The similarity value of $2k$-gram for users X and Y is stated in Equation (3).

$$\delta_{(XY)_k} = \sqrt{\sum_{0}^{2^k-1} (X_j - Y_j)^2} \tag{3}$$

3.4. ESVM-based AVO for plagiarism detection

During the plagiarism detection processes, the datasets are split into training and testing datasets with a ratio of 80:20. Here the training datasets are employed to establish the detection model and the testing datasets are utilized to verify the performances of the established detection model. In this paper, an ensemble SVM (ESVM)-based African vulture optimization (AVO) algorithm is employed in detecting plagiarism. The significant steps involved in ESVM and AVO algorithms are discussed in the subsequent section.

3.4.1. Ensemble support vector machine

The ensemble support vector machine (SVM) is one of the classifications of machine learning approaches established by Vapnik et al. in 1990. The SVM can transfer a nonlinear-independent issue into a nonlinear-independent issue using kernel functions. The SVM approach is employed widely to solve various classification problems. The mathematical theory involved in the SVM classifier is stated as follows.

The SVM was established initially to resolve dual-class issues. The significant objective involves determining an optimal theoretical hyperplane for differentiating two different types of samples, namely the positive sample and negative sample. Therefore,

$$\text{Minimum } \psi(\omega) = 0.5\,||\omega||^2 + P_f \sum_{j=1}^{n} S_v \tag{4}$$

Such that

$$Z_j[\omega.k(y_j.y_k) + \beta] + S_v{}^3 1; \quad j = 1 \text{ to } n \tag{5}$$

From Equations (4) and (5), ω and β signify the weight function and bias function, respectively. The penalty factor and the slack variables are represented by P_f and S_v, respectively. $k(y_j.y_k)$ indicates the Gaussian radial basis kernel function that is expressed as

$$k(y_j.y_k) = \text{Exp}\,[-||y_j - y_k||^2/2\mu^2] \tag{6}$$

Equation (6), signifies the kernel width. The optimization issue requires a pair of parameters, namely P_f and μ, respectively. Here, the trade-off between the classification accuracy and the complexity of the classifier is characterized through the penalty factor.

Meanwhile, the single SVM classifier is employed in solving unbalanced datasets, multi-class classification issues, etc. Furthermore, applying complex models, containing high-dimensional noises to the classification issues using a single SVM results in limited consistency of features that further minimizes the prediction accuracy. To conquer such drawbacks of single SVM classifiers, an ensemble-based classifier provides a strong potency in numerous sectors. Hence, this paper utilized ensemble-based SVM classifier to obtain classification results. Figure 2 depicts the schematic architecture of the ensemble-based SVM classifier.

The ensemble SVM (ESVM) approach integrates a single classifier set to enhance the robustness and accuracy of the SVM. The execution of the ensemble SVM approach depends on two significant factors. The first factor involves the construction of every member classifier and the second factor involves the fusing of member classifiers to obtain a strong classifier with high accuracy. The step-by-step process to obtain ensemble classified output is depicted below.

Initially, the member classifier is obtained by the random selection of the training dataset and the rest datasets are considered as temporary testing datasets for performance evaluation. The member classifiers choose the features following the support vector rate. Finally, the output obtained from the member classifiers is combined to fuse the decisions using the weighted voting principle. The obtained ensemble classifier is employed in obtaining the classifying results. The mathematical expression to evaluate the performance based on classification accuracy is obtained in Equation (7).

$$\text{classifiaction accuracy} = \frac{n(\text{correct})}{n(\text{all})} \times 100\% \quad (7)$$

From the above equation, the total number of samples classified correctly and the entire samples are represented by n(correct) and n(all), respectively.

The classification accuracy of every member classifier is obtained by computing the testing dataset. In addition to this, the voting weight α_j of every member classifier is expressed in the following equation.

$$\alpha_j = \frac{\lambda_j}{\sum \lambda_j} \quad \text{where } j = 1 \text{ to } n \quad (8)$$

From Equation (8), the classification accuracy is λ_j. This paper utilized the ensemble SVM classifier containing 10 member classifiers to ensure the balance among the simplicity and diversity of the member classifier. Each SVM layer consists of the weight function that plays a significant role and in this paper, an African vulture optimization (AVO) algorithm is employed in determining the optimal output.

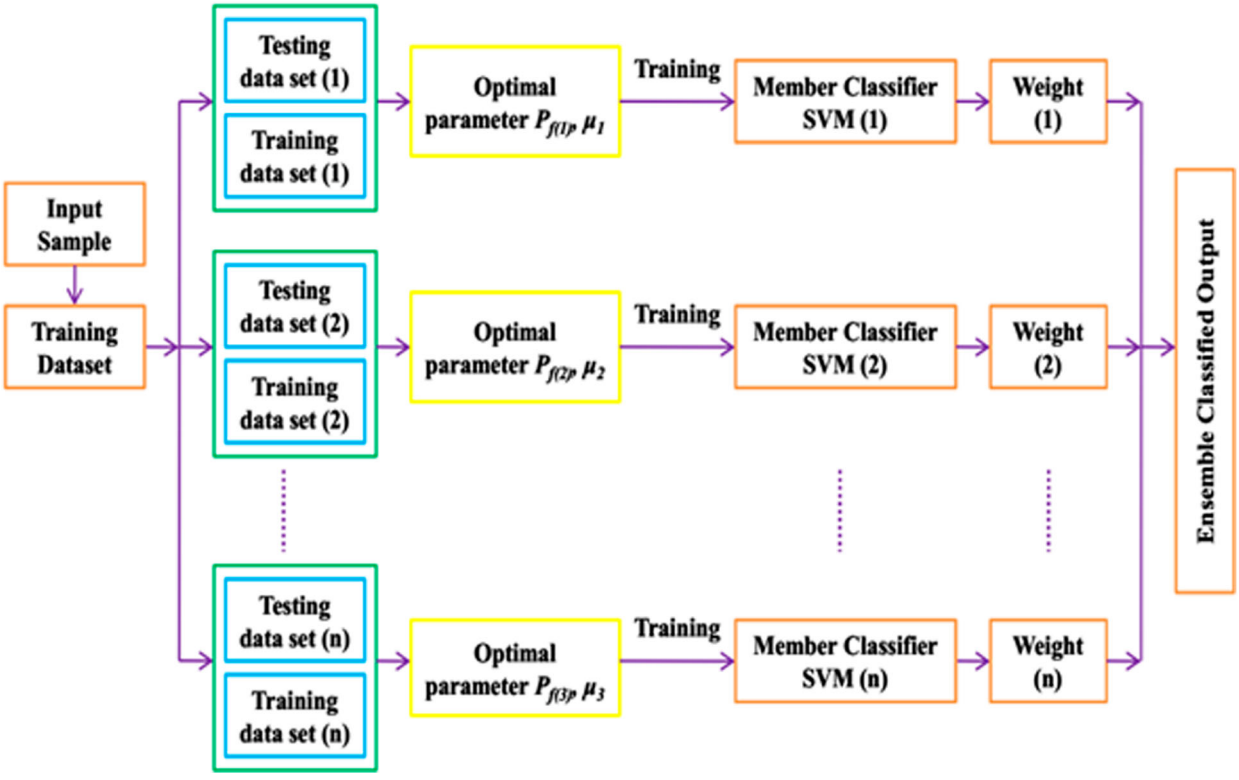

Figure 2. Ensemble SVM architecture.

3.4.2. African vulture optimization (AVO) algorithm

The vultures belong to the hunting bird varieties that are found in Europe, Africa and Asia. Most of the vultures are hairless with irregular feathers and can be found mostly on the battlefield. Nowadays the vultures are considered endangered species and among them African vultures are becoming more extinct. The African vultures consist of a few peculiar features and are classified into three different types: The initial type contains the vultures that are physically stronger than all other vultures. The vultures that are weaker than type 1 vultures are considered the second type vultures. The vultures that are weaker than type 1 and type 2 are categorized under the third type. The algorithm that imitates the features, physical behaviours and battling characteristics of the African vulture is referred to as the African vulture optimization (AVO) algorithm [13]. The step-by-step process and principles involved in the AVO algorithm are discussed as follows.

Step 1: determination of the best vulture. After determining the initial population and fitness function f_j, the best optimal solution is obtained by selecting the first best vulture BV_1 for the first group and the next best vulture BV_2 for the second group. The total population is recomputed for every fitness function. Equation (9) provides the movement of the best solution $K(j)$ for the first and second groups.

$$K(j) = \begin{cases} BV_1 & \text{if } \delta_j = M_1 \\ BV_2 & \text{if } \delta_j = M_2 \end{cases} \tag{9}$$

From the above equation, M_1 and M_2 are the parameters which are to be evaluated before the search process. The probability δ_j of selecting the best solution is obtained by the Roulette wheel technique. Thus,

$$\delta_j = \frac{f_j}{\sum_{j=1}^{N} f_j} \tag{10}$$

Step 2: starvation rate of vultures. The vultures frequently search for food with high energy levels that can search food for long distances. On the other hand, the vultures become aggressive when they feel hungry and they search for food next to the strong vulture. The satiated rate z of vultures is expressed in Equation (11).

$$z = q \times \left(\sin^x \left(\frac{\Pi}{2} \times \frac{itr_j}{\max itr} \right) + \cos \left(\frac{\Pi}{2} \times \frac{itr_j}{\max itr} \right) - 1 \right) \tag{11}$$

From Equation (11)

$$f = (2 \times \Re_1 + 1) \times Z \times \left(1 - \frac{itr_j}{\max itr} \right) + z \tag{12}$$

From the above equations, the current and maximum number of iterations are denoted by *itr* and *maxitr*, respectively. Z and q indicate the random value ranging from [−1,1] and [−2,2], respectively. The parameter x provides a fixed set of numbers during the process of optimization.

Step 3: exploration phase. In the AVO algorithm, the vultures are capable of examining diverse random strategies and the selection of strategy is referred to as δ_1. Before operation, the parameter must be evaluated and it must contain either 0 or 1. A random number that ranges from 0 to 1 is generated by choosing any of the strategies from the random exploration phase $\Re_{\delta 1}$. Equation (14) is employed if the number is greater than or equal to δ_1. Else, Equation (16) is selected.

$$\delta(j+1) = \begin{cases} \text{eqn (14)} & \text{if } \delta_1 \geq \Re_{\delta 1} \\ \text{eqn (16)} & \text{if } \delta_1 < \Re_{\delta 1} \end{cases} \tag{13}$$

$$\delta(j+1) = K(j) - E(j) \times f \tag{14}$$

From Equation (14)

$$E(j) = |Y \times K(j) - \delta(j)| \tag{15}$$

From the above equations, the position vector of the vulture is denoted by $\delta(j+1)$. f indicates the satiated vulture rate. One of the best vultures is denoted by $K(j)$. The vultures moving randomly to save the food from the neighbouring vultures are denoted by Y. Then the current position vector of the vulture is expressed in the following equation.

$$\delta(j+1) = K(j) - f + \Re_2 \times ((Lb - Ub) \times \Re_3 + Lb) \tag{16}$$

From Equation (16), the random value ranges from 0 to 1 are $\Re_2$ and $\Re_3$, respectively. Lb and Ub signify the upper and lower boundary rates, respectively.

Step 4: exploitation phase. The efficiency of the AVO algorithm is determined in the exploitation phase. The exploitation phase is carried out under two different strategies, namely the rotating flight strategy and the siege flight strategy. Before operation, the parameter must be evaluated and it must contain either 0 or 1. A random number that ranges from 0 to 1 is generated by choosing any of the strategies from the random exploitation phase $\Re_{\delta 2}$. The rotating flight strategy is employed if the number is greater than or equal to δ_1. Otherwise, the siege flight strategy is selected.

$$\delta(j+1) = \begin{cases} \text{eqn (10)} & \text{if } \delta_2 \geq \Re_{\delta 2} \\ \text{eqn (13)} & \text{if } \delta_2 < \Re_{\delta 2} \end{cases} \tag{17}$$

a. Struggle for foodstuff

When several vultures depend on one source of food, there occurs a conflict among the vultures. The vulture with strong physical energy does not share the food whereas the weaker vultures fight with the strongest vultures for the foodstuff. The conflicts among various vultures for the foodstuff are determined in Equations (18) and (19).

$$\delta(j+1) = E(j) \times (f + \Re_4) - g(z)c \tag{18}$$

$$g(z) = K(j) - \delta(j) \tag{19}$$

From the above equation, $\Re_4$ signifies the random number ranges from 0 to 1.

b. Vultures revolving voyage

The spiral motion is determined by the vultures by frequent rotational flight. The spiral expression is established among the two best vultures. Thus,

$$r_1 = K(j) \times \left(\frac{\Re_5 \times \delta(j)}{2\Pi}\right) \times \cos(\delta(j)) \tag{20}$$

$$r_2 = K(j) \times \left(\frac{\Re_6 \times \delta(j)}{2\Pi}\right) \times \sin(\delta(j)) \tag{21}$$

$$\delta(j+1) = K(j) - (r_1 + r_2) \tag{22}$$

In the above equations, $\Re_5$ and $\Re_6$ signify the random number ranging from 0 to 1. Equations (20) and (21) provide the value for r_1 and r_2, respectively.

Step 5: the second phase of exploitation. During this phase, aggressive and siege strife are performed to identify the food source.

$$\delta(j+1) = \begin{cases} \text{eqn } 26 & \text{if } \delta_3 \geq \Re_{\delta 3} \\ \text{eqn } 27 & \text{if } \delta_3 < \Re_{\delta 3} \end{cases} \tag{23}$$

In addition to this, the vulture movement towards the food is investigated. Equations (24) and (25) provide the movement formulation of vultures.

$$C_1 = BV_1(j) - \frac{BV_1(j) \times \delta(j)}{BV_1(j) - \delta(j)^2} \times f \tag{24}$$

$$C_2 = BV_2(j) - \frac{BV_2(j) \times \delta(j)}{BV_2(j) - \delta(j)^2} \times f \tag{25}$$

In the above equations, $BV_1(j)$ and $BV_2(j)$ indicate the best vulture of the first and second groups in the present iteration. The vector position of the vulture for the preceding iteration $\delta(j+1)$ is derived in Equation (26). Thus,

$$\delta(j+1) = \frac{C_1 + C_2}{2} \tag{26}$$

On the contrary, the vultures move in various directions that are expressed as

$$\delta(j+1) = K(j) - |g(z) \times f \times Ly(g) \tag{27}$$

From Equation (27), the distance between the vultures is represented by $g(z)$. $Ly(g)$ indicates the levy flight that is determined in Equation (28).

$$Lyf(e) = 0.01 \times \frac{\nu \times \rho}{\left|i\right|^{\frac{1}{\alpha}}},$$

$$\rho = \left(\frac{\eta(1+\alpha) \times \sin\left(\frac{\Pi\alpha}{2}\right)}{\eta(1+\alpha_2) \times \alpha \times 2\left(\frac{\alpha-1}{2}\right)}\right)^{\frac{1}{\alpha}} \tag{28}$$

From the above equation, the random numbers ranging from 0 to 1 are denoted by u and i, respectively. The fixed value ranging from 1.5 is denoted by α.

Thus, the proposed ESVM-based AVO approach to detect plagiarism signifies that the performance based on detection is enhanced and the execution time in

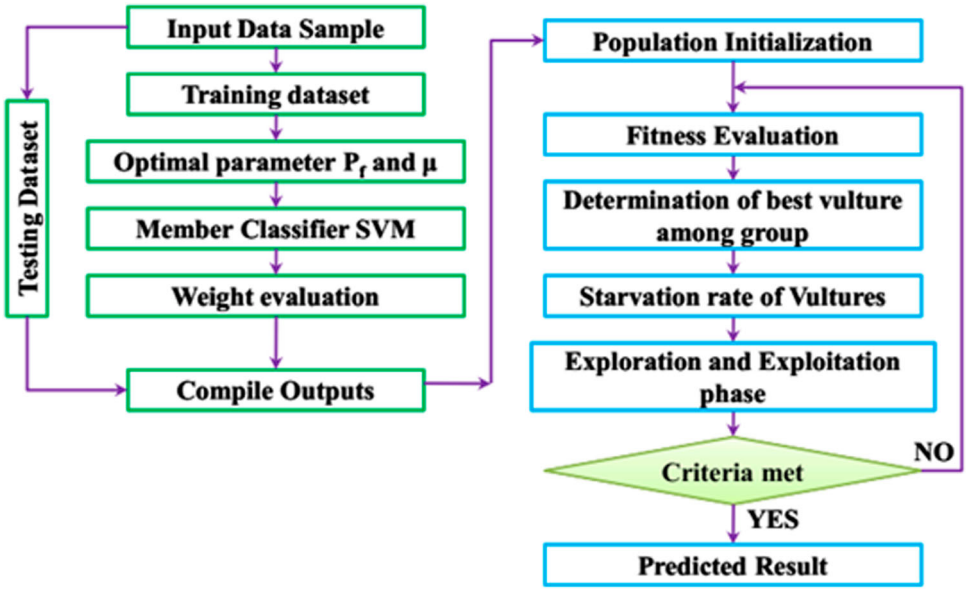

Figure 3. ESVM-based AVO for plagiarism detection.

obtaining a high rate of detection accuracy is very low. Figure 3 depicts the flow diagram of the proposed ESVM-based AVO for plagiarism detection.

4. Results and discussions

This section depicts the evaluation results for various simulation parameters to determine the effectiveness of the proposed approach. In addition to this, the upcoming section portrays the comparative analysis of other existing detection techniques with our proposed approach. The dataset utilised for experimental investigation is real-time Twitter dataset.

4.1. Parameter settings of the proposed approach

Table 1 depicts the parameter configurations involved in the ESVM-AVO approach. The parameters and their specific ranges are discussed below.

Table 1. Parameter settings of the ESVM-AVO approach.

Parameters	Ranges
Size of population	50
Maximum number of iterations	100
Member classifier	10
Gaussian radial kernel rate	0.5
M_1	0.7
M_2	0.2
δ_1	2.5
δ_2	0.7
δ_3	0.5

4.2. Evaluation measures

This section provides the mathematical formulations of various simulation measures, namely the accuracy, precision, recall and f-measure to evaluate the performances of the proposed ESVM-AVO approach.

$$\text{Accuracy} = \frac{Tr(\text{Pos}) + Tr(\text{Neg})}{Tr(\text{Pos}) + Fa(\text{Pos}) + Fa(\text{Neg}) + Tr(\text{Neg})} \tag{29}$$

$$\text{Recall} = \frac{Tr(\text{Pos})}{Tr(\text{Pos}) + Fa(\text{Neg})} \tag{30}$$

$$\text{Precision} = \frac{Tr(\text{Pos})}{Tr(\text{Pos}) + Fa(\text{Pos})} \tag{31}$$

$$F - \text{measure} = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{32}$$

4.3. Performance analysis

The evaluation results of the ESVM-AVO approach for various performance measures, namely accuracy, precision, recall and f-measure, are depicted in Figure 4. The performance values are validated and the resulting outcome revealed that this approach provides 95.29% of accuracy, 93.67% of precision, 91.58% of recall and 93.29% of f-measure, respectively. Thus from the evaluation results, the resulting outcome revealed that the proposed approach provides a better detection accuracy rate for various simulation measures.

Figure 5(a)–(d) depicts the graphical analysis based on various performance measures, such as accuracy, precision, recall and f-measure for various approaches,

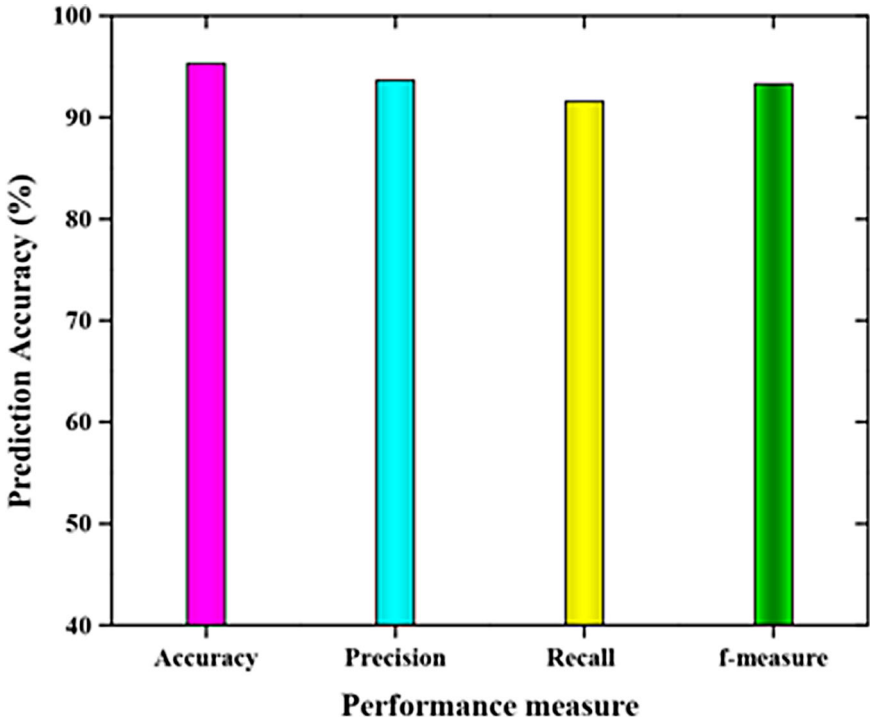

Figure 4. Performance analysis of the proposed approach.

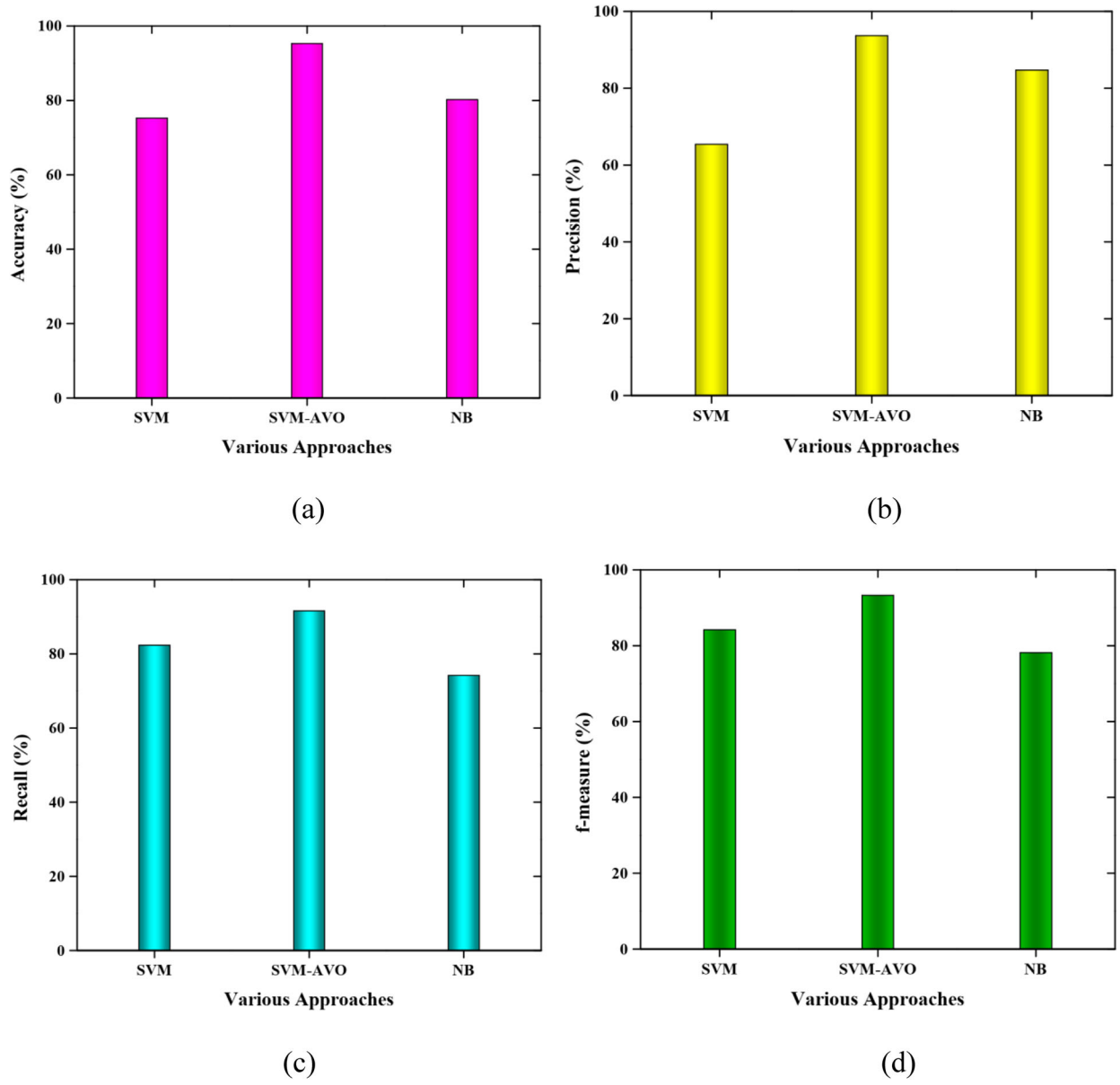

Figure 5. Comparative analysis for (a) accuracy, (b) precision, (c) recall, and (d) *f*-measure.

namely Naive Bayes [12], support vector machine (SVM) [14] and the proposed approach, to detect the performances. The performance rates are evaluated and the analysis revealed that the proposed approach provides higher performance than other existing plagiarism detection approaches.

5. Conclusion

This paper proposed an ensemble support vector machine-based African vulture optimization (ESVM-AVO) approach to detect plagiarism. Usually, the comments posted by the user on social media will be very short with limited words and it becomes a tedious task to analyze a short single comment. Nowadays, the formation of plagiarism is increasing tremendously which occurs by rewriting or repeating one's work. Therefore, this paper utilized four different phases to detect plagiarism with a high rate of accuracy. During the plagiarism detection processes, the datasets are split into training and testing datasets with a ratio of 80:20. Here the training datasets are employed to establish the detection model and the testing datasets are utilized to verify the performances of the established detection model. Finally, the experimental evaluation and comparative analysis are performed for various simulation measures, namely the accuracy, precision, recall and *f*-measure to evaluate the performances of the proposed ESVM-AVO approach. Thus from the evaluation results, the resulting outcome revealed that the proposed approach provides a better detection accuracy rate for various simulation measures. In the future, a novel hybrid metaheuristic algorithm will be proposed to enhance the robustness and accuracy and minimize the error value rate.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Albrektsson F. Detecting sockpuppets in social media with plagiarism detection algorithms; 2017.

[2] Kurniawan MA, Surendro K. Similarity measurement algorithms of writing and image for plagiarism on Facebook's social media. IOP Conf Ser: Mater Sci Eng, IOP Publ. 2018;403(1):012074. doi:10.1088/1757-899X/403/1/012074

[3] Hunt E, Janamsetty R, Kinares C, et al. Machine learning models for paraphrase identification and its applications on plagiarism detection. 2019 IEEE International Conference on Big Knowledge (ICBK), IEEE; 2019. p. 97–104.

[4] Zrnec A, Lavbič D. Social network aided plagiarism detection. Br J Educ Technol. 2017;48(1):113–128. doi:10.1111/bjet.12345

[5] Peng J, Choo K-KR, Ashman H. Bit-level n-gram based forensic authorship analysis on social media: identifying individuals from linguistic profiles. J Netw Comput Appl. 2016;70:171–182. doi:10.1016/j.jnca.2016.04.001

[6] Chu SKW. Developing 21st century skills with plagiarism-free inquiry learning, collaborative teaching, social media, and gamification. Learning and Teaching Expo; 2014.

[7] Olivia-Dumitrina N, Casanovas M, Capdevila Y. Academic writing and the internet: cyber-plagiarism amongst university students. J New Approaches Educa Res (NAER Journal). 2019;8(2):112–125.

[8] Pratama H, Prastyaningrum I. Effectiveness of the use of Integrated Project Based Learning model, Telegram messenger, and plagiarism checker on learning outcomes. J Phys: Conf Ser, IOP Publ. 2019;1171(1):012033. doi:10.1088/1742-6596/1171/1/012033

[9] Fernando S, Stevenson M. A semantic similarity approach to paraphrase detection. Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics; 2008. p. 45–52.

[10] Rosales F, García A, Rodríguez S, et al. Detection of plagiarism in programming assignments. IEEE Trans Educ. 2008;51(2):174–183. doi:10.1109/TE.2007.906778

[11] Ali W, Ahmad T, Rehman Z, et al. A novel framework for plagiarism detection: a case study for Urdu language. 2018 24th International Conference on Automation and Computing (ICAC), IEEE; 2018. p. 1–6.

[12] Hemlatha AM, Subha M. A study on plagiarism checking with appropriate algorithm in data mining. Int J Res Comput Appl Robot. 2014;2(11):50–58.

[13] Abdollahzadeh B, Gharehchopogh FS, Mirjalili S. African vultures optimization algorithm: A new nature-inspired metaheuristic algorithm for global optimization problems. Comput Ind Eng. 2021;158:107408. doi:10.1016/j.cie.2021.107408

[14] Wang R, Li W, Li R, et al. Automatic blur type classification via ensemble SVM. Signal Process, Image Commun. 2019;71:24–35. doi:10.1016/j.image.2018.08.003