Single Valued Neutrosophic Sets with Optimal Support Vector Machine for Sentiment Analysis

Mohammed I. Alghamdi

Department of Computer Science, Al-Baha University, Al-Baha City, Kingdom of Saudi Arabia

Email: mialmushilah@bu.edu.sa

Abstract

Sentiment analysis (SA) is mainly employed to investigate the polarity of the sentiment existent in a content. It assist in understanding the opinions or feelings expressed by people. It find useful in several application areas such as e-commerce, education, etc. Natural language processing (NLP) and machine learning tools can be employed for SA. In this view, this study develops a Single Valued Neutrosophic Sets with Optimal Support Vector Machine (SVNS-OSVM) model for SA. The major intention of the SVNS-OSVM model is to identify the existence of sentiments exist in the data. The SVNS-OSVM model initially performs data pre-processing to transform the input data into a useful format. In addition, SVNS model is applied to derive word embeddings. Then, SVM model is applied for the detection and classification of sentiments. At the last level, the improved particle swarm optimization (IPSO) algorithm is used to fine tune the parameters involved in the SVM model. For ensuring the improved outcomes of the SVNS-OSVM model, a wide range of simulations were performed and the results are inspected under several aspects. The comparative study highlighted the betterment of the SVNS-OSVM model compared to recent approaches.

Keywords: Sentiment analysis; Classification; Machine learning; Neutrosophic Sets; Support Vector machine.

1. Introduction

Sentiment analysis (SA) or opinion mining is the computational study of individual perspective, sentiments, feelings, appraisals, and mentalities towards elements like items, administrations, associations, people, issues, occasions, themes, and their characteristics [1]. The commencement and fast development of the field match with those of the web-based media on the Web, e.g., surveys, gathering conversations, sites, microblogs, Twitter, and interpersonal organizations, in light of the fact that without precedent for mankind's set of experiences, we have an enormous volume of stubborn information recorded in advanced structures [2]. SA has become one of the most dynamic examination regions in natural language processing (NLP). This expansion is because of the way that suppositions are fundamental to practically all human exercises and are key forces to be reckoned with of our practices. These days, to purchase a purchaser item, one is not generally restricted to approaching one's loved ones for conclusions since there are numerous client audits and conversations regarding the item in open gatherings on the Web [3, 4].

For an association, it might presently not be necessary to lead reviews, assessments of public sentiment, and center gatherings to assemble popular suppositions since there is an overflow of such data freely accessible [5]. As of late, we have seen that stubborn postings in web-based media have reshaped organizations, and influence public sentiments and feelings, which have significantly affected on our social and political frameworks. Such postings have likewise prepared masses for political changes, for

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

example, those occurred in some Arab nations in 2011 [6]. It has along these lines turned into a need to gather and concentrate on suppositions. Notwithstanding, finding and observing assessment destinations on the Web and refining the data contained in them stays an imposing assignment due to the expansion of different locales [7]. Each website ordinarily contains an enormous volume of assessment text that isn't in every case handily translated in lengthy sites and discussion postings. The normal human will experience issues recognizing pertinent destinations and removing and summing up the sentiments in them [8]. Computerized SA frameworks are consequently required. Along these lines, there are numerous new businesses zeroing in on giving SA administrations. Several organizations have likewise constructed their own in-house capacities. These pragmatic applications and modern interests have given solid inspirations to explore in SA. Existing exploration has created various strategies for different assignments of SA, which incorporate both directed and solo techniques [9, 10].

Yang et al. [11] presented a novel SA method-SLCABG, that is depending on the sentimental lexicon and integrates attention-based Bidirectional Gated Recurrent Unit (BiGRU) and Convolution Neural Network (CNN). Interms of method, the SLCABG models combine the advantage of deep learning and sentiment lexicon technologies, and overcome the shortcoming of current SA of product review. Phan et al. [12] developed a novel method-based feature ensemble technique associated with twitters contain fuzzy sentiment by considering fundamentals like word-type, lexical, position, sentiment polarity, and semantic of words. The suggested technique was investigated on actual information. Basiri et al. [13] presented an Attention-based Bidirectional CNN-RNN Deep Method (ABCDM). With this two autonomous GRU bi-directional and LSTM layers, the presented method extracts previous and upcoming context by taking temporal data flow in two directions into account. As well, the attention model is employed on the output of bi-directional layer of ABCDM to place relatively importance on distinct words. The researchers in [14] employ topic recognition and SA explores a massive amount of twitters in both countries with a higher amount of spreading and deaths in COVID19. They rated ten topics and studied the context deliberated on Twitter for 4 months offering an calculation of the progress development.

This study develops a Single Valued Neutrosophic Sets with Optimal Support Vector Machine (SVNS-OSVM) model for SA. The major intention of the SVNS-OSVM model is to identify the existence of sentiments exist in the data. The SVNS-OSVM model initially performs data pre-processing to transform the input data into a useful format. In addition, SVNS model is applied to derive word embeddings. Then, SVM model is applied for the detection and classification of sentiments. At the last level, the improved particle swarm optimization (IPSO) algorithm is used to fine tune the parameters involved in the SVM model. For ensuring the improved outcomes of the SVNS-OSVM model, a wide range of simulations were performed and the results are inspected under several aspects.

2. Design of SNVS-OSVM Model

In this study, a new SVNS-OSVM model has been developed to identify the existence of sentiments exist in the data. The SVNS-OSVM model initially performs data pre-processing to transform the input data into a useful format. In addition, SVNS model is applied to derive word embeddings. Then, SVM model is applied for the detection and classification of sentiments. At the last level, the IPSO algorithm is used to fine tune the parameters involved in the SVM model.

2.1 Process involved in SVNS Model

SVNS value, as determined in [15] a membership function for entity. A neutrosophic set $A \in X$ considered as an indeterminacy membership function IA, truth-membership function TA and a falsity-membership function FA. The membership function ranges from (0,1). It characterizes the likelihood of component X belongs to class and should be autonomous. For SVNS computation, we utilize the feature vector extracted from the intermediatory layer of trained neural network. This vector provides a feature-rich arithmetical depiction of twitter instance. As this vector is extracted from model trained through supervised information, sample belongs to the similar class might contain same characteristics. Then utilize the feature for separating the feature into cluster respective to the corresponding class. Later, employed k-means clustering model. When, we attain the cluster, we discover the cluster centre by averaging the feature vector. Next, determine SVNS through this feature vector and cluster centre.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

Consider the cluster centre as $(C_{positive}, C_{negative}, C_{neutral})$, SVNS value for each sample as $(SVNS_{positive}, SVNS_{negative}, SVNS_{neutral}$ and then feature vector F of N samples:

$$SVNS_{positive} = \frac{Cosine\_distance(C_{positive}, F_n)}{2} \quad (1)$$

$$SVNS_{negative} = \frac{Cosine\_distance(C_{negative}, F_n)}{2} \quad (2)$$

$$SVNS_{neutral} = \frac{Cosine\_distance(C_{neutral}, F_n)}{2} \quad (3)$$

$$Cosine\ distance = 1 - Cosine\_similarity(C_\chi, F_n) \quad (4)$$

for $x \in$ (positive, negative, neutral) and $n \in N$, the amount of instances.

2.2 SVM based Classification

In this study, the SVM classifier is used to analyse the sentiments. Assume a binary classification process: $\{xi, yi\}$, $i = 1, \dots, l, yi \in \{-1,1\}$, $xi \in R^d$, whereas xi denotes data point, and yi indicates the respective labels. They are divided as a hyperplane as $w^T x + b = 0$, in which w denotes a d-dimension coefficient vector viz. standard to the hyperplane, and b indicates the offset from the origin, as shown in Fig. 1. The linear SVM attains an optimum separation margin by resolving the subsequent optimization process [16]:

$$Ming(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (5)$$

$$s.t., y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

By presenting Lagrangian multiplier $\alpha_i (i = 1,2, \dots, n)$, the prime issue is minimized to a Lagarangian dual issue:

$$\max_\alpha \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j x_i^T X_j \quad (6)$$

$$s.t., \alpha_i \geq 0, i = 1, \dots, n, \sum_{i=1}^{n}\alpha_i\, y_i = 0$$

It is clear that a quadratic optimization issue with linear constraint. From the KarushKuhnTucker (KKT) conditions, consider: $\alpha_i(y_i(w^T x_i + b) - 1) = 0$. When $\alpha_i > 0$. The respective data point is named SV. Therefore, the solution takes the subsequent formula: $w = \sum_{i=1}^{n}\alpha_i\, y_i x_i$, in which n indicates the amount of SVs. Here, b is attained from $y_i(w^T x_i + b) - 1 = 0$, here x_i is SV. Afterward w and b is described, the linear discriminative function is represented as follows.

$$g(x) = sgn\left(\sum_{i=1}^{n}\alpha_i\, y_i x_i^T x + b\right) \quad (7)$$

In each case, the two classes could not be divided linearly. The linear learning machine works well in nonlinear case, a common concept is presented. Specifically, the original input space is mapped with high- dimension feature space in which the trained set is separated linearly:

$$g(x) = sgn(\sum_{i=1}^{n}\alpha_i\, y_i \emptyset(x_i)^T \emptyset(x) + b) \quad (8)$$

Now $x_i^T x$ indicates the input space is denoted by the form $\emptyset(x_i)^T\emptyset(x)$ indicates the feature space. It isn't essential to know the function form of the mapping $\varphi(X_i)$ because it is indirectly determined by the chosen kernel: $K(x_i, x_j) = \emptyset(x_i)^T\varphi(x_j)$. Therefore, the decision function is denoted by the following equation:

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

$$g(x) = sgn\left(\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b\right) \tag{9}$$

Generally, positive semi- definite function that satisfies Mercer condition [16].

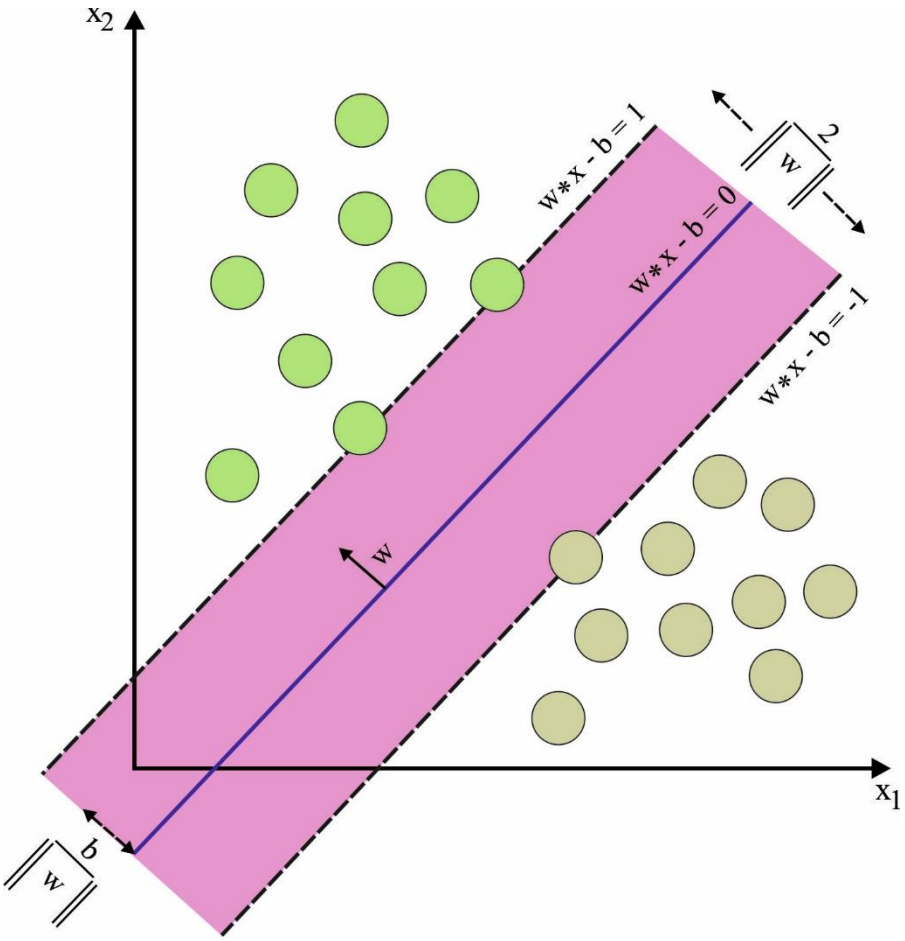

Figure 1: Hyperplanes in SVM

2.3 IPSO based Parameter optimization

In order to proficiently tune the SVM parameters, the IPSO algorithm has been applied. Consider that the particle turned by one dimension and δ. Next, the interrelated position (x) is estimated by the following equation [17]:

$$\mathrm{x} = p \pm \frac{L}{2} \ln\left(\frac{1}{u}\right) \tag{10}$$

in which p denotes the particle motion centre. The variable L should be appropriately calculated through IPSO method whereas such attributes are estimated as follows:

$$L_{\mathrm{i,j}} = 2\beta \cdot \left\| mbest_j - x_{i,j} \right\| \tag{11}$$

In which

$$mbest = \frac{1}{N} \sum_{i=1}^{N} pbest_i \tag{12}$$

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

Here $pbest_i$ denotes the optimal position of individual from a particle x_i and β denotes a Contraction-Expansion (CE) factor [17]. This parameter should be minimalized during the execution process. In IPSO approach, all the particles consume the weighted maximal position of individual past optimal position and better location of group history as the corresponding attractive point. This approximation might lead to particle motion trajectory results. Consequently, the particle is constrained for a rectangle and vertices $pbest_{i,t}$ and$gbest_{i,t}$. The $attractor_{i,t}$ search for $best_{i,t}$. Accordingly, the algorithm doesn't move from local optimal existing in the target level. Next, it can be represented as

$$attractor_{i,t} = u_{\mathrm{i,t}}\, pbest_{i,t} + \left(1 - u_{\mathrm{i,t}}\right) pbest_{b,t} + \Delta_{\mathrm{i},t} \tag{13}$$

In the equation, $u_{i,t}$ characterizes an arbitrary number and distributed function in [0, 1]. The subscript i represent the value of randomly selected particles involving optimum fitness measure. Moreover, the range of particle is selected by $m \in (0,1]$. $\Delta_{\mathrm{i},t} = \left\{\Delta_{i,t}^{1}, \Delta_{i,t}^{2}, \ldots, \Delta_{i,t}^{D}\right\}$ represent a perturbation vector formulated by

$$\Delta_{i,t} = \frac{pbest_{a,t} - pbest_{c,t}}{2} \tag{14}$$

Now b and c subscripts are called as arbitrarily selected particles from the group and a ≠ b ≠ c ≠ i. The upgrade function for particle position in IPSO method is given by

$$x_{i,t} = attractor_{i,t} \pm 2\beta \left\| mbest_j - x_{i,t} \right\| \tag{15}$$

From the changed approach, some information is assumed as an appropriate about individual particle and global optimal position could be misplaced with the algorithm. The workflow of PSO algorithm is shown in Fig. 2.

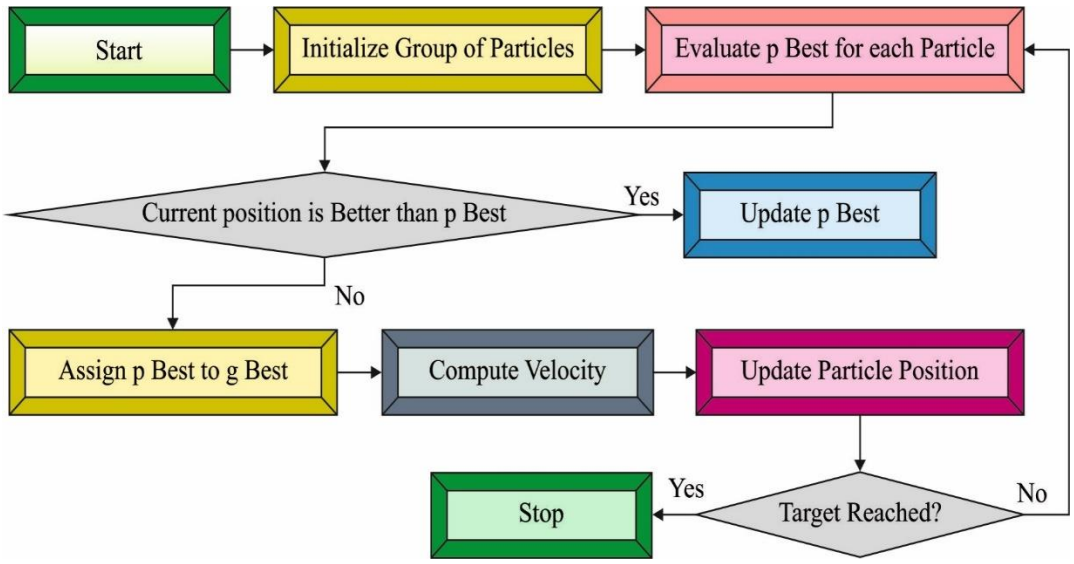

Figure 2: Process involved in PSO algorithm

3. Performance Validation

The performance validation of the SVNS-OSVM model is validated using benchmark dataset [18]. Fig. 3 showcases the confusion matrices offered by the SVNS-OSVM model on the test datasets. Fig. 3a indicates that the SVNS-OSVM model has identified 19109 samples into neural, 14560 samples into positive, and 1047 samples into negative class. Next, Fig. 3b designates that the SVNS-OSVM model has identified 8255 samples into neural, 6205 samples into positive, and 444 samples into negative class on 30% of testing dataset.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

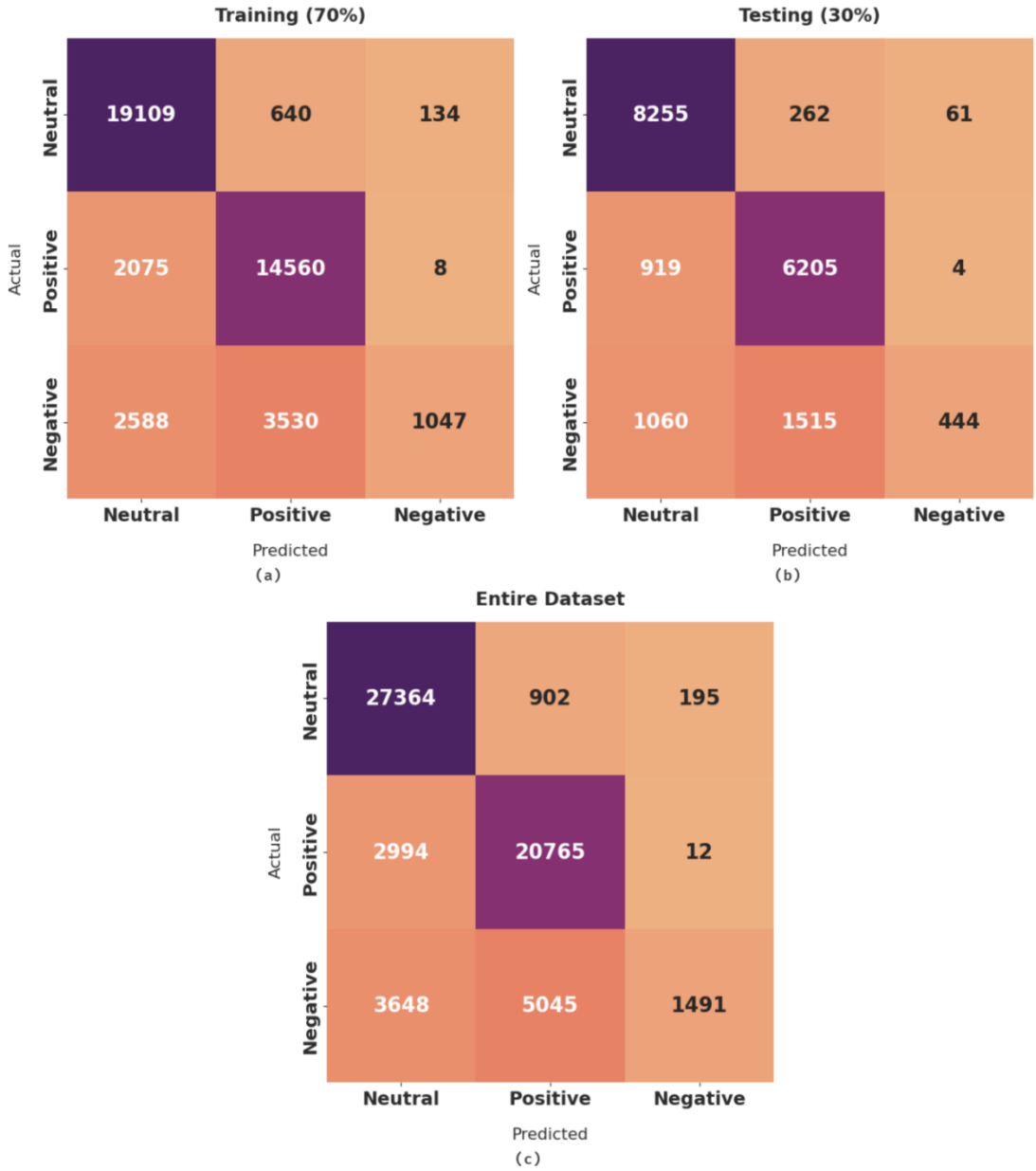

Figure 3: Confusion matrix of SVNS-OSVM model

Table 1 reports a detailed classification outcomes of the SVNS-OSVM model on the test dataset. The experimental values indicated that the SVNS-OSVM model has resulted to maximum outcome on all datasets. On the entire dataset, the SVNS-OSVM model has classified neural instances with $accu_y, prec_n$, $reca_l$, and F_{score} of 87.60%, 80.47%, 96.15%, and 87.61% respectively. Besides, the SVNS-OSVM model has classified positive instances with $accu_y, prec_n$, $reca_l$, and F_{score} of 85.66%, 77.74%, 87.35%, and 82.27% respectively.

Fig. 4 demonstrates an average SA of the SVNS-OSVM model on the entire dataset. The figure indicated that the SVNS-OSVM model has accomplished average $accu_y, prec_n$, $reca_l$, and F_{score} of 86.33%, 82%, 66.05%, and 64.99% respectively. Therefore, it is clear that the SVNS-OSVM model has shown effective outcome on entire dataset.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

Table 1: Classifiers outcomes of SVNS-OSVM model

Class Labels	Accuracy	Precision	Recall	F-Score
Entire Dataset				
Neutral	87.60	80.47	96.15	87.61
Positive	85.66	77.74	87.35	82.27
Negative	85.74	87.81	14.64	25.10
Average	86.33	82.00	66.05	64.99
Training (70%)				
Neutral	87.56	80.38	96.11	87.55
Positive	85.69	77.74	87.48	82.32
Negative	85.67	88.06	14.61	25.07
Average	86.31	82.06	66.07	64.98
Testing (30%)				
Neutral	87.71	80.66	96.23	87.76
Positive	85.58	77.74	87.05	82.13
Negative	85.90	87.23	14.71	25.17
Average	86.40	81.88	66.00	65.02

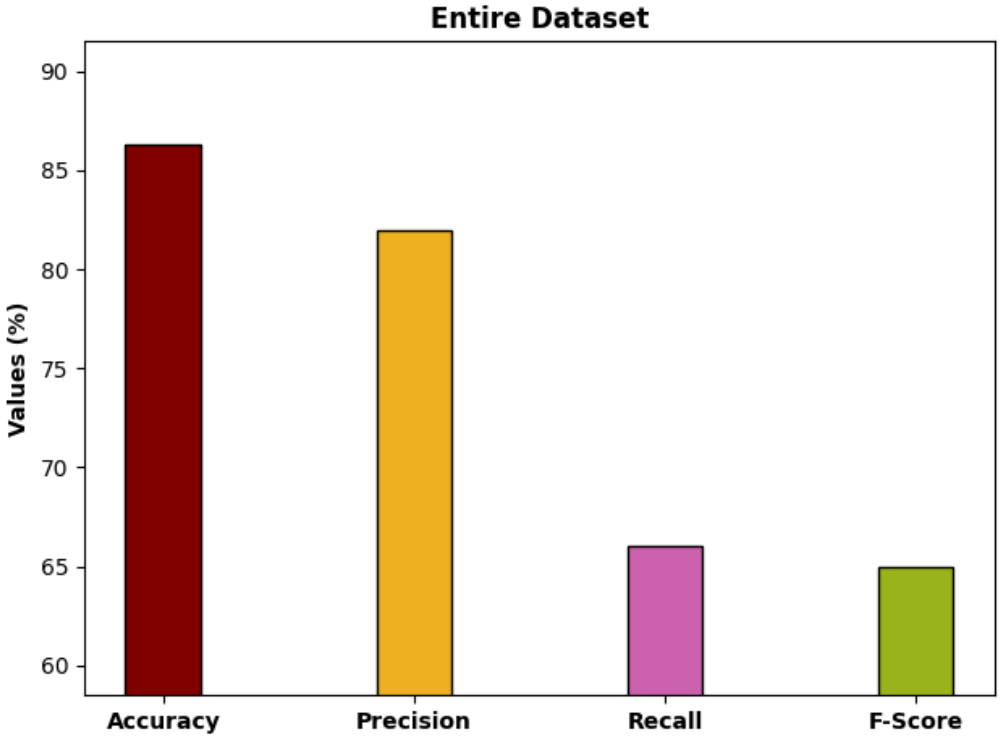

Figure 4: Average SA analysis of SVNS-OSVM model on entire dataset

Fig. 5 validates an average SA of the SVNS-OSVM model on the 70% of training dataset. The figure indicated that the SVNS-OSVM model has attained average $accu_y, prec_n$, $reca_l$, and F_{score} of 86.31%, 82.06%, 66.07%, and 64.98% respectively. Therefore, it is perfect that the SVNS-OSVM model has revealed operative outcome on 70% of training dataset.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

Fig. 6 displays an average SA of the SVNS-OSVM model on 30% of testing dataset. The figure designated that the SVNS-OSVM model has accomplished average $accu_y, prec_n, reca_l$, and F_{score} of 86.40%, 81.88%, 66%, and 65.02% respectively. Therefore, it is clear that the SVNS-OSVM model has shown effective outcome on 30% of testing dataset.

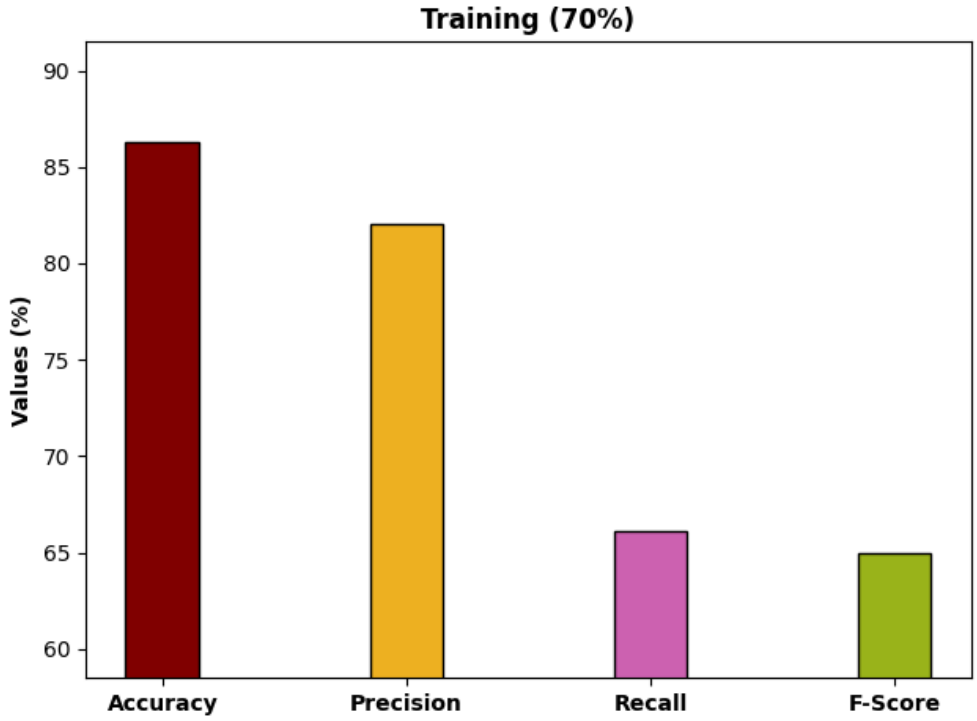

Figure 5: Average SA analysis of SVNS-OSVM model on 70% of training dataset

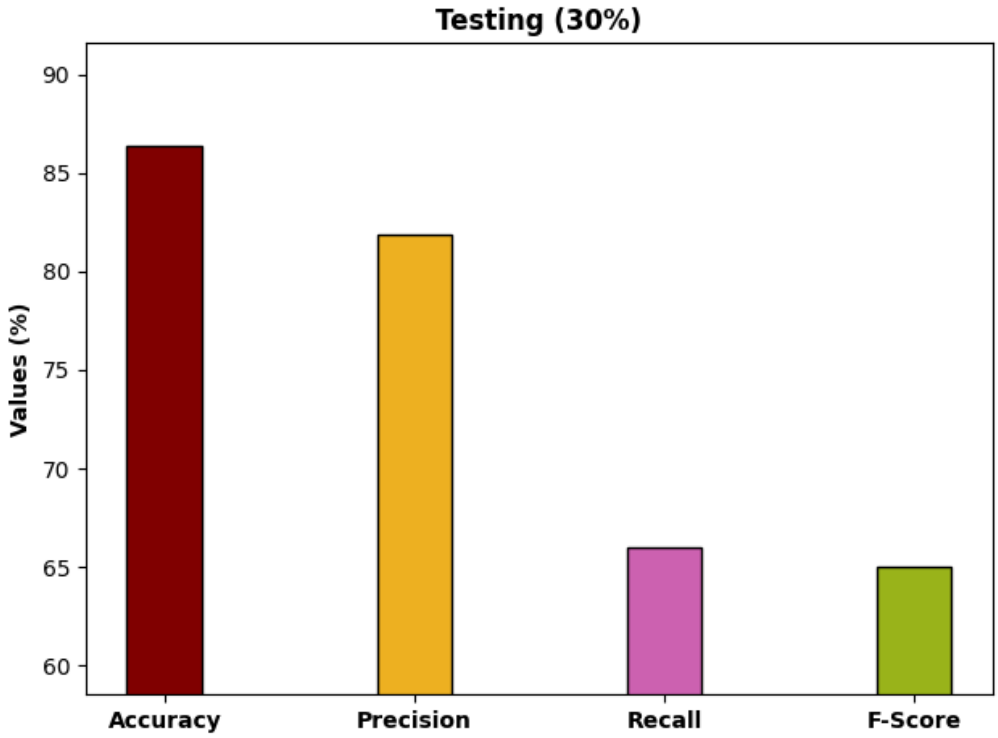

Figure 6: Average SA analysis of SVNS-OSVM model on 30% of testing dataset

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

A detailed comparative SA outcomes of the SVNS-OSVM with recent models are made in Table 2 and Fig. 7 [19]. The results indicated that the BiLSTM-Softmax and BiLSTM-SVNSGRU models have reached lower clasisifcation performance. Followed by, the BERT-Softmax model has resulted to slightly improved classifier results. In line with, the BERT-SVNSCLS model has accomplished even increased classification performance.

Table 2: Comparative SA Results of SVNS-OSVM model

Methods	Precision	Recall	Accuracy	F-Score
BiLSTM-Softmax	69.01	71.51	59.48	60.64
BiLSTM-SVNSGRU	71.12	70.58	59.24	59.79
BERT-Softmax	70.17	72.65	60.49	63.40
BERT-SVNSCLS	71.27	73.99	61.29	63.17
Stacked Ensemble-Softmax	71.69	75.98	62.97	64.15
Stacked Ensemble-PLM	74.41	74.29	63.99	64.75
SVNS-OSVM	86.40	81.88	66.00	65.02

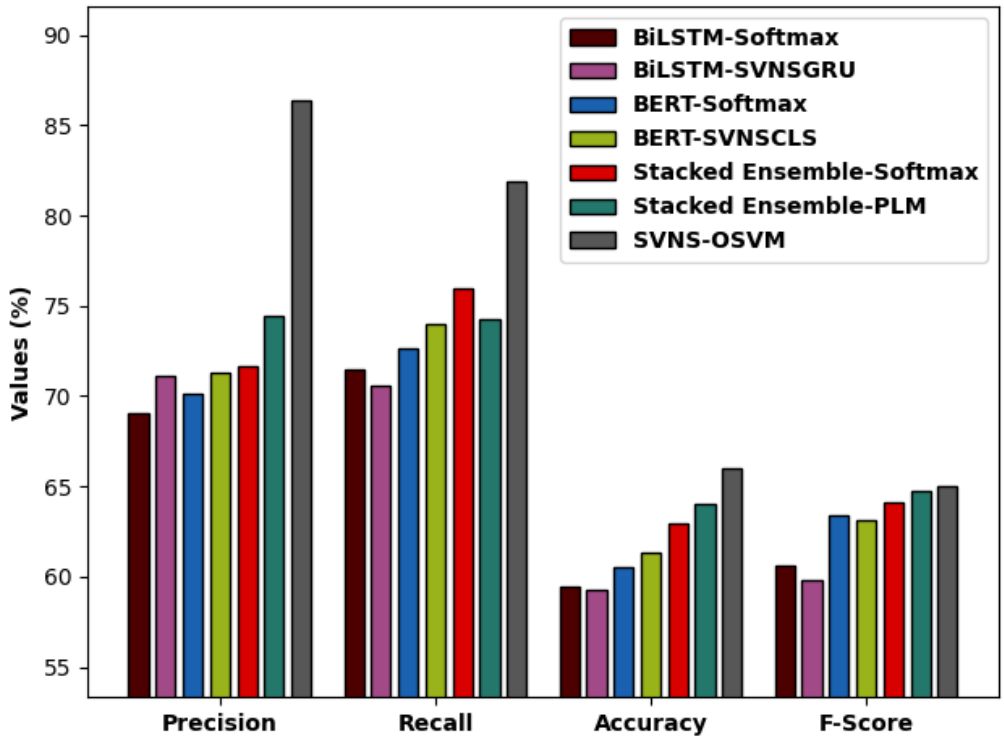

Figure 7: Comparative analysis of SVNS-OSVM model with recent models

Though the Stacked Ensemble-Softmax and Stacked Ensemble-PLM models have accomplished reasonable outcome, the SVNS-OSVM model has resulted to maximum performance with $prec_n, reca_l$, $accu_y$, and F_{score} of 86.40%, 81.88%, 66%, and 65.02% respectively. Therefore, the SVNS-OSVM model has accomplished maximum results over the other techniques.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

3. Conclusion

In this study, a new SVNS-OSVM model has been developed to identify the existence of sentiments exist in the data. The SVNS-OSVM model initially performs data pre-processing to transform the input data into a useful format. In addition, SVNS model is applied to derive word embeddings. Then, SVM model is applied for the detection and classification of sentiments. At the last level, the IPSO algorithm is used to fine tune the parameters involved in the SVM model. For ensuring the improved outcomes of the SVNS-OSVM model, a wide range of simulations were performed and the results are inspected under several aspects. The comparative study highlighted the betterment of the SVNS-OSVM model compared to recent approaches. In future, data clustering techniques can be integrated to the SVNS-OSVM model to improve the classification results.

References

[1] Yadav, A. and Vishwakarma, D.K., 2020. Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), pp.4335-4385.

[2] Dang, N.C., Moreno-García, M.N. and De la Prieta, F., 2020. Sentiment analysis based on deep learning: A comparative study. Electronics, 9(3), p.483.

[3] Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T. and Trajanov, D., 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE access, 8, pp.131662-131682.

[4] Jiang, Q., Chen, L., Xu, R., Ao, X. and Yang, M., 2019, November. A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6280-6285).

[5] Jindal, K. and Aron, R., 2021. A systematic study of sentiment analysis for social media data. Materials today: proceedings.

[6] Phan, M.H. and Ogunbona, P.O., 2020, July. Modelling context and syntactical features for aspect-based sentiment analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 3211-3220).

[7] 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.

[8] Oueslati, O., Cambria, E., HajHmida, M.B. and Ounelli, H., 2020. A review of sentiment analysis research in Arabic language. Future Generation Computer Systems, 112, pp.408-430.

[9] 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.

[10] Habimana, O., Li, Y., Li, R., Gu, X. and Yu, G., 2020. Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences, 63(1), pp.1-36.

[11] Yang, L., Li, Y., Wang, J. and Sherratt, R.S., 2020. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE access, 8, pp.23522-23530.

[12] Phan, H.T., Tran, V.C., Nguyen, N.T. and Hwang, D., 2020. Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. IEEE Access, 8, pp.14630-14641.

[13] Basiri, M.E., Nemati, S., Abdar, M., Cambria, E. and Acharya, U.R., 2021. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Generation Computer Systems, 115, pp.279-294.

[14] Garcia, K. and Berton, L., 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied soft computing, 101, p.107057.

[15] Sharma, M., Kandasamy, I. and Vasantha, W.B., 2021. Comparison of neutrosophic approach to various deep learning models for sentiment analysis. Knowledge-Based Systems, 223, p.107058.

[16] Zhou, J., Qiu, Y., Zhu, S., Armaghani, D.J., Li, C., Nguyen, H. and Yagiz, S., 2021. Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Engineering Applications of Artificial Intelligence, 97, p.104015.

[17] Qin, T., Guo, J., Jing, Z., Han, P. and Qi, B., 2021. Hybrid IPSO-IAGA-BPNN algorithm-based rapid multi-objective optimization of a fully parameterized spaceborne primary mirror. Applied Optics, 60(11), pp.3031-3043.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022

[18] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
[19] Sharma, M., Kandasamy, I. and Vasantha, W.B., 2021. Comparison of neutrosophic approach to various deep learning models for sentiment analysis. Knowledge-Based Systems, 223, p.107058.

Doi: https://doi.org/10.54216/IJNS.1803020
Received: February 26, 2022 Accepted: May 11, 2022