Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhancing Sentiment Analysis: A Novel Machine Learning Framework Integrating Textual Features and Document Metadata

Aishwary Awasthi[1], Dr. Sangeeta Devanathan[2], Ish Kapila[3], Archana Verma[4], Priti Marwah[5], Shikhar Gupta[6]

[1]Research Scholar, Department of Mechanical Engineering, Sanskriti University, Mathura, Uttar Pradesh, India, Email Id- aishwary@sanskriti.edu.in, Orcid Id- 0000-0002-4985-2584
[2]Associate Professor, Department of Marketing, JAIN (Deemed-to-be Univesity), Bangalore, Karnataka, India, Email Id- sangeeta.d@cms.ac.in, Orcid Id-0000000206461153
[3]Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India. ish.kapila.orp@chitkara.edu.in https://orcid.org/0009-0001-0976-6585
[4]Assistant Professor, Master of Computer Applications, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, Email Id-archanaverma.mca@niet.co.in, Orcid Id- 0000-0002-1539-4642
[5]Assistant Professor, Department of Management Studies, Vivekananda Global University, Jaipur, India, Email Id- priti.marwah@vgu.ac.in, Orcid Id- 0009-0002-7310-1067
[6]Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh-174103 India. shikhar.gupta.orp@chitkara.edu.in, https://orcid.org/0009-0004-0138-3987

Introduction: In general, sentiment evaluation has focused on text content analysis. Traditional sentiment analysis (SA) excludes document metadata and focuses on text. The lack of research in this area prevents textual material and metadata from working together to create synergies that could lead to reliable sentiment assessment solutions.
Method: We exhibit the potential uses of our integrated methodology across multiple areas and prove its efficacy through testing and evaluation on a different datasets. Data preprocessed using a multiple techniques, such as “stemming, tokenization, and stop word removal”. The “term frequency-inverse document frequency (TF-IDF)”methodology is utilized for feature extraction.
Result: The outcomes of our experiments reveal that our suggested method is successful in collecting sensitive sentiment such as” accuracy, recall, precision, and f1-score”. This is a promising development in the field of sentiment assessment exploration.
Conclusion: In conclusion, this study proposed a novel approach, the DDO-SVM, which addresses the limitations of conventional SA by incorporating both textual and metadata features. These components integrate well, improving sentiment evaluation across measurements. SA becomes more reliable and effective in actual applications with this breakthrough.

Keywords: Sentiment Assessment, Metadata; Text, TF-IDF, Machine Learning (ML), Dynamic Dragonfly Optimization-Tuned Support Vector Machine (DDO-SVM).

1. Introduction

A wide range of techniques is being investigated in the changing field of data retrieval in an effort to identify accurate and effective ways of gathering data from large databases. The merging of document information and text-based components is a specific method that has increased in popularity. [(1)] It becomes distinct that a document's content and the metadata contribute to creating a synergy that can reveal new levels of effectiveness and understanding in systems for retrieving data. [(2)] The primary purpose of data retrieval is to match users with the most appropriate content. Standard techniques frequently utilize the analysis of text, evaluating the text of the written material to determine its applicability to the user's inquiry. However, the method presents difficulties in handling unclear inquiries or files with intricate construction. [(3)] A complete strategy is offered by combining textual characteristics with document metadata, thereby improving the search procedure through consideration of the internal content and the external data included in the metadata. [(4)] Metadata and text characteristics collectively provide massive potential for system efficiency in addition to being beneficial for end users. By enabling methods to dynamically adjust the changing user behavior and data landscapes, this combination raises the Information Retrieval (IR) technique's overall efficacy and relevancy. [(5, 6)] Recognizing the difficulties this method presents is essential. However, reliable processes and procedures are necessary for the sophisticated management and analysis of an extensive range of metadata from different text sources. [(7, 8)] Assigning values to textual features and information during retrieval is similarly difficult. To ensure that the integration enhances the effectiveness of the IR method rather than diminishes it, this equilibrium must be achieved. [(9)]

Study [(10)] proposed Deep Learning (DL)-based topic-level. It employed web-hidden semantics indexing and normalization restrictions to identify phrase topics. Research [(11)] developed a new DL model that used word embedding and DL approaches to assess and conduct multi-class SA on tweets from six major airlines. Paper [(12)] examined that word count and readability affect sentiment classification. They employed convolution neural network (CNN), long short-term memory (LSTM), Simple Recurrent Network (SRN), and DL methods. Article [(13)] proposed a Hierarchical Deep Fusion (HDF) method for investigating cross-modal word social links. Research [(14)] proposed the "Local Search Improvised Bat" method-based Elman Neural Network for SA of internet evaluations of products. Study [(15)] presented a mixed emotional evaluation method using the "Multi-view Attention Network" (MVAN) and a developing emotional structure to capture the text's deep semantics. Study [(16)], proposed a new ML-systems independently identified and extracted computational pseudo-codes and technical phrases. Research [(17)] developed a Deep Neural Network (DNN) model for SA using LSTM, CNN, and grid search. We tested CNN, K-nearest neighbor, LSTM, and neural networks. Study [(18)] proposed Capsule Networks Matrix factorization (CapsMF). A recommendation system Capsule Networks stacked atop “bidirectional Recurrent Neural Networks (Bi-RNN)” for robust representations of product and user texts enhanced a DNN text assessment model. Paper [19] presented a supervised topic-level sentiment model for broad SA tasks.

This paper's aim is to enhancing SA by creating a novel dynamic dragonfly optimization-tuned support vector machine (DDO-SVM). To improve the precision or efficiency of SA, the proposed method integrates textual features and document metadata.

2. Materials and Methods

In this paper, we gathered the sentiment dataset and employed data preprocessing techniques. The proposed Dynamic Dragonfly Optimization–tuned Support Vector Machine (DDO-SVM) presents a unique and potentially more effective approach for higher-quality Sentiment Analysis. Figure 1 depicts the proposed method architecture.

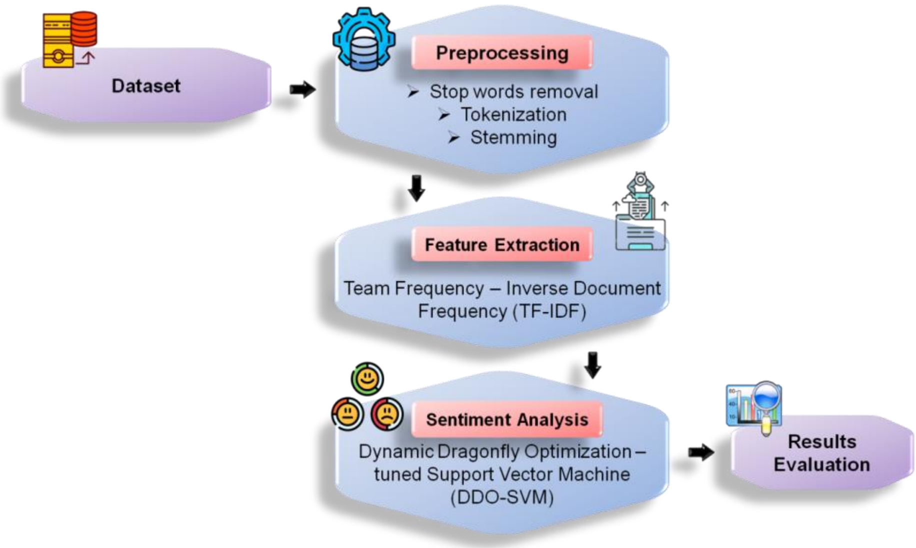

Figure 1.Proposed method architecture (Source: Author)

Dataset

This dataset comprising 351 174 English language records, the source frequencies are in Table 1. This dataset shows uneven source representation. About 48,06 % are Twitter tweets, 27,7 % are news, and 24,24 % are Facebook (FB) posts. These differences in source frequency are due to the system's overall compliance with privacy laws. Thus, data gathering is limited to open accounts or Online Social Networks (OSNs) where information is exchanged. Twitter has a sample surplus. Daily, all English newspaper websites, digital newspapers, and related sources are cataloged and examined. Table 2 presents sentiment frequencies across data annotated by five participants with reduced bias through randomization.

Table 1. Globally SDG research bibliometric performance metric (Source: Author)

Origin	Records	Proportion
Tweets	168 774	48,06 %
Articles of news	97 275	27,7 %
FB shares	49 832	14,19 %
FB comments	35 293	10,05 %
Total	351 174	100 %

Table 2. Sentiments dataset (Source: Author)

Sentiment Class	Frequency
Favorable	2,00 %
Impartial	79,75 %
Unfavorable	18,25 %

Data pre-processing

In this Study, three essential preprocessing steps such as, tokenization, stop word removal, and stemming. Tokenization converts words into usable terms. Any IR technique needs tokenization. Just sentences, figures, signs, and other characters are removed from a document. Stop words are eliminated during IR. Positive or negative, these phrases have no value in SA. They must be removed from the data set. The process of stemming involves identifying several word morphological variations and reducing them down to a single root.

Feature extraction

In this section, we utilize TF-IDF method, which is used to determine word significance by evaluating the frequency of documents.

Term frequency-inverse document frequency (TF-IDF)

TF-IDF approach, a widely used text mining method, to assign weight values to words in documents. The TF-IDF calculation involved counting word appearances and computing the Inverse Document Frequency (IDF), which is the inverse of Document Frequency (DF). DF reflects the visual importance of each word in a document. Equation (1) defines the IDF rate.

$$\boldsymbol{IDF_S = log\,\frac{N}{DF_s}} \tag{1}$$

Here, N is the total amount of records in the file, and TFIDF is calculated using Equation (2).

$$\boldsymbol{TFIDF_S = TF_S \times IDF_S} \tag{2}$$

Sentiment Analysis

In this section, we integrate DDO with SVM, enhancing SA by integrating textual features and document metadata. DDO improves SA by optimizing SVM parameters, enhancing accuracy and efficiency in sentiment classification tasks.

Dynamic Dragonfly Optimization (DDO)

The unique and improved swarm behavior of dragonflies served as the model for the DA.

DA Swarm behavior consists of tracking as well as movement. Let us assume the existence of M Draconian. Equation (3) provides the "Non-Disclosure Agreement's (NDA)" placement.

$$\boldsymbol{z_m = \left(z_m^1, z_m^e, \dots, z_m^O\right)} \tag{3}$$

Whereas the numerals "$\boldsymbol{k}$= $\boldsymbol{1, 2, 3, \dots, O}$and $\boldsymbol{M}$" represent the number of search brokers, the symbol "$\boldsymbol{w_j^c}$"specifies where the "$\boldsymbol{m^{th}}$" DA is located among the "$\boldsymbol{e^{th}}$" searchable dimension.

Utilizing the primary position evaluations, the fitness factor randomly generated and estimated among the lower and upper factor limits. There are differences in the starting values of the variables "V (separation weight), e (cohesion), D (alignment), G (food), and H (opponent factors)". Dragonfly location and velocity updates are determined using factors derived from Equations (4) to (6).

$$\boldsymbol{V_m = -\sum_{k=1}^{O} z - z_m} \tag{4}$$

$$\boldsymbol{D_m = \frac{\sum_{k=1}^{O} V_i}{O}} \tag{5}$$

$$\boldsymbol{C_m = \frac{k=1}{O} - z} \tag{6}$$

The location and speed of the $\boldsymbol{m^{th}}$ individual are represented by the variables $\boldsymbol{V_m and\ Z_m}$, respectively. Whereas $\boldsymbol{V_m}$shows the person's current location, $\boldsymbol{N}$shows the members of the group that are nearby. Equations (7) and (8) can be utilized to calculate $\boldsymbol{G_m}$, which indicates a pull towards food, and $\boldsymbol{F_j}$, which indicates Distraction from Opponents.

$$\boldsymbol{H_m = Z^+ - Z} \tag{7}$$

$$\boldsymbol{G_m = Z^- + Z} \tag{8}$$

$\boldsymbol{W}$ Denotes the person's present position, "$\boldsymbol{Z^-}$denotes the enemy source",and "$\boldsymbol{Z^+}$ denotes the food source", the Euclidean distances between the "$\boldsymbol{N}$ dragonflies" to estimate separation from one another,the distance is given by equation (9) and denoted by "$\boldsymbol{r_{nm}}$".

$$\boldsymbol{r_{nm} = \sqrt{\sum_{k=1}^{e} (z_m, k - z_m, k)^2}} \tag{9}$$

Equation (9) is the same as the PSO position calculation. We'll provide an updated address for the DA. This will be performed using Equations (10-12), which are similar to the velocity formulation for Particle Swarm Optimization (PSO).

$$\boldsymbol{\Delta Z_{u+1} = (wW_m + dD_m + aA_m + gG_m + hG_m) + y\Delta z_v} \tag{10}$$

$$\boldsymbol{Z_{u+1} = Z_u + \Delta Z_{u+1}} \tag{11}$$

The DA location will be adjusted in the area via the "Levy Flight" calculation;this is contained in Equation (12).It improves the "dragonflies' capacity for worldwide hunting". Their behavior becomes even more unpredictable and inconsistent as a result.

$$\boldsymbol{Z_{u+1} = Z_u + levy\ (c) Z_u} \tag{12}$$

The measure of fitness can be determined through making use of different positions and rates.

Support Vector Machine

This section provides a clear description of the SVM classifier. The purpose is to use the idea of margin maximization to categorize the training data into two categories given the training set$\{(\boldsymbol{w_1}, \boldsymbol{z_1}), \dots, (\boldsymbol{w_k}, \boldsymbol{z_k})\}$, where$\boldsymbol{w_j \epsilon q^m}$ ***and*** $\boldsymbol{z_j \epsilon \{-1, 1\}, j = 1, \dots, k}$. The following optimization challenge is formulated (equation (13-14)):

$$\underset{x,a,\xi}{min} \frac{1}{2} \left||x|\right|^2 + D \sum_{j=1}^{k} \xi_j \tag{13}$$

$$s.t. \quad z_j \left(\left(x \cdot \phi(w_j) \right) + a \right) \geq 1 - \xi_j, \tag{14}$$

$$\xi_j \geq 0, j = 1, \dots, k, \tag{15}$$

Sample $\boldsymbol{w_j}$ is mapped onto a high-dimensional space through the mapping function, where x and a are the separating "hyperplane's normal vector and offset," and $\boldsymbol{D > 0}$ is the "negative parameter of the error term"$\boldsymbol{\xi}; \boldsymbol{\phi(w_j)}$.

The Lagrange dual problem Equations (14-15) is translated (equations (16-18)) as follows.

$$\underset{\alpha}{min} \frac{1}{2} \sum_{j=1}^{k} \sum_{i=1}^{k} z_j z_i \alpha_j \alpha_i \left(\phi(w_j) \cdot \phi(w_i) \right) - \sum_{i=1}^{k} \alpha_i \tag{16}$$

$$s.t. \ \sum_{j=1}^{k} z_j \alpha_j = 0, \tag{17}$$

$$0 \leq \alpha_j \leq D, \ j = 1, \dots, k. \tag{18}$$

After deriving the Lagrange multipliers $\boldsymbol{\alpha_j} \in (\boldsymbol{0}, \boldsymbol{D})$ from equation (19), the classification decision function $\boldsymbol{e(w)}$can be built as follows (equation (19)):

$$e(w) = sign \ \left(\sum_{j=1}^{k} \alpha_j z_j \left(\phi(w_j) \cdot \phi(w) \right) + a \ \right) \tag{19}$$

The kernel function is typically referred as $\boldsymbol{L(w_j, w) = \left(\phi(w_j) \cdot \phi(w) \right)}$. The Gaussian kernel function is the primary focus of this work because it can simulate the majority of kernel functions using a chosen kernel parameter. It can be represented (equation (20)) in the following:

$$L(w_j, w_i) = exp \left(-\gamma \left| |w_j - w_i| \right|^2 \right), \tag{20}$$

Where the kernel parameter is $\boldsymbol{\gamma}$. The classification accuracy is heavily impacted by the ideal SVM configurations.

Dynamic Dragonfly Optimization –tuned Support Vector Machine (DDO-SVM)

The approach integrates Dragonfly Optimization (DDO) and Support Vector Machines (SVM) to improve SA by incorporating textual features. The dynamic nature of DDO adapts to changing data patterns, optimizing SVM parameters. Algorithm 1 shows the pseudo code for DDO-SVM.

Algorithm 1: Pseudocode for Dynamic Dragonfly Optimization –tuned Support Vector Machine

Function DDO_SVM_Optimization ():

Initialize Dragonflies ()

Repeat until convergence or max iterations:

Optimize With DDO ()

Train SVM with BestSolution ()

End repeat

Return Best SVM Model

Function Optimize With DDO ():

Evaluate Fitness ()

Prey Search ()

Update Positions ()

functionTrain SVM With Best Solution ():

Svm_parameters = Extract Parameters ()

BestSVMModel = TrainSVM (svm_parameters)

3. Results

The recommended task is executed in Jupiter using the Pandas Library, requiring installation with Python. The evaluation of the proposed approach in terms of Precision, Recall, and F1-Score. Comparative analysis with other existing methods, including Random Forest [11] and DNN + CNN [11].

Precision is a fundamental parameter utilized in the fields of statistics and machine learning to evaluate performance. The metric evaluates the precision of positive sentiment predictions generated by a computational model. The equation of precision as follows:

$$Precision = \frac{TP}{(TP+FP)} \tag{21}$$

Table 3 and Figure 2 show the precision result. While comparing the proposed method (DDO-SVM – 90,87 %) with the other existing method (Random Forest – 85,6 %, DNN + CNN -86,5 %), it shows that our proposed method is superior to other methods in SA.

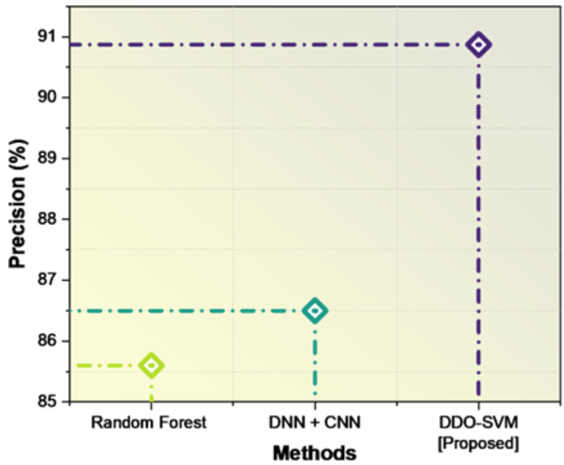

Figure 2.Precision (Source: Author)

Table 3. Precision (Source: Author)

Methods	Precision (%)
Random Forest	85,6
DNN + CNN	86,5
DDO-SVM [Proposed]	90,87

Recall is a performance metric used in data categorization and machine learning that represents the proportion of actual positive sentiments that a model properly detects. The Equation of recall as follows:

$$Recall = \frac{TP}{(TP+FN)} \qquad (22)$$

The recall result is shown in Table 4 and Figure 3. When comparing the suggested approach (DDO-SVM – 91,8 %) to the other existing method (Random Forest – 86,5 %, DNN + CNN – 87,33 %), it shows that our suggested approach outperforms the different approaches in terms of recall in SA.

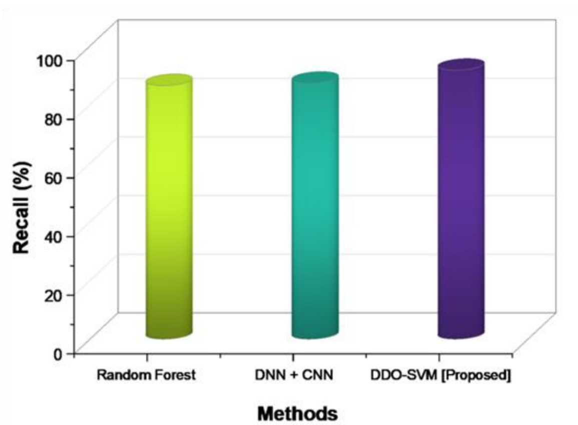

Figure 3.Recall (Source: Author)

Table 4. Recall (Source: Author)

Methods	Recall (%)
Random Forest	86,5
DNN + CNN	87,33
DDO-SVM [Proposed]	91,8

The F1-score improves a model's ability to distinguish between favorable and unfavorable attitudes. False positives as well as negatives are included in the evaluation of sentiment categorization skills. Figure 4 and Table 5 show the f1-score results. F1-score equation as follows:

$$\text{F1-score}=2\times\frac{Precision\times Recall}{Precision+Recall} \quad (23)$$

Our suggested method DDO-SVM has achieved 92,68 % When comparing the other existing methods (Random Forest – 86,5 %, DNN + CNN -87,66 %), it shows that our method outperforms the others in terms of F1-score in SA.

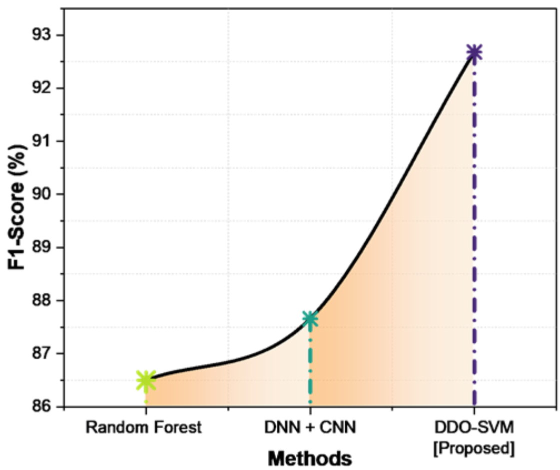

Figure 4.F1-Score (Source: Author)

Table 5. F1-Score (Source: Author)

Methods	F1-Score (%)
Random Forest	86,5
DNN + CNN	87,66
DDO-SVM [Proposed]	92,68

4. Conclusion

We introduced a novel dynamic dragonfly optimization-tuned support vector machine (DDO-SVM) method, which enhanced the SA. The proposed machine learning (ML) technique combines textual and metadata features. The gathered dataset was preprocessed, and TF-IDF was used for feature extraction. Experimental result shows Precision (90,87 %), Recall (91,8 %), and F1-Score (92,68 %). The results were compared to previous algorithms, DDO-SVM proved more effective for SA tasks. The proposed approach would have needed help growing to accommodate demands for real-time processing or massive data volumes. In future research, intelligent data sampling can develop representative subsets of the large dataset for training and testing. It can reduce computational load while preserving model performance.

References

1. Basha SM, Rajput DS. A supervised aspect level sentiment model to predict overall sentiment on tweeter documents. International Journal of Metadata, Semantics and Ontologies. 2018;13(1):33-41.https://dx.10.1504/IJMSO.2018.096451
2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of big data. 2019 Dec;6(1):1-25.https://dx.10.1186/s40537-019-0217-0
3. Gajanayake GM, Sandanayake TC. Trending pattern identification of youtube gaming channels using sentiment analysis. In2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer) 2020 Nov 4 (pp. 149-154). IEEE. https://dx.10.1109/ICTer51097.2020.9325476
4. Guo J. Deep learning approach to text analysis for human emotion detection from big data. Journal of Intelligent Systems. 2022 Jan 12;31(1):113-26.
5. Habimana O, Li Y, Li R, Gu X, Yu G. Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences. 2020 Jan;63:1-36. https://dx.10.1007/s11432-018-9941-6
6. Hasib KM, Habib MA, Towhid NA, Showrov MI. A novel deep learning based sentiment analysis of twitter data for us airline service. In2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) 2021 Feb 27 (pp. 450-455). IEEE. https://dx.10.1109/ICICT4SD50815.2021.9396879
7. Katarya R, Arora Y. Capsmf: a novel product recommender system using deep learning based text analysis model. Multimedia Tools and Applications. 2020 Dec;79(47-48):35927-48.https://dx.10.1007/s11042-020-09199-5
8. Korovesis K, Alexandridis G, Caridakis G, Polydoras P, Tsantilas P. Leveraging aspect-based sentiment prediction with textual features and document metadata. In11th Hellenic Conference on Artificial Intelligence 2020 Sep 2 (pp. 168-174). https://dx.10.1142/S0129065721500465
9. Krallinger M, Rabal O, Lourenco A, Oyarzabal J, Valencia A. Information retrieval and text mining technologies for chemistry. Chemical reviews. 2017 Jun 28;117(12):7673-761. https://dx.10.1021/acs.chemrev.6b00851
10. Li L, Goh TT, Jin D. How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. Neural Computing and Applications. 2020 May;32:4387-415. https://dx.10.1007/s00521-018-3865-7
11. Pathak AR, Pandey M, Rautaray S. Topic-level sentiment analysis of social media data using deep learning. Applied Soft Computing. 2021 Sep 1;108:107440. https://dx.10.1016/j.asoc.2021.107440
12. Priyadarshini I, Cotton C. A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis. The Journal of Supercomputing. 2021 Dec;77(12):13911-32. https://dx.10.1007/s11227-021-03838-w
13. Rintyarna BS, Sarno R, Fatichah C. Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context. International Journal of Information and Decision Sciences. 2020;12(1):75-101. https://dx.10.1504/IJIDS.2020.104992

14. Safder I, Hassan SU, Visvizi A, Noraset T, Nawaz R, Tuarob S. Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Information processing & management. 2020 Nov 1;57(6):102269. https://dx.10.1016/j.ipm.2020.102269
15. Shahare FF. Sentiment analysis for the news data based on the social media. In2017 International Conference on Intelligent Computing and Control Systems (ICICCS) 2017 Jun 15 (pp. 1365-1370). IEEE. https://dx.10.1109/ICCONS.2017.8250692
16. Xu J, Huang F, Zhang X, Wang S, Li C, Li Z, et al. Sentiment analysis of social images via hierarchical deep fusion of content and links. Applied Soft Computing. 2019 Jul 1;80:387-99. https://dx.10.1016/j.asoc.2019.04.010
17. Yang X, Feng S, Wang D, Zhang Y. Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia. 2020 Nov 2;23:4014-26. https://dx.10.1109/TMM.2020.3035277
18. Zeng ML, Hong Y, Clunis J, He S, Coladangelo LP. Implications of knowledge organization systems for health information exchange and communication during the COVID-19 pandemic. Data and Information Management. 2020 Sep 1;4(3):148-70. https://dx.10.2478/dim-2020-0009
19. Zhao H, Liu Z, Yao X, Yang Q. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. Information Processing & Management. 2021 Sep 1;58(5):102656.https://dx.10.1016/j.ipm.2021.102656