

Sentiment classification using gradient modulation and layered attention

Bagiyalakshmi Natarajan, T. Veeramakali

Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology,
Kattankulathur, India

Article Info

Article history:

Received Apr 29, 2025

Revised Aug 1, 2025

Accepted Sep 7, 2025

Keywords:

Attention sentiment optimization

Graded multi-head attention

Hierarchical layer analysis

Natural language processing

Selective gradient adjustment

Sentiment analysis

ABSTRACT

Sentiment analysis is a technique for evaluating text to ascertain whether a statement is positive, negative, or neutral. Currently, transformer-based models capture the contextual relationships among words in a phrase and accomplish sentiment analysis in a nuanced manner via multi-head attention. This approach, with a fixed number of layers and heads, struggles to find the complex relationships between phrases and their semantic structures. To mitigate this issue, the suggested technique incorporates the graded multi-head attention model (GMHA) at the base of the distilled bidirectional encoder representations from transformers (DistilBERT) model. It is employed to augment the layers and heads progressively, capturing the relationships between sentences in a sophisticated manner. By increasing the layers and heads the proposed model extracts long-term and hierarchical relationships from the sentence. Additionally, the attention sentiment optimization technique is introduced, which improves model learning by giving more weight to important words in a sentence. During training, the process checks to see which words ("amazing" or "worst") get more attention and gives them more weight in the model update. This makes it easier for the model to understand important emotions. Our suggested model enhances performance in sentiment exploration, with an accuracy of 96.53%. This interpretation includes a comparison analysis with another contemporary framework.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Bagiyalakshmi Natarajan

Department of Data Science and Business Systems, School of Computing

SRM Institute of Science and Technology

Kattankulathur, India

Email: bn7569@srmist.edu.in

1. INTRODUCTION

Sentiment analysis, commonly referred to as opinion mining, is utilized in natural language processing (NLP). Sentiment analysis is a task used to processing textual data, including product reviews, consumer comments, social media material, and news [1]. The sentiment can be categorized into three classifications [2]. Positive content words suggest a good attitude or contentment; negative emotion phrases signify disappointment, critique, or adverse perspectives; and neutral sentiment the text conveys no specific emotions or lacks clarity [3]. Aspect-based sentiment analysis (ABSA) is a technique in the field of NLP designed to discern the sentiment directed at particular elements or attributes of a product, service, or subject within a specified sentence [4].

Initially, rule-based systems employed lexicons for the analysis of sentiment in text. The rule-based methodology exhibits constraints such as substantial development effort, limited flexibility, and challenges in

processing complex sentences [5]. As a result, machine learning emerged. Diverse techniques, such as support vector machines and naive Bayes, are employed for sentiment classification [6]. However, it necessitates sufficient training data and is challenging to locate context-specific data [7].

Deep learning appeared significantly to mitigate contextual dependency. Numerous deep learning techniques utilized in sentiment analysis comprise recurrent neural network (RNN), long short-term memory (LSTM), and bidirectional encoder representations from transformers (BERT) models [8]. RNNs and LSTMs encountered difficulties with long-range word dependencies, leading to the development of attention models, which proficiently discern crucial words in sentences and encapsulate extended dependencies in text [9]. Multi-head attention requires significant computational resources and several hyperparameters, which are difficult to optimize [10]. The conventional multi-head attention distributes uniform attention among all heads and layers, leading to heightened memory usage and computational requirements [11].

To overcome this limitation, our proposed model incorporates graded multi-head attention (GMHA), which gradually rises the number of focused heads throughout layers. This enhancement allows the model to effectively capture both semantic and syntactic patterns at varying levels of granularity. The model uses the contextual relevance filter (CRF) were employed to eliminate superfluous and relevant information before classification. This allows the model to focus on sentiment relevant attributes while discarding extraneous data. Additionally, this study introduces the novel approach is attention-sentient gradient optimization (ASGO), which gives more important to the sentiment rich terms during the training process. It modifies gradients of sentiment rich words based on attention scores instead of treating all words equally. This work enhances sentiment classification systems by incorporating various attention layers, filtering techniques, and gradient optimization approaches, leading to improved robustness and interpretability.

2. METHOD

2.1. Data pre-processing

In NLP, text preprocessing is a vital process that includes cleaning and transforming the data into a machine-readable format [12]. Among the many tasks involved are stop word elimination, tokenization, lemmatization, and stemming. These steps reduce the noise in the data, thereby rendering it easier to navigate and more useful for analysis [13]. The next step involves transforming the text into token embedding using the DistilBERT embedding technique, which establishes the context-dependent connection between the words in the sentence [14].

2.2. DistilBERT model

DistilBERT is a pretrained model appropriate for recent applications, especially sentiment analysis. It achieves a good balance between efficiency and accuracy, making it useful for real-time tasks and situations where computational resources are inadequate [15]. Despite its small size, it retains a high level of accuracy in emotion classification. The DistilBERT model is a more compact interpretation of the BERT model, being 40% smaller and 60% faster [16]. Although BERT has 12 transformer encoders, DistilBERT employs only 6 encoders [17], however, it achieves performance comparable to BERT in a nuanced fashion, as demonstrated in the Figure 1. This study examines the text utilizing the 6-transformer model with 12 attention heads, commencing with the translation of sentences into tokens by DistilBERT embedding, and subsequently transforming each token into vectors [18]. This feature vector navigates the six transformer layers, each consisting of multi-head self-attention and a feedforward network [19]. Each self-attention layer calculates the attention score for every token, so capturing the inter-relationships among tokens while concurrently emphasizing different facts of the text [20]. Each token in the sentence is represented as w_1, w_2, \dots, w_n , and word embedding can be expressed by (1). Let X be the input sentence with the n words.

$$X = \{w_1, w_2, w_3 \dots w_n\} \quad (1)$$

Following tokenization and word embedding, we acquire the input representation matrix E . E is the embedding vector with n rows and d columns, with d indicating the embedding dimension of the DistilBERT model, which is 768 as in (2) [21].

$$E = [e_1, e_2, \dots, e_n] \in R^{n \times d} \quad (2)$$

The attention score for each token is computed using the (3).

$$A_d = \text{Softmax} \left(\frac{qk^T}{\sqrt{d_k}} \right) \quad (3)$$

In this instance, A_d signifies the attention score for each token, while Q , K , and V denote the input words. The hidden state value of DistilBERT, referred to as H_d , is calculated using the (4) and relayed to the GMHA layer [22].

$$H_d = A_d V \quad (4)$$

2.3. Proposed graded multi-head attention

The transformer-based model faces challenges with complex and long-range dependencies and existing model possesses a predetermined quantity of layers and heads [23]. In this context, each attention layer treats all tokens equally, lacking a focus on varying levels of abstraction. therefore, this study incorporated GMHA following the final hidden layer of DistilBERT. In the GHMA method, the layers are progressively increased, effectively capturing the semantic information in the text with nuance. The initial layer consists of two heads, designed to learn the syntactic structure and basic word dependencies within the sentence. In the second layer, the number of heads increases by four to emphasize phrases and nearby relationships. The third layer further expands the heads by eight to capture contextual dependencies among the words. Finally, the fourth layer increases the heads by twelve to address long-range dependencies, with a strong focus on sentiment-rich words.

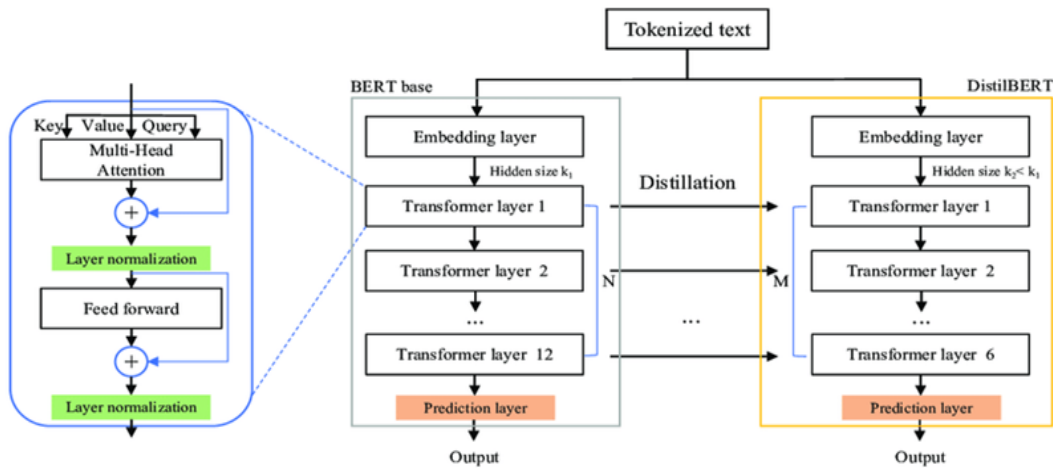


Figure 1. Architecture of transformer model [24]

The DistilBERT model is designed for wide application; however, its graded mechanism is specifically optimized for sentiment analysis, effectively capturing language nuances [25]. The proposed architecture represented in Figure 2. In the proposed architecture, the pre-processed review is embedded using DistilBERT embedding, which is subsequently sent to the DistilBERT model. This model comprises a fixed number of layers and heads, specifically six layers and twelve heads, allowing the review to be processed through all these layers and the attention score for each token to be calculated. These layers capture the contextual relationships and dependencies among the words and generate hidden information, which is then transmitted to the GMHA module. By applying multiple attention heads hierarchically across layers, the GHMA improves feature extraction and enables the model to capture longer-range and deeper dependencies. This strategy improves its ability for understanding complicated semantic patterns in a text. GHMA effectively improves the contextual knowledge, which is subsequently utilized for additional analysis such as attention sentiment optimization technique. The attention score for each token is calculated by acquiring the hidden information H_d from the base model and multiplying it by the learnable weight parameter W^h , according to (5) and (6). Were, $Q_h = H_d W_Q^h$, $K_h = H_d W_K^h$, $V_h = H_d W_V^h$.

$$A_h = \text{Softmax} \left(\frac{Q_h K_h^T}{\sqrt{d_k}} \right) \quad (5)$$

The updated concealed representation result of GMHA H_h can be computed utilizing the successive formula.

$$H_h = A_h V_h \quad (6)$$

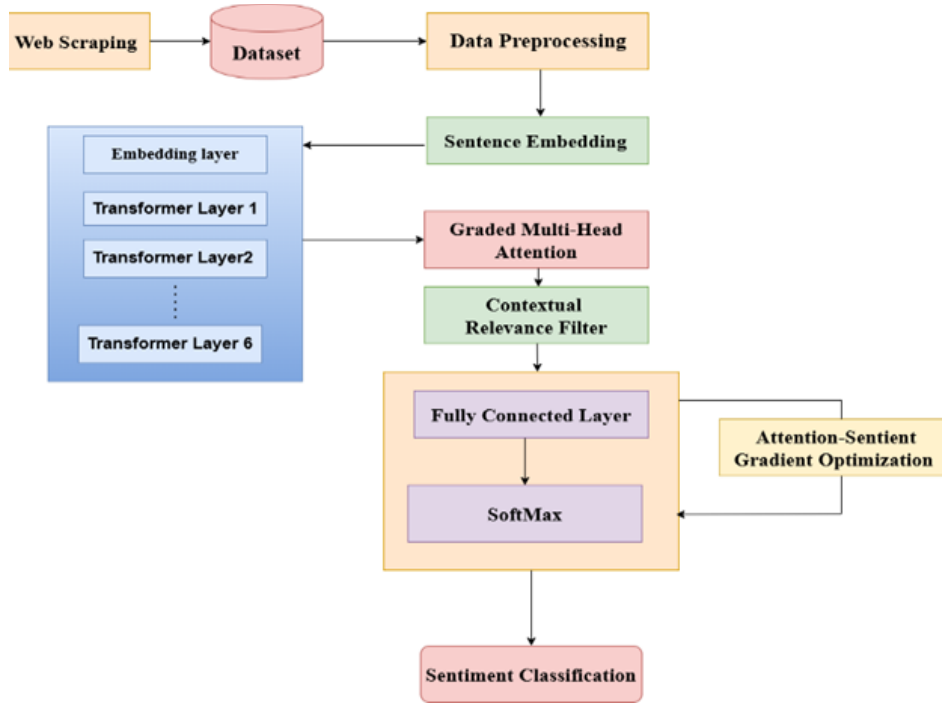


Figure 2. Proposed framework for GMHA architecture

2.4. Attention controller

The attention controller determines where to stop layer incrementation and how many layers should be added to the GHMA. The attention score for each level is computed and compared with the previous layer. The cosine similarity measure was used to compare this score with the score for the previous layer; if the score is less than or equal to the threshold value of 0.001, it indicates that the model has already learned enough about the context and is not benefiting much from the addition of more layers. The controller then stops to add layers. The threshold value is set at 0.001, which is useful for addressing long-range dependencies. This aids the model in avoiding pointless calculations and concentrating solely on pertinent data.

2.5. Contextual relevance filter

GMHA generates attention scores while also offering superfluous information and deceptive patterns. The contextual filter methodically eliminates noise and irrelevant information before sending data to the classifier. The formula for obtaining the hidden information is shown in (7).

$$G_f = \sigma(W_g H_h) \quad (7)$$

The W_g denotes the trainable weight matrix, while H_g signifies hidden information derived from GMHA. The G_f provides the assessed weight for each feature. σ denotes the sigmoid activation function. The filtered output is then formulated in (8).

$$H_g = G_f \odot H_h \quad (8)$$

The aforementioned formula denotes the cumulative filtered features transmitted to the subsequent level for categorization. \odot denotes element-wise multiplication.

2.6. Attention sentient gradient optimization

This optimization strategy increases sentiment analysis quality by updating the weights of the most essential terms in the sentence while training. Traditional optimizers, such as Adam, update the weights of words during training, but they treat all tokens equally. Our suggested method finds the most essential word in the text using the attention score from DistilBERT and GHMA. For example, "The product is amazing but the battery is not good." In this statement, the words "amazing" and "not good" will receive more attention and weight than the other words. This additional information is provided to the traditional optimizer, which

gives higher weight to key words during backpropagation. By focusing on the appropriate words, the model learns more effectively and improves its ability to understand feelings. The model uses the standard gradient with the Adam optimizer, employing the (9).

$$\nabla L = \frac{\partial L}{\partial w} \quad (9)$$

Next the ASGO reweight the gradient using the (10), then the amend gradient $\nabla L'$ used to rejuvenate the model parameter using optimizer like Adam.

$$\nabla L' = \nabla L \odot (A_d + A_h) \quad (10)$$

3. RESULTS AND DISCUSSION

The proposed work includes collecting review comments from e-commerce platforms through application programming interfaces (APIs) in the categories of mobile phones and accessories, as well as clothing and accessories, and performing data annotation. We gathered more English language comments from online sources, including positive, negative, and neutral sentiments, as depicted in Figure 3. The yielded dataset details represented in the Table 1, it shows our dataset is reasonably balanced and suitable for sentiment analysis.

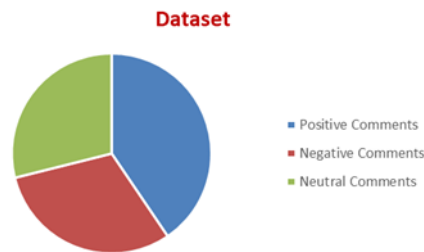


Figure 3. Review comments dataset

Table 1. Dataset distribution

Class	Number of comments
Positive	36965
Negative	27929
Neutral	26376

The model was trained and deployed with this dataset to conduct sentiment analysis on the review comments. The model utilizes the Adam optimizer to identify the parameters necessitating updates. The principal study illustrates that the model utilizes both GMHA and an Attention sentiment optimization method, alongside a pretrained model, hence improving the efficacy of sentiment analysis tasks. The existing pretrained model, includes only self-attention layer, it uses fixed number of nodes and layers for semantic analysis, but various processing layer essential to process composite information contained in the sentence. So, in our proposed model used progressive attention on the sentences, it increases the precision, recall and accuracy of the model. The model has a precision of 96.23%, an F1-score of 96.32%, a recall of 96.53%, and an accuracy of 96.53%. The depicted GMHA can manage the complex relationships among the texts.

We utilized our dataset with several established sentiment analysis models, sentiment analysis-attention gated recurrent unit (GRU), sentiment analysis-attention BERT. Sentiment analysis-attention robustly optimized BERT pretraining approach (RoBERTa) and compared the results with our suggested model. Table 2 illustrates the comparison of precision across different sentiment analysis models.

Table 2. Performance evaluation of sentiment analysis models

Model	Precision	Accuracy
SA-attention GRU	82.6	83.3
SA-attention BERT	86.3	86.6
SA-attention RoBERTa	90.2	90.4
GMHA	96.23	96.53

This proposed research investigates the effects of consecutive layers and parallel attention. The proposed model exhibits superior results compared to an alternative model, with the precision comparison illustrated in Figure 4. The current approaches employing attention mechanisms are assessed, as the innovative attention mechanism is effortlessly included into the existing model relevant to our suggested framework. The current model primarily enhances the optimization of neural networks. The existing attention system encodes phrases and evaluates the sentiment strength of words; nevertheless, its efficacy is limited. Our suggested model utilizes the attention mechanism alongside an optimization method to improve weight adjustment during backpropagation, hence allocating greater weight to significant words inside the phrase. Figure 5 illustrates the accuracy association between our suggested model and the established model.

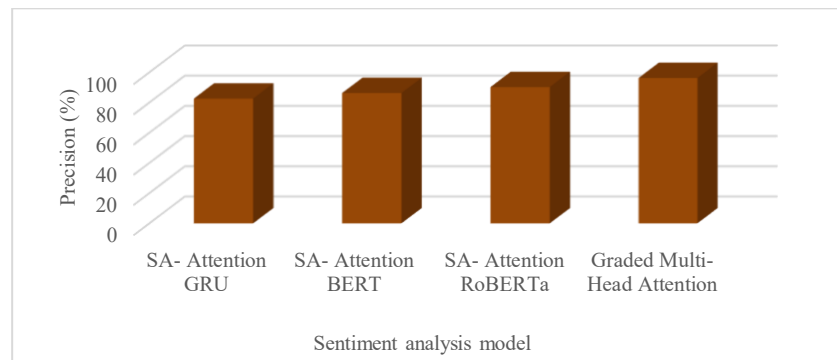


Figure 4. Precision comparison

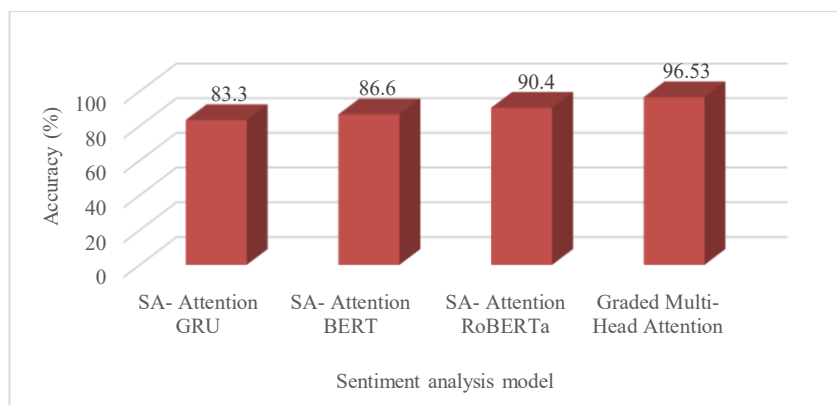


Figure 5. Accuracy comparison

4. CONCLUSION

The GMHA approach accomplishes sentiment analysis on review comments with sophistication. This novel method utilizes a GMHA mechanism and an optimization strategy to capture the contextual connections among words in a phrase. It systematically describes the relationship, incorporating syntactic and semantic attributes from the text through the gradual enhancement of attention layers. This approach enhances the model's efficacy and exhibits strong adherence to recognized NLP paradigms. It gathers more precise data to enhance the model's efficacy. The proposed methodology was evaluated using a real dataset, producing effective results employing our optimization techniques. We employed the same real dataset for the present models and performed analyses; based on the outcomes of this experiment, we compared them with our proposed model. The proposed work employs attention sentiment optimization strategies that prioritize sentiment intensity words during hyperparameter tuning, so highlighting the more significant components of the phrase in various aspects of review comments and producing more accurate findings in a nuanced manner. Future endeavors will include integrating additional linguistic external knowledge into the model and endeavoring to implement the model using a multidomain dataset.

FUNDING INFORMATION

The authors confirm that no funding, money, or other assistance was obtained in order to prepare this work.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Bagiyalakshmi Natarajan	✓	✓	✓	✓	✓	✓		✓	✓	✓				✓
T. Veeramakali		✓				✓		✓	✓	✓	✓	✓		

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: a systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.

[2] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: a review of sentiment analysis based on sequential transfer learning," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 749–780, Apr. 2022, doi: 10.1007/s10462-022-10183-8.

[3] A. Daza, N. D. G. Rueda, M. S. A. Sánchez, W. F. R. Espíritu, and M. E. C. Quiñones, "Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: a bibliometric analysis, systematic literature review, challenges and future works," *International Journal of Information Management Data Insights*, vol. 4, no. 2, Nov. 2024, doi: 10.1016/j.jjimei.2024.100267.

[4] M. E. Mowlaei, M. S. Abadeh, and H. Keshavarz, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," *Expert Systems with Applications*, vol. 148, Jun. 2020, doi: 10.1016/j.eswa.2020.113234.

[5] N. Saraswathi, T. S. Rooba, and S. Chakaravarthi, "Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews," *Measurement: Sensors*, vol. 29, Oct. 2023, doi: 10.1016/j.measen.2023.100888.

[6] M. Z. Mekonen *et al.*, "An opinionated sentiment analysis using a rule-based method," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 726–732, Feb. 2025, doi: 10.11591/eei.v14i1.8568.

[7] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/access.2022.3152828.

[8] C. Gupta, G. Chawla, K. Rawlley, K. Bisht, and M. Sharma, "Senti_ALSTM: Sentiment analysis of movie reviews using attention-based LSTM," in *Proceedings of 3rd International Conference on Computing Informatics and Networks*, 2021, pp. 211–219, doi: 10.1007/978-981-15-9712-1_18.

[9] O. Ndama, I. Bensassi, and E. M. En-Naimi, "The impact of BERT-infused deep learning models on sentiment analysis accuracy in financial news," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1231–1240, Apr. 2025, doi: 10.11591/eei.v14i2.8469.

[10] G. Zhao, Y. Luo, Q. Chen, and X. Qian, "Aspect-based sentiment analysis via multitask learning for online reviews," *Knowledge-Based Systems*, vol. 264, Mar. 2023, doi: 10.1016/j.knosys.2023.110326.

[11] Y. Huang, X. Bai, Q. Liu, H. Peng, Q. Yang, and J. Wang, "Sentence-level sentiment classification based on multi-attention bidirectional gated spiking neural P systems," *Applied Soft Computing*, vol. 152, Feb. 2024, doi: 10.1016/j.asoc.2024.111231.

[12] C. Gan, X. Fu, Q. Feng, Q. Zhu, Y. Cao, and Y. Zhu, "A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis," *Expert Systems with Applications*, vol. 242, May 2024, doi: 10.1016/j.eswa.2023.122731.

[13] R. K. Dey and A. K. Das, "Neighbour adjusted dispersive flies optimization based deep hybrid sentiment analysis framework," *Multimedia Tools and Applications*, vol. 83, no. 24, pp. 64393–64416, Jan. 2024, doi: 10.1007/s11042-023-17953-8.




[14] F. Wang, Q. Bao, Z. Wang, and Y. Chen, "Optimizing transformer based on high-performance optimizer for predicting employment sentiment in American social media content," in *2024 5th International Conference on Machine Learning and Computer Application*, Oct. 2024, pp. 414–418, doi: 10.1109/icmlca63499.2024.10753783.

[15] J. Zimmermann, L. E. Champagne, J. M. Dickens, and B. T. Hazen, "Approaches to improve preprocessing for latent Dirichlet allocation topic modeling," *Decision Support Systems*, vol. 185, Oct. 2024, doi: 10.1016/j.dss.2024.114310.




- [16] A. Wendland, M. Zenere, and J. Niemann, "Introduction to text classification: impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique," in *Systems, Software and Services Process Improvement*, 2021, pp. 289–300, doi: 10.1007/978-3-030-85521-5_19.
- [17] I. A. Kandhro, F. Ali, M. Uddin, A. Kehar, and S. Manickam, "Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement," *Knowledge and Information Systems*, vol. 66, no. 7, pp. 3639–3669, Apr. 2024, doi: 10.1007/s10115-024-02104-8.
- [18] M. Jojoa, P. Eftekhari, B. N.-Kia, and B. G.-Zapirain, "Natural language processing analysis applied to COVID-19 open-text opinions using a DistilBERT model for sentiment categorization," *AI & SOCIETY*, vol. 39, no. 3, pp. 883–890, Nov. 2022, doi: 10.1007/s00146-022-01594-w.
- [19] A. Areshey and H. Mathkour, "Exploring transformer models for sentiment classification: a comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet," *Expert Systems*, vol. 41, no. 11, Aug. 2024, doi: 10.1111/exsy.13701.
- [20] B. Pattanayak, A. Majumder, B. Jothi, and K. S., "Text summarization with DistilBERT-LSTM," in *2025 International Conference on Intelligent Systems and Computational Networks*, Jan. 2025, pp. 1–8, doi: 10.1109/iciscn64258.2025.10934207.
- [21] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *Journal of Big Data*, vol. 10, no. 1, Jun. 2023, doi: 10.1186/s40537-023-00781-w.
- [22] V. Vajrobal, N. Aggarwal, U. Shukla, G. J. Saxena, S. Singh, and A. Pundir, "Explainable cross-lingual depression identification based on multi-head attention networks in Thai context," *International Journal of Information Technology*, vol. 17, no. 5, pp. 2997–3012, Oct. 2023, doi: 10.1007/s41870-023-01512-3.
- [23] A. Aljofey, S. A. Bello, J. Lu, and C. Xu, "BERT-PhishFinder: a robust model for accurate phishing URL detection with optimized DistilBERT," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 4, pp. 4315–4329, Jul. 2025, doi: 10.1109/tdsc.2025.3545771.
- [24] H. Adel *et al.*, "Improving crisis events detection using DistilBERT with Hunger Games search algorithm," *Mathematics*, vol. 10, no. 3, Jan. 2022, doi: 10.3390/math10030447.
- [25] A. R. Nair, R. P. Singh, D. Gupta, and P. Kumar, "Evaluating the impact of text data augmentation on text classification tasks using DistilBERT," *Procedia Computer Science*, vol. 235, pp. 102–111, 2024, doi: 10.1016/j.procs.2024.04.013.

BIOGRAPHIES OF AUTHORS



Bagiyalakshmi Natarajan    received her Bachelor's degree in Information Technology from Adhiparasakthi Engineering College. She subsequently obtained her master's degree in Information Technology. She served as an Assistant Professor in the Department of Computer Science and Engineering. She is currently pursuing full-time research at the SRM Institute of Science and Technology. Her research interests include natural language processing, text summarization, and deep learning. She contacted at email: bn7569@srmist.edu.in.



T. Veeramakali    working as an Associate Professor in the Department of Data Science and Business Systems, School of Computing at SRM Institute of Science and Technology. She graduated in Information Technology in 2003 at Sri Siva Subramaniya Nadar (SSN) College of Engineering, Chennai, Tamilnadu, India. She secured Master of Technology in Information Technology in 2007 at Sathyabama University, Chennai, India. She completed her Ph.D. in the field of Cognitive Radio Network at Anna University in 2018, Chennai, India. She is in teaching profession for more than 20 years. She has published many papers in SCI/Scopus indexed journals and presented number of papers in national and international conference. Her main area of interest includes networks, machine learning, image processing, and internet of things. She is a life time member of the professional bodies such as CSE, ISTE, and IETE. She contacted at email: veeramat@srmist.edu.in.