IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Efficiently Learning an Encoder that Classifies Token Replacements and Masked Permuted Network-Based BIGRU Attention Classifier for Enhancing Sentiment Classification of Scientific Text

MUHAMMAD INAAM UL HAQ[1], KHALID MAHMOOD[2], (Senior Member, IEEE), QIANMU LI[1], (Senior Member, IEEE), ASHOK KUMAR DAS[3,4], (Senior Member, IEEE), SACHIN SHETTY[5,6], (Senior Member, IEEE) AND MAJID HUSSAIN[7], (Senior Member, IEEE)

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]Graduate School of Intelligent Data Science, National Yunlin University of Science and Technology, Douliu 64002, Taiwan
[3]Center for Security, Theory and Algorithmic Research, International Institute of Information Technology at Hyderabad, Hyderabad 500032, India
[4]Department of Computer Science and Engineering, College of Informatics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, South Korea
[5]Department of Modeling, Simulation and Visualization Engineering, Virginia Modeling, Analysis and Simulation Center, Old Dominion University, Suffolk, VA23435, USA
[6]Center for Cybersecurity Education and Research, Old Dominion University, Suffolk, VA 23435, USA
[7]Department of Computer Science, The University of Faisalabad, Faisalabad, Pakistan

Corresponding authors: Ashok Kumar Das (e-mail: iitkgp.akdas@gmail.com, ashok.das@iiit.ac.in); Qianmu Li (e-mail: liqianmu374@gmail.com, qianmu@njust.edu.cn)

ABSTRACT The exponential growth of scientific literature in digital repositories poses challenges in interpreting complex attitudes within academic texts. Traditional sentiment analysis methods often struggle with nuanced word meanings due to contextual variations. To address this, we propose the Electra-MPNet-based BiGRU attention classifier that extracts the high-level semantic features from citation sentences using the combined strength of Electra and MPNet encoder layers. These features are then combined to extract long-range dependencies through a stacked BiGRU layer. A linear attention mechanism is imposed to estimate the attention weights and context vector which enables the model to selectively focus on relevant information. The proposed model overcomes inherent constraints and adeptly manages contextual information, enhancing the model's understanding of sequential data and improving predictive accuracy. We evaluate the proposed model on a dataset of 8736 citation sentences extracted from scientific articles spanning multiple domains. Our model outperforms state-of-the-art models like LSTM-GRU, Bert-BiLSTM, Bert-LSTM-CNN and MPNet, Electra, BiGRU, and machine learning models in terms of accuracy, precision, recall, F1 and kappa measure which further solidifies its superiority in scientific text sentiment analysis tasks.

INDEX TERMS citation sentences, sentiment analysis, attention mechanism

I. INTRODUCTION

This investigation of academic scholars' emotions in scientific citation texts is an interesting and relevant topic in the current digital era, which is defined by the enormous amount of textual material that floods the World Wide Web [1]. This abundance of data contains a gold mine of insightful information that is just waiting to be uncovered by careful examination catered to need. Sentiment analysis, which is sometimes confused with opinion mining, aims to identify and classify the wide range of emotions—whether favorable, negative, or neutral—expressed in textual content [2]. Extracting thoughts from citation text in research articles is especially important since it strengthens arguments and creates intellectual connections. Over the past decade,

numerous scholars have underscored the significance of sentiment analysis, or opinion mining, thereby establishing it as a pivotal area of research within citation analysis. This growing interest highlights the necessity of understanding sentiment in citation sentences, as it provides insights into the influence and reception of scientific work. The analysis of sentiments expressed in citation sentences helps discern the nuanced opinions researchers hold about particular studies, contributing to a deeper comprehension of academic discourse and its impact. This increasing recognition reflects the broader trends in the academic community toward leveraging advanced computational techniques to extract meaningful patterns from scholarly texts, reinforcing the crucial role of sentiment analysis in the evolving landscape of scientific citation text analysis [3].

Recognizing the distinct linguistic features of the scientific domain, we explicitly aim to solve the problem of sentiment analysis in citation sentences from various research articles. There is a gap in our knowledge of sentiment in scientific writing because previous research has primarily concentrated on sentiment analysis in genres such as English and Chinese [4]. The existing approaches for scientific citation text analysis use machine learning and deep learning models [1,4,5]. Among them, machine learning approaches use the Bag of Words (BoW) model, which utilizes N-grams but fails to capture the contextual meaning of words. To address this limitation, word embeddings were introduced [5], enabling the representation of words in a way that captures their semantic relationships. However, these models require extensive vocabularies and substantial computational resources, highlighting the trade-off between improved contextual understanding and the demands for computational power and data. The nuanced expression of sentiment in citation sentences, the objective writing style common in this discipline, the neutral tone, and the writers' prejudices that must be considered are challenges found in scientific texts [1].

Furthermore, sentiment identification is made more difficult by the usage of dual-mode language, where praise may be followed by criticism. This problem is not exclusive to scientific literature [6]. Word2Vec and GloVe embeddings are limited by static, non-contextual representations, vocabulary constraints, inability to capture complex semantic and syntactic relationships, poor handling of morphological variations, and lack of task-specific fine-tuning. These challenges are not effectively handled by the existing approaches [1,2,5]. These problems motivate us to propose a more advanced sentiment analysis model to solve these challenges of the existing approach in scientific text sentiment analysis. In existing work researchers also use transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) for designing hybrid sentiment analysis approaches [7,8], however, BERT does not capture the dependency between predicted tokens hence BERT based approaches are less efficient. MPNet (Permuted Language Modeling) and ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) outperform BERT by providing better contextual representations as MPNet captures dependency between predicted tokens and ELECTRA uses replaced token detection mechanism it efficiently learns from few samples. So, we propose a more advanced approach combining the power of these two models and further enhancing it by employing Bidirectional GRU (BIGRU) and attention layers [9,10,11,12].

In this study, a citation sentence indicates that there are references to other academic publications included in the text of a particular academic publication; the cited work is called the former and the citing article is called the latter. The "Harvard Style" is a well-known example of how citation references are supplied in addition to citation sentences. It involves listing the author's last name followed by the year of publication [2]. In this work, we employed two annotated corpora of citation phrases to perform sentiment analysis on scientific citations. To provide the citation sentence's polarity, these corpora were annotated using pre-established procedures and then further refined. We built a complex model architecture on top of these foundations by fusing the powerful properties of the BIGRU, Electra, MPNet, and additional linear attention layers. The main contributions of our study can be summarized as follows:

- **i)** Enhance sentiment analysis accuracy: The goal is to leverage Electra's and MPNet's combined strength to identify intricate patterns and contextual cues from textual inputs, thereby improving sentiment analysis accuracy. By integrating BIGRU layers, the model will better handle contextual data and enhance its understanding of sequential information, leading to improved forecast precision.
- **ii)** Develop the Electra-MPNet-based Scientific Sentiment analysis model architecture: The primary objective is to propose a novel model architecture that combines BIGRU layers with the Electra and MPNet encoders along with the attention mechanism. This fusion aims to overcome the limitations of current sentiment analysis approaches of scientific text sentiment analysis in capturing subtle differences in word meanings within sentences due to context.
- **iii)** Compare performance against existing methods: Evaluate the performance of the proposed model against state-of-the-art machine learning approaches and various deep learning benchmarks, including LSTM-GRU, BERT-BiLSTM, BERT-LSTM-CNN, MPNet, Electra, and BiGRU. Particularly focus on instances where positive and negative sentiments are underrepresented to demonstrate the superiority of the proposed approach.

iv) Enable insights extraction from scholarly writings: Provide researchers with a robust tool for sentiment analysis in scientific literature that seamlessly integrates modern methodologies. By utilizing the proposed approach, researchers can extract valuable insights from the expanding corpus of scholarly writings, facilitating knowledge retrieval and interpretation in academic domains.

The rest of the paper is structured as follows: Section 2 discusses the literature review and related works. Section 3 elaborates on the specific implementation process of the proposed method. Section 4 presents the experiments conducted in this study. Section 5 explains the results. Section six has a discussion and in the 7th section, we conclude this study.

II. LITERATURE REVIEW

This section reviews the existing schemes proposed in the literature.

A. TEXT CLASSIFICATION

Natural Language Processing (NLP) techniques, widely utilized for text classification [13] [14] [15] [16] [17] [18] [19] find applications across various domains like health, social science, business, marketing, and law. Researchers analyze text data such as chat messages, notes, and social media posts to uncover human activity, employing text classification at document, paragraph, sentence, and subsentence levels. The process involves feature extraction, dimension reduction, classification selection, and evaluation phases. The initial conversion of unstructured text to structured data, followed by cleaning to preserve essential characters and words, constitutes feature extraction. Optionally, dimensionality reduction may be applied to handle large, pre-processed data efficiently. Text classification employs machine learning, deep learning, or ensemble-based models, with the trained model evaluated for performance [20].

Rocchio compares document frequency vectors with a prototype but retrieves only a few documents. Boosting adjusts weights based on performance, and Bagging[43] combines sub-sample predictions; both are computationally intensive and lack interpretability. Logistic regression (LR) predicts using a decision boundary, Naïve Bayes (NB) based on word frequencies, assuming data distribution. K-Nearest Neighbors (KNN) predicts via document similarity but requires significant memory. Support Vector Machine (SVM) creates a hyperplane boundary but is time and memory-intensive. Decision trees (DT) are categorized based on node entropy but are sensitive to data perturbations. Random Forest (RF) builds multiple trees for voting but has longer training times. Conditional Random Field (CRF) is effective on text but complex and cannot handle unseen data. Semi-supervised learning leverages labeled and unlabeled data, often utilizing clustering techniques [20,46].

Deep learning surpasses traditional methods by eliminating manual feature extraction and domain knowledge dependency. Techniques like feed-forward neural networks convert text into low-dimensional feature vectors through multiple layers, with final predictions made by classifiers. Recurrent neural networks (RNNs), especially (Long Short-Term Memory) LSTM variants, capture temporal patterns, while convolutional neural networks (CNNs) excel in spatial comprehension. Capsule networks (Caps Nets) view entities as groups of neurons with distinct attributes. Attention-based models weigh the importance of vectors for target prediction, while memory-augmented networks extend this with external memory for input comprehension. Graph neural networks construct natural language structures like syntax [21,22,46].

B. SENTIMENT ANALYSIS APPROACHES

Sentiment analysis, a crucial NLP technique, aims to discern the positive, negative, or neutral attitudes expressed in textual data, such as product reviews or social media discussions [23] [24]. This analysis proves valuable in understanding public opinion on various topics and can be extended to emotion recognition within text documents, facilitating insights into individuals' reactions and responses across different contexts [25]. The classification of sentiment and emotions often relies on domain-specific knowledge and contextual understanding, employing a range of NLP techniques, including statistical, machine learning, and deep learning approaches.

Three primary methods are employed: supervised, lexicon-based, and semantic-based techniques. In supervised methods, sentiment is predicted based on labeled datasets, utilizing text classification techniques with appropriate feature engineering. Lexicon-based methods leverage pre-existing dictionaries with sentiment orientations to make predictions. Semantic-based approaches, on the other hand, rely on evaluating conceptual and contextual semantics through word co-occurrence patterns in text. External semantic knowledge, such as semantic networks and word clustering, assists in sentiment prediction by capturing semantic relationships. Semantic networks represent sentiment-conveying words, while WordNet exploits ontological structures [23].

C. LARGE LANGUAGE MODELS IN SENTIMENT ANALYSIS

Sentiment analysis, a crucial aspect of natural NLP, has been significantly advanced with the advent of large language models (LLMs). These models have brought unprecedented improvements in accuracy, contextual understanding, and robustness. The introduction of transformer architecture by [26] revolutionized NLP by enabling models to capture long-range dependencies in text efficiently. This architecture is foundational to the development of LLMs [26].

[27] introduced BERT, a pre-trained model that set new benchmarks for various NLP tasks, including sentiment analysis. BERT's bidirectional approach allows it to understand the context of a word with its surrounding words, significantly outperforming traditional models in sentiment classification tasks [27]. [28] Optimized BERT by training on more data and using dynamic masking, leading to the development of the Robustly Optimized BERT Approach (RoBERTa). This model showed further improvements in sentiment analysis tasks, achieving higher accuracy and robustness [28]. [29] introduced Generative Pre-trained Transformer 3 (GPT-3), a

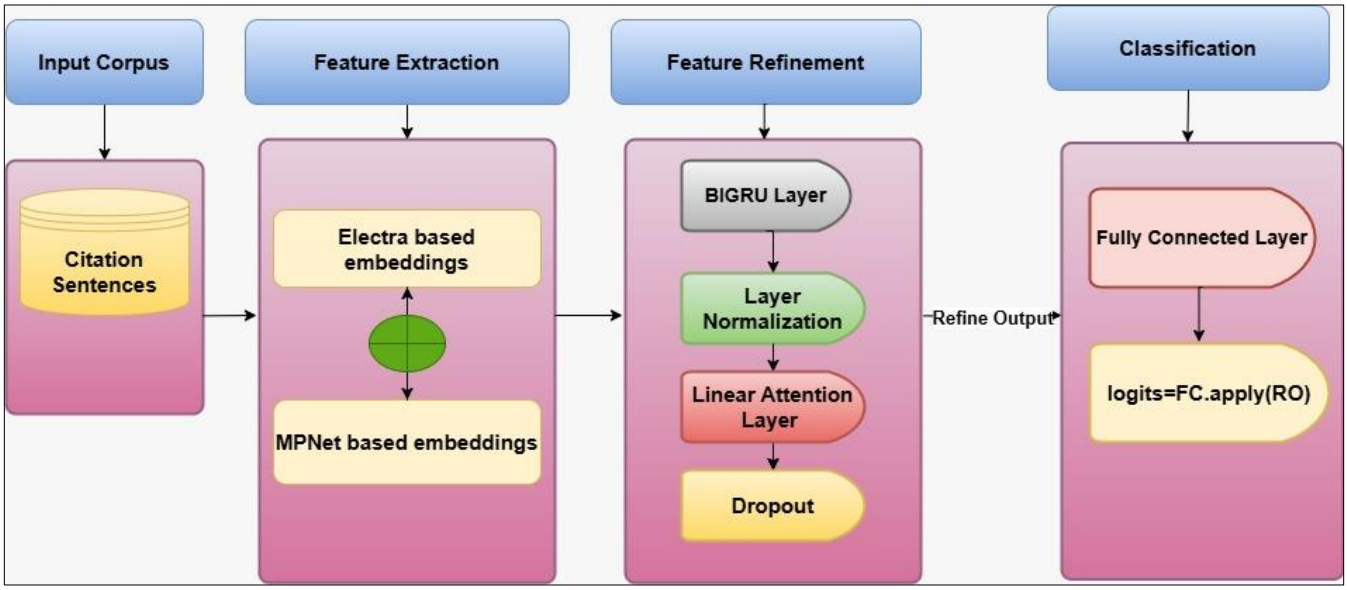


FIGURE 1. Proposed Electra-MPNet-based scientific sentiment classification framework

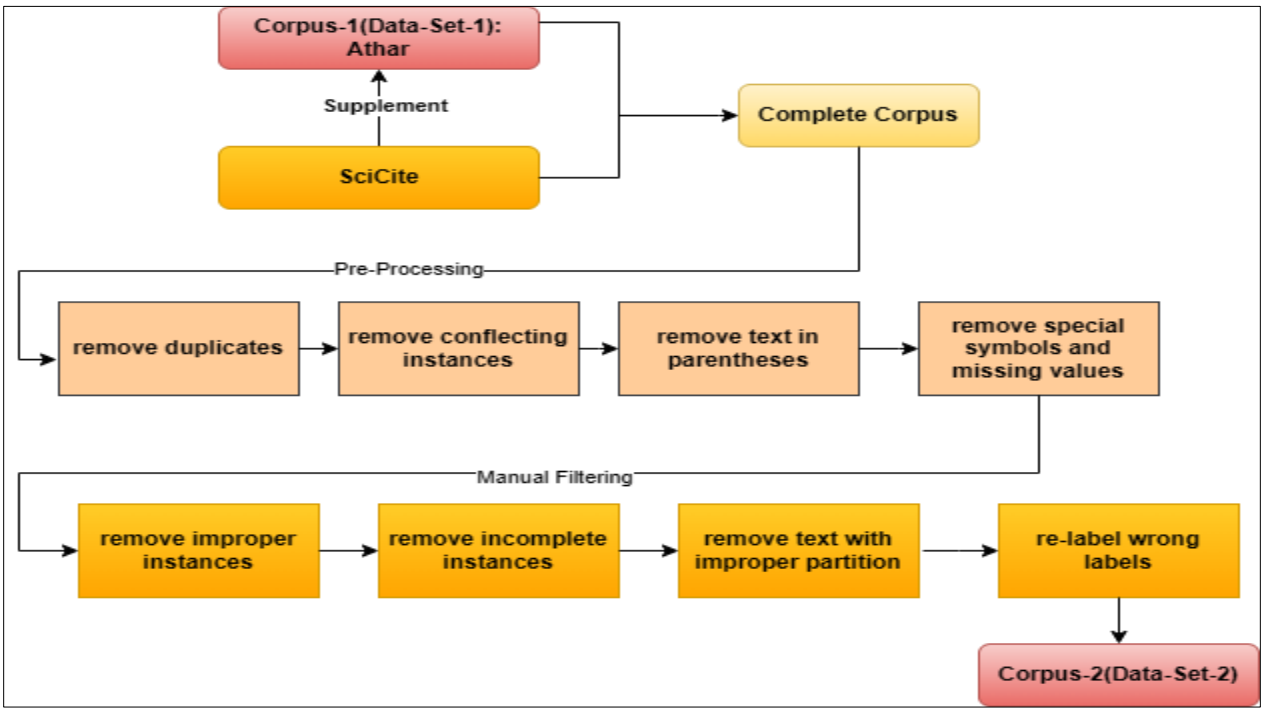


FIGURE 2. Diagram showing the detail of preprocessing steps and supplementation for data set-2

generative pre-trained transformer with 175 billion parameters. Known for its text generation capabilities, GPT-3 has also been effectively used for sentiment analysis due to its extensive training and ability to understand nuanced language [29]. [30] proposed the Text-To-Text Transfer Transformer (T5) model, which frames all NLP tasks as text-to-text transformations. This flexible framework has been successfully applied to sentiment analysis, demonstrating strong performance across various datasets [30].

LLMs are applied in various sentiment analysis domains. [31] utilized BERT for analyzing sentiments on social media platforms, effectively tracking public opinion and brand sentiment [31]. [32] demonstrated the application of RoBERTa for analyzing customer reviews, providing insights into customer satisfaction and product performance [32]. [33] employed GPT-3 to analyze market trends by processing large volumes of text data from news articles and forums, aiding businesses in making informed decisions based on public sentiment [33].

Algorithm-1: Electra-mpnet-based bigru-attention classifier scientific text sentiment classification procedure

Input: Dataset of citation sentences single label (multi-class: neutral, positive, negative)
Output: Scientific text sentiment classification results

Step 1: Initialized the model's Electra and MPNet encoder layers, a bidirectional GRU layer, layer normalization, a dropout layer, an attention mechanism, and a fully connected layer.

Step 2: Get the embeddings through the electra and mpnet encoders of scientific citation sentences separately and then concatenate them as features.

Step 3: Capture the bidirectional dependence of the embeddings generated in step 2 through the BiGRU layer

Step 4: Apply the layer normalization process to normalize the outputs of the BiGRU layer

Step 5: Impose a linear attention mechanism to refine the sequence quality generated in step-4 computing attention scores and derive a weighted context vector

Step 6: Apply dropout to the attention output (Step 5) to prevent overfitting by randomly setting some of the elements to zero.

Step 7: Use a fully connected layer to map the step-6 output to the number of output classes and return logits, as raw predictions of the model before application of loss function.

Step 8: Perform the training and evaluation of the proposed model using evaluation matrices

Despite their capabilities, LLMs present challenges. The computational demands of LLMs are substantial. Studies like [34] focus on developing more efficient models, such as Big Bird, which maintain performance while reducing resource requirements [34]. [35] emphasized the need for models that adapt to specific domains. Their work on domain-adaptive pre-training has shown promising results in enhancing model performance in specialized contexts [35]. [36] discussed the importance of interpretable models. As LLMs become more complex, understanding their decision-making processes is crucial for trust and adoption, especially in sensitive applications [36]. [37] raised ethical concerns about the use of LLMs, including bias in training data and potential misuse. Addressing these issues is critical for developing fair and transparent models [37].

D. RELATED WORK ON SCIENTIFC TEXT SENTIMENT ANALYSIS

Several researchers have investigated sentiment analysis on scientific text using both machine learning and deep learning techniques. For instance, in [1] and [2], machine learning models were utilized to analyze sentiment in scientific text citation sentences. [5] examined sentiment analysis on harassment sentences in the scientific literature using the LSTM-GRU model. Notably, there is limited research on advanced transformer-based models in this context; [38,44,45] utilized the traditional Bert model for sentiment analysis. As per the recommendations of this study [2] our contributions to the field of sentiment analysis from scientific text have been diverse. First, we look at two scientific citation sentence corpora, each one carefully annotated with attitudes classified as neutral, negative, or positive based on strict guidelines. On top of this base, we combined the potent characteristics of the Electra and MPNet with Bidirectional BiGRU layers to create an intricate model architecture along with an attention mechanism. Electra and MPNet are the foundation for feature extraction from textual inputs in this architecture, utilizing its ability to identify subtle patterns and situations.

In addition, the contextualized features are skillfully processed by the BiGRU layer, and then the linear attention mechanism is employed after the BiGRU layer, which improves the model's comprehension of sequential data and boosts its predictive power. The innovation of this paper is to build a new RNN-based architecture on the ELECTRA and MPNet combined strength along with the attention layer to improve the classification process of scientific text sentiment analysis.

III. Proposed Electra-MPNet-Based BiGRU Attention Classifier

This section details the proposed approach as illustrated in (Figure 1 and Algorithm-1).

A. ELECTRA

The ELECTRA model is an innovative transformer-based architecture for natural language processing that improves upon traditional models like BERT. Instead of using the masked language modeling (MLM) approach, ELECTRA employs a replaced token detection task, where a generator replaces some tokens in the input sequence and a discriminator determines whether each token is original or replaced. This allows ELECTRA to leverage more training signals from each input, making it more sample-efficient and

computationally efficient. Compared to BERT, ELECTRA achieves better performance on various NLP benchmarks with fewer resources, due to its ability to learn from both the original and replaced tokens during training [10].

B. MPNET

MPNet is a cutting-edge language model designed to enhance the capabilities of traditional models like BERT. It introduces a unique training approach that combines masked language modeling (MLM) and permuted language modeling (PLM), allowing the model to capture token dependencies more effectively by predicting masked tokens in a permuted context. This method enables MPNet to understand the context better than BERT, which treats each masked token prediction independently. Consequently, MPNet shows superior performance on various natural language understanding tasks, such as text classification and question answering, while being more sample-efficient and maintaining lower computational costs. Overall, MPNet offers improved accuracy, efficiency, and performance, making it a significant advancement over BERT in natural language processing applications [12].

C. FEATURE EXTRACTION USING ELECTRA AND MPNET

We employed a dual-model approach to extract features from text, leveraging the strengths of both the Electra and MPNet models. Each model was used independently to generate features from the same text input. Specifically, Electra was utilized for its discriminative pretraining capabilities, allowing for precise representation learning. Simultaneously, MPNet was employed to enhance context modeling by integrating masked and permuted language modeling, thereby capturing intricate word relationships more effectively. Once the features were extracted from each model separately, they were concatenated to form a unified feature representation. This concatenation leverages the complementary strengths of Electra's robust discrimination and MPNet's sophisticated context understanding, resulting in a comprehensive feature set. This combined feature representation was then used in downstream tasks, demonstrating improved performance in text sentiment analysis.

D. ELECTRA MPNET-BASED BIGRU ATTENTION CLASSIFIER AS AN INTEGRATED APPROACH

In the proposed model we extracted features from the citation sentences using the electra and mpnet encoders separately. The diverse set of features is then concatenated as the final features. To comprehend the complex relationships and dependencies found in the text, bidirectional data processing is critical, and this is made possible by the BiGRU layer. The proposed model leverages depth (number of BIGRU layers) and width (hidden size) to balance complexity and performance. Adding BIGRU layers enables capturing deeper text relationships for nuanced sentiment, but a single layer avoids overfitting and long training. A hidden size of 128 strikes a balance, capturing sufficient semantic details. These design choices optimize the model's ability to capture scientific sentiment effectively and efficiently. The BIGRU layer simultaneously processes this input sequence forward and backward. In this way, it records not only the context of the current token but also the context before and following it, enabling a more comprehensive comprehension of the sequence. As data moves across the network, each GRU unit in the BIGRU keeps track of internal states, such as cell states and hidden states. Because of this, the BIGRU extracts contextually important information and dependencies from various locations across the input sequence. After the normalization of BIGRU output linear attention mechanism is employed. This is performed in several steps:

The attention scores e_t are computed for each time step t using a linear transformation as given in Eq. (1)

$$e_t = Linear(h_t) \quad (1)$$

where h_t is the hidden state at time step t.

The softmax function is applied across the sequence length to obtain attention weights α_t given in Eq. (2)

$$\alpha_t = softmax(e_t) \quad (2)$$

Finally, the context vector c is computed as the weighted sum of the hidden states, weighted by the attention weights Eq. (3)

$$c = \sum_{t=1}^{T} \alpha_t . h_t \quad (3)$$

where T is the sequence length.

The linear attention mechanism allows the model to focus on the most relevant parts of the input sequence when making predictions. These final context vectors pass to a fully connected layer for classification and prediction. The fully connected layers are used for further classifications and label predictions. In the training process, cross-entropy optimization is employed. This approach is particularly suitable for classification tasks. The function combines the functions of normalizing output probabilities and evaluating their divergence from the true labels.
The Cross-Entropy loss L for classification is given by:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (4)$$

where

- y is the true label distribution.
- $\hat{y}$ Is the predicted label distribution?
- C is the number of classes or labels.

The goal during training is to minimize this loss:

$$\min L(y, \hat{y})$$

IV. EXPERIMENTS

In this section, we elaborated on the experimental setup and results of the proposed framework as follows.

A. DATASET

We employed two datasets for our study. The first data set we chose was created in [4]. In the first data set Association for Computational Linguistics (ACL) anthology journal and conference articles are included [39,40]. About 21,800 papers were housed in the anthology at the time of investigation in [4]. There are 8736 citation sentences in the complete corpus of data set 1. The method of annotating data set-1 followed certain rules. Three sorts of citation sentences were identified: neutral, negative, and positive. The following is a summary of the annotation guidelines used for this task:

1) NEGATIVE SENTIMENT ANNOTATION GUIDELINES

a. When a fault or deficiency of the referenced paper is directly mentioned, mark it as negative.
b. If the referenced paper outperforms the referenced one, mark it as negative.
c. If the referenced paper offers a better evaluation than the referenced paper, mark it as negative.

2) NORMS FOR POSITIVE SENTIMENT ANNOTATION

a. If a direct reference to a favorable aspect of the referenced work is made, mark it as positive.
b. If the referenced work does not get better than the referenced one, mark it as positive.
c. If there's no outperform evaluation provided by the citing paper to the cited paper, mark it as positive.

3) NEUTRAL SENTIMENT ANNOTATION GUIDELINES

a. If opinions about the citation sentence are neither favorable nor unfavorable, mark it as neutral.

TABLE 1. Statistics for corpus annotation of dataset-1.

Sentiment	Count	Percentage
Positive	829	9.5
Negative	280	3.2
Neutral	7627	87.3
Total	8736	100

TABLE 2. Statistics for corpus annotation of dataset-2.

Sentiment	Count	Percentage
Positive	1237	15.63
Negative	347	4.39
Neutral	6328	79.98
Total	7912	100

The data set mentioned in Table 1 is the most recommended data set for citation sentiment analysis proposed by [4]. To further improve the generalization of the dataset and showcase the model performance another data set proposed in [41] is added for experimentation in the current study as shown in Table 2. To address potential biases, the authors in [41] apply data supplementation techniques and other pre-processing steps to enhance dataset diversity and quality. For example, remove missing values, discard instances with the same text but different labels, and delete duplicate entries with identical text and labels, strip text within parentheses, eliminate all special characters. This approach aims to balance representation, helping the model better capture diverse scientific perspectives and reduce bias in sentiment analysis. For conducting the data supplementation SCICite data set was utilized as proposed in [42]. In this process, about 1000 sentences from SCICite are used to supplement the corpus proposed in [4] as a result new corpus was generated as given in Table 2, and a detailed process is shown in (Figure 2). This is more diverse and generalized to better model experimentation and validation. Both data sets are divided as 60% of these citation sentences are randomly selected for training the classifier, while the remaining 5% are reserved for validation. The remaining 35% of the data is earmarked for testing purposes.

B. EXPERIMENTAL SETUP

The proposed method was implemented on Google Colab, leveraging its access to NVIDIA T4 GPUs, which offer efficient deep learning capabilities at no cost for prototyping and experimentation. Google Colab provides a convenient cloud-based environment for executing machine learning and deep learning tasks, allowing researchers to utilize powerful hardware like the T4 GPU without local resource constraints. With pre-installed support for essential libraries such as TensorFlow, PyTorch, CUDA, and TensorRT, Colab streamlines model development and testing. This cloud platform is particularly valuable for projects requiring quick setup, seamless collaboration, and access to high-performance hardware, making it an ideal choice for our experimentation. The detailed experimental configuration is given in Table 3.

We trained our model using the AdamW optimizer with a learning rate and weight decay. A linear learning rate scheduler with warmup adjusts the learning rate over num_training_steps. Early stopping with a patience level of 3 halts training if validation loss doesn't improve for three consecutive epochs, preventing overfitting. All models used as comparison baselines for the proposed model are evaluated on the same dataset used in the current study.

C. EVALUATION MATRICES AND SOTA BASELINES

Any research project's evaluation establishes its position and quality within its field. An overview of the measures used to evaluate our developed sentiment analysis system is given in this section. The accuracy of the sentiment analysis system's

TABLE 3. The elements employed in the proposed model implementation

Component	Description	Parameters
Electra Model	Pre-trained Electra model for initial embedding generation	Model: google/electra-base-discriminator
MPNet Model	Pre-trained MPNet model for additional embedding generation	Model: microsoft/mpnet-base
GRU Layer	Bi-directional GRU layer to capture temporal dependencies in concatenated embeddings	Hidden size: 128, Num layers: 1, Bidirectional: True, Batch first: True
Layer Normalization	Layer normalization applied to the GRU output	Dimension: hidden_size * 2 (256)
Dropout Layer	Dropout for regularization to prevent overfitting	Dropout rate: 0.5
Attention Mechanism	Custom attention layer to focus on significant parts of the GRU output	Attention weights layer: Linear(hidden_size * 2, 1)
Fully Connected Layer (FC)	Final linear layer for class prediction	Input size: hidden_size * 2 (256), Output size: num_classes (3)
Cross-Entropy Loss	Loss function for multi-class classification	None
Optimizer	AdamW optimizer	Learning rate: 2e-5, Weight decay: 1e-4
Scheduler	Linear scheduler with warmup for learning rate adjustment	Warmup steps: 0, Total steps: num_epochs * len(train_loader)
Training Batch Size	Size of each training batch	32
Max Sequence Length	Maximum length of input sequences	50
Number of Epochs	Total number of training epochs	100
Patience for Early Stopping	Number of epochs to wait for improvement before stopping	3
Device	Hardware for model training and inference	CUDA

classification results are used to determine how effective it is. Overall accuracy is determined by the percentage of correctly classified items as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

Precision can be defined as the ratio of requirements in the class to requirements that the model correctly identified with the appropriate label. This is how the prediction measure is calculated:

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

The recall is determined by dividing the total number of relevant labels that should have been anticipated by the number of properly predicted labels, which is defined as

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

We calculated the Macro-F Score as well as the Micro-F Score throughout our evaluation phase. In this case, FP (False Positive) errors are classified as type-1 errors, and FN (False Negative) errors are classified as type-2 errors. The F-score, which is frequently used in these kinds of evaluations, is the harmonic means of recall and precision, and it is defined as

$$\text{F1} - \text{Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad (8)$$

where TP (True Positive) is the number of correctly predicted positive instances, TN (True Negative) is the number of correctly predicted negative instances, FP (False Positive) is the number of incorrectly predicted positive instances (predicted positive but negative), and FN (False Negative) is the number of incorrectly predicted negative instances (predicted negative but positive).

Macro-F1 score: The Macro-F1 score is an evaluation metric used to assess the performance of classification models. It is the arithmetic mean of the F1 scores of each class. The F1 score itself is the harmonic mean of precision and recall. The Macro-F1 score treats all classes equally by calculating the F1 score for each class independently and then averaging them. This approach ensures that each class has the same impact on the final score, regardless of its frequency in the dataset. The importance of the Macro-F1 score lies in its ability to give equal importance to all classes. This is particularly crucial in datasets with imbalanced classes where some classes may be underrepresented. By focusing on the average performance across all classes, the Macro-F1 score ensures that the model performs well not only in the majority classes but also in the minority ones. This makes it a valuable metric for applications where balanced performance across all categories is necessary, such as medical diagnosis or fraud detection [4,1].

Micro-F1 score: The Micro-F1 score, on the other hand, aggregates the contributions of all classes to compute the average metric. It calculates the overall precision and recall by summing up the true positives, false positives, and false negatives for all classes, and then computes the F1 score from these aggregated values. This method gives more weight to the performance of the model on the larger classes. The Micro-F1 score is important because it reflects the overall performance of the classification model by considering the total number of true positives, false

positives, and false negatives across all classes. It is particularly useful when the class distribution is imbalanced, and the focus is on the overall accuracy of the model. By aggregating the contributions of all classes, the Micro-F1 score provides a more comprehensive evaluation of the model's performance, especially in scenarios where the correct classification of all instances, regardless of their class, is critical [4,1].

Cohen's Kappa quantifies inter-rater reliability by adjusting for agreement expected by chance.

$$\text{Cohen's Kappa } (\kappa)=\frac{P_0-P_e}{1-P_e} \qquad (9)$$

where P_o is the observed agreement, and P_e is the expected agreement by chance.

We compare our proposed approach with **SOTA** models for scientific text sentiment analysis and other transformer-based sentiment analysis models on this data set to prove the effectiveness of the proposed approach as follows:

1) SOTA BASELINES

- **i)** NB [4] and others [1] utilize NB, SVM, LR, DT, KKN, and RF algorithms for sentiment analysis of scientific articles using citation sentences on the data set-1.
- **ii)** LSTM-GRU [5]: We implemented this model on our data set-1 and 2 under the same settings suggested by the author in [5] using glove embeddings to compare it with our approach as the SOTA model because it addresses a similar problem context which is scientific literature sentiment analysis.
- **iii)** Bert-BiLSTM [8]: We implemented this model on our data set-1 and 2 under the same settings suggested by the author in [8] using Bert-based embeddings to compare it with our approach as the SOTA model because this model is proposed for sentiment analysis of investors and consumers statements.
- **iv)** Bert-LSTM-CNN [7]: We implemented this model on our data set-1 and 2 under the same settings suggested by the author in [7] using Bert-based embeddings to compare it with our approach as the SOTA model because this model is proposed for sentiment analysis of the Indonesian language e-commerce platform consumer review data.

2)ABLATION ANALYSIS

- **i)** BIGRU [9]: It is a type of Recurrent Neural Network (RNN), in which processors input sequences in both forward and reverse directions, taking into account the dependencies from both past and future contexts at the same time. In a nutshell, it is an extension of the GRU model.
- **ii)** Electra [10]: ELECTRA is a pre-training approach aimed at addressing natural language understanding tasks in the context of natural language.
- **iii)** MP Net [12]: MPNet a pre-training method, combines BERT and XLNet strengths while addressing their limitations. It outperforms previous models on downstream tasks, including BERT and XLNet.

V) RESULTS

The result section is divided into two sub-sections in the first section we discuss the comparison between the proposed model and the SOTA baseline models and in the second section we discuss the ablation analysis of the proposed model.

TABLE 4. Overall performance analysis of the proposed model with the machine, deep learning models based on F1-Macro, and F1-Micro score for data set-1.

Sr. No	Models	F1-Macro	F1-Micro	Kappa
1	**Proposed Model**	**0.6452**	**0.8950**	**0.5231**
2	LSTM-GRU(Glove) [5]	0.3891	0.8676	0.1045
3	Bert-LSTM-CNN [7]	0.5846	0.8829	0.4673
4	Bert-BiLSTM [8]	0.6008	0.8898	0.4855
5	Random Forst [1]	0.44	0.88	0.1510
6	Decision Tree [1]	0.49	0.85	0.3122
7	Naïve Bayes [4]	0.30	0.78	0.0000
8	SVM [1]	0.37	0.88	0.1903
9	KNN [1]	0.33	0.87	0.1224
10	LR [1]	0.49	0.88	0.2263

A. COMPARATIVE ANALYSIS OF THE PROPOSED MODEL WITH SOTA BASELINES

1) EXPERIMENTS WITH DATA SET-1

The proposed approach is compared to various SOTA baseline models on the dataset-1 to appreciate its effectiveness in improving the scientific text sentiment classification as shown in (Table 4). In comparison to other established models, the proposed model shows significant improvements in performance metrics when compared to other existing models. As an example, when compared with the LSTM-GRU (Glove) [5], the proposed model exhibits an impressive 65.76% improvement in F1-Macro and a 3.16% improvement in F1-Micro score. Against the Bert-LSTM-CNN [7], the proposed model exhibits a 10.36% improvement in F1-Macro and a 1.37% improvement in F1-Micro. When compared to Bert-BiLSTM [8], the proposed

TABLE 5. The classification report of the proposed model and others depicting the precision, recall, and f1-score of various classes for data set 1.

Model	Matrices	Neutral	Positive	Negative
Proposed Model	**Precision**	**0.93**	**0.66**	**0.44**
	Recall	**0.96**	**0.56**	**0.34**
	F1-Score	**0.94**	**0.61**	**0.39**
	Accuracy	**0.8950**		
MP Net [12]	Precision	0.95	0.49	0.27
	Recall	0.90	0.68	0.36
	F1-Score	0.92	0.57	0.31
	Accuracy	0.8548		
Electra [10]	Precision	0.94	0.52	0.39
	Recall	0.93	0.65	0.21
	F1-Score	0.94	0.58	0.27
	Accuracy	0.8780		
BiGRU [9]	Precision	0.88	0.56	0.00
	Recall	0.99	0.15	0.00
	F1-Score	0.93	0.24	0.00
	Accuracy	0.8676		
LSTM-GRU(Glove) [5]	Precision	0.90	0.54	0.00
	Recall	0.97	0.33	0.00
	F1-Score	0.93	0.41	0.00
	Accuracy	0.8738		
Bert-BiLSTM [8]	Precision	0.93	0.60	0.45
	Recall	0.96	0.55	0.24
	F1-Score	0.94	0.57	0.31
	Accuracy	0.8898		
Bert-LSTM-CNN [7]	Precision	0.92	0.60	0.39
	Recall	0.96	0.49	0.22
	F1-Score	0.94	0.54	0.28
	Accuracy	0.8829		

model achieves a 7.39% increase in F1-Macro and a 0.58% increase in F1-Micro, demonstrating better attention mechanisms and hybrid architecture efficiency.

In the proposed approach when compared to the RF, there is a notable improvement in F1-Macro of 46.18% and in F1-Micro of 1.70%. In addition, the DT also falls short, with the proposed model showing a 31.67% and 5.29% improvement in F1-Macro and F1-Micro, respectively [1,4].
In comparison with NB, the proposed model achieved a remarkable 79.22% improvement in F1-Macro and a 2.87% increase in F1-Micro [4,1]. It was found that the proposed model outperformed SVM in terms of F1-Macro by 74.11 percent and in terms of F1-Micro by 1.70 percent. The proposed model showed the highest relative improvement in comparison with the KNN model achieving a 95.21% better F1-Macro score and a 2.87% better F1-Micro score. Finally, in comparison to LR, the proposed model shows a 31.07% enhancement in F1-Macro and a 1.70% enhancement in F1-Micro [1,4]. It can be concluded that the proposed model consistently outperforms other models, demonstrating its superior ability to handle the classification task more efficiently than other models.

TABLE 6. Overall performance analysis of the proposed model with the machine, deep learning models based on F1-Macro, and F1-Micro score for data set-2.

Sr. No	Models	F1-Macro	F1-Micro	Kappa
1	**Proposed Model**	**0.8239**	**0.9336**	**0.7740**
2	LSTM-GRU(Glove) [5]	0.5310	0.8787	0.5616
3	Bert-LSTM-CNN [7]	0.7505	0.8960	0.7595
4	Bert-BiLSTM [8]	0.7798	0.9133	0.7341
5	Random Forst	0.5572	0.8727	0.4870
6	Decision Tree	0.5994	0.8332	0.4874
7	Naïve Bayes	0.3507	0.8240	0.1039
8	SVM	0.6270	0.8834	0.5530
9	KNN	0.4877	0.8382	0.3249
10	LR	0.5386	0.8711	0.4612

2) EXPERIMENTS WITH DATA SET-2

The proposed approach is also evaluated as compared to other SOTA baselines on data set-2 to validate its effectiveness for scientific text sentiment analysis. The results are shown in Table 6. It can be seen that the proposed model outperforms as compared to other models with an F1-Macro score of 0.8239, and an F1-Micro score of 0.9336 in addition we also calculated the Kappa score for the proposed and all other baseline models. The proposed model showed a higher value for the kappa score (0.7740).

B. ABLATION ANALYSIS

In our proposed model, each component plays a distinct and essential role in enhancing text representation and classification accuracy. For example, ELECTRA contributes primarily through its efficient token-level understanding. Unlike traditional masked language models, ELECTRA is trained to predict real tokens rather than masked ones, making it highly sensitive to nuanced linguistic details at the token level. MPNet brings sentence-level coherence and structural context, bridging local details with broader semantic meaning. By combining permuted language modeling and masked language modeling techniques, MPNet effectively captures dependencies within a sentence, making it ideal for tasks requiring both fine-grained and contextual understanding. The bidirectional GRU adds depth

by modeling sequential information in both forward and backward directions. This enables the model to understand the flow of information throughout the text, capturing dependencies between tokens across longer spans. BiGRU thus enriches the context by preserving order and sequence, which improves the model's understanding of complex sentence structures. Linear attention improves interpretability and model efficiency and serves as a final refinement layer.

1) EXPERIMENTS WITH DATA SET-1

In the ablation analysis, the proposed model was compared against several other standalone components based on its performance in terms of F1-Macro scores and F1-Micro scores, to evaluate its effectiveness as shown in (Table 7). As compared to all other models in both metrics, the proposed model has reached an F1-Macro score of 0.6452 and an F1-Micro score of 0.8950, outperforming all others in all metrics on dataset-1. It has been observed that the BiGRU model [9] showed an F1-Macro score of 0.3891 and an F1-Micro score of 0.8676. In comparison to the BiGRU model, the proposed model shows a remarkable improvement of 65.76% in F1-Macro and 3.16% in F1-Micro. While the Electra model [10] performed better than BiGRU, it still fell short of the proposed model when measured in F1-Macro and F1-Micro scores with an F1-Macro score of 0.5954 and an F1-Micro score of 0.8780. In terms of F1-Macro and F1-Micro, this indicates that the proposed model is 8.36% better than Electra in F1-Macro and 1.94% better than F1-Micro. It can be noted that the MPNet model [12] also produced an F1-Macro score of 0.5989, as well as an F1-Macro score of 0.8548.

A comparison of the proposed model to MPNet demonstrates a 7.74% improvement in F1-Macro and a 4.70% improvement in F1-Micro. This study shows that the proposed model consistently outperforms the other models in both F1-Macro and F1-Micro scores, highlighting its superior performance in the given task compared to other models. Based on this comprehensive comparison, it is evident that the proposed model is both effective and robust.

TABLE 7. The ablation analysis study for data set-1.

Sr. No	Models	F1-Macro	F1-Micro	Kappa
1	Proposed Model	0.6452	0.8950	0.5231
2	BiGRU [9]	0.3891	0.8676	0.0053
3	Electra [10]	0.5954	0.8780	0.4310
4	MPNet [12]	0.5989	0.8548	0.4898

2) EXPERIMENTS WITH DATA SET-2

Similarly in the ablation analysis, the proposed model was compared against several other standalone components on data set-2 as shown in Table 8. For example, in isolation, ELECTRA achieved a respectable F1-Macro of 0.5619, which highlights its ability to strengthen foundational representations. The MPNet as a standalone model reached an F1-Macro of 0.7906, showing its significant contribution in forming coherent sentence embeddings and improving overall model stability. However, as seen in Table 8, BiGRU alone (F1-Macro of 0.4946) struggles with complex representations, underscoring the need for integration with transformer models like ELECTRA and MPNet. Linear attention improves interpretability and model efficiency which boosts the F1-Macro score to 0.8239 in our combined model. This attention mechanism balances the model's focus on meaningful features, enabling more precise predictions and improving overall robustness.

C.COMPARISION OF PROPOSED MODEL WITH SOTA MODELS AND OTHERS FOR HANDLING IMBALANCED CLASSES

1) EXPERIMENTS WITH DATA SET-1

The sentiment analysis presents several difficulties in handling the imbalance classes found in scientific text. The proposed model handles it effectively. By combining the advantages of the Electra MPNet and Bidirectional GRU (BiGRU) layers along with linear attention, this model enables precise sentiment classification and a more sophisticated comprehension of scientific textual input. In comparison with other models (Table 5), the proposed model

TABLE *8*. The ablation analysis study for data set 2.

Sr. No	Models	F1-Macro	F1-Micro	Kappa
1	Proposed Model	0.8239	0.9336	0.7740
2	BiGRU[9]	0.4946	0.8393	0.4742
3	Electra[10]	0.5619	0.8920	0.6639
4	MPNet[12]	0.7906	0.9180	0.7487

shows the best overall balance between precision, recall, and F1-score across all classes, especially for the Positive and Negative classes, and overall accuracy when compared to other models.

It shows a significant improvement in handling imbalanced classes, with an accuracy of 0.8950, which is higher than the accuracy achieved by the other models that were evaluated. It is important to note that the BiGRU [9] and LSTM-GRU [5] models, in particular, struggle when it comes to identifying instances of the Negative class. There is a reasonable level of performance for both Electra [10] and MPNet [12] models but they do not perform as well as the proposed model in the Positive and Negative classes.

In comparison with Bert-BiLSTM [8] and Bert-LSTM-CNN [7], the proposed model exhibits higher precision, recall, and F1-Score across different classes, particularly for the minority Negative class. For the Neutral class, all models show high precision, recall, and F1-Score, indicating they handle the majority class well, but the proposed model maintains a slightly better balance across metrics. For the positive class, the proposed model shows a

notable improvement in recall (0.56) and F1-Score (0.61) compared to Bert-BiLSTM (recall: 0.55, F1-Score: 0.57) and Bert-LSTM-CNN (recall: 0.49, F1-Score: 0.54), suggesting it is better at identifying instances of this class and balancing precision and recall. The most significant improvement is observed in handling the Negative class,

TABLE 9. The classification report of the proposed model and others depict the precision, recall, and f1-score of various classes for dataset-2

Model	Matrices	Neutral	Positive	Negative
Proposed Model	**Precision**	**0.95**	**0.89**	**0.74**
	Recall	**0.98**	**0.70**	**0.63**
	F1-Score	**0.96**	**0.82**	**0.69**
	Accuracy	**0.9336**		
MP Net [12]	Precision	0.96	0.78	0.59
	Recall	0.95	0.83	0.63
	F1-Score	0.96	0.81	0.61
	Accuracy	0.9180		
Electra [10]	Precision	0.96	0.64	0.00
	Recall	0.94	0.87	0.00
	F1-Score	0.95	0.74	0.00
	Accuracy	0.8920		
BiGRU [9]	Precision	0.91	0.52	0.00
	Recall	0.92	0.63	0.00
	F1-Score	0.91	0.57	0.00
	Accuracy	0.8393		
LSTM-GRU(Glove) [5]	Precision	0.91	0.69	0.00
	Recall	0.97	0.63	0.00
	F1-Score	0.94	0.66	0.00
	Accuracy	0.8787		
Bert-BiLSTM [8]	Precision	0.96	0.78	0.56
	Recall	0.94	0.83	0.61
	F1-Score	0.95	0.80	0.58
	Accuracy	0.9133		
Bert-LSTM-CNN[7]	Precision	0.98	0.64	0.67
	Recall	0.91	0.93	0.47
	F1-Score	0.94	0.76	0.55
	Accuracy	0.8960		

where the proposed model achieves a higher F1-Score (0.39) than both Bert-BiLSTM (0.31) and Bert-LSTM-CNN (0.28), with better precision (0.44 vs. 0.45 and 0.39, respectively) and recall (0.34 vs. 0.24 and 0.22, respectively). This indicates that the proposed model is more effective at managing class imbalance and ensuring that even less frequent classes are accurately identified and classified. Overall, the proposed model's superior handling of imbalanced classes leads to more balanced and robust classification outcomes, ensuring more reliable and accurate sentiment classification, making it a preferable choice over Bert-BiLSTM and Bert-LSTM-CNN as well for tasks involving class imbalance.

2) EXPERIMENTS WITH DATA SET-2

It can be noted from Table 9 that the proposed model performs well in handling the class imbalance problem. Its accuracy is further enhanced in dataset-2, particularly in handling the classification of negative or minority classes as compared to dataset-1. Hence the proposed model continuously outperformed as compared to baseline models for handling the class imbalance problem.

D.COMPLEXITY ANALYSIS

We estimated the training time on the dataset-2 and no. of parameters of the proposed model and others to report the complexity analysis as shown in Table 10.

The proposed model has approximately 220M parameters and it takes approximately 352.95 seconds for training. While the Bert-BiLSTM has approximately 111M parameters it takes about 225.83 seconds for training. The proposed model has a significantly higher number of parameters compared to other deep learning-based models, yet it outperformed all of them. This highlights the trade-off between computational cost and model effectiveness. While a larger number of parameters can enable the model to learn more complex patterns, it also demands more computational resources.

TABLE 10. The comparative analysis of the computational cost of deep learning models on dataset 2

Models	No. of Parameters (million)	Time(s)
Proposed Model	**˜220**	**352.95**
MPNet	˜110	338.09
Electra	˜110	138.75
LSTM-GRU-Glove	˜120	16.12
Bert-BiLSTM	˜111	225.83
Bert-LSTM-CNN	˜111	221.26

VI) DISCUSSIONS

The discussion revolves around a comprehensive comparison of the proposed model (Electra-MPNet-BiGRU attention Classifier) against both traditional machine learning algorithms and advanced deep learning architectures [5] and transformer-based models [7,8] for sentiment analysis within scientific text. Leveraging a diverse corpus of 8,736 citation sentences, the proposed model consistently outperforms conventional methods such as Decision Trees, Random Forest, Logistic Regression, and Support Vector Machines [1], [4] and advanced methods of deep learning [5] and transformer-based [07,08] across all sentiment categories—positive, negative, and neutral. This superiority is evident in metrics like Macro and Micro F1-score, indicating the proposed model's adeptness at capturing

the nuanced sentiment nuances prevalent in scholarly discourse.

One of the remarkable strengths of the proposed model lies in its ability to navigate imbalanced class distributions, a common challenge in sentiment analysis tasks. Despite the inherent complexity posed by class imbalance, the proposed model demonstrates remarkable resilience, accurately classifying both positive and negative sentiments with high precision. This robust performance underscores the model's reliability and its potential to deliver dependable sentiment categorization outcomes without compromising precision, thus providing researchers with valuable insights into the sentiment landscape of scientific text.

Moreover, the proposed model capitalizes on the synergies between the Electra [10], MPNet [12], and Bidirectional GRU [09] with linear attention layers [11], facilitating precise sentiment classification and a nuanced understanding of textual inputs. By seamlessly integrating these layers, the proposed model achieves superior performance metrics across various evaluation benchmarks, particularly excelling in discerning positive and negative sentiments. The ability of the proposed model to harness the pre-trained contextual understanding from Electra and MPNet, combined with the sequence learning capabilities of BiGRU and the focusing power of attention mechanisms, results in an architecture that is both robust and insightful.

An insightful ablation study highlights the pivotal role of each component, with Electra [10], MPNet [12], Bidirectional GRU [9], and attention layers [11] exhibiting lower performance when analyzed independently. This nuanced understanding of model components underscores the efficacy of the proposed approach in enhancing sentiment analysis within scientific texts, offering not only enhanced accuracy but also improved interpretability compared to traditional algorithms and other deep learning models. The study reveals that the combination of these advanced techniques allows the model to not only understand the contextual depth of each sentence but also maintain the temporal relationships within the text, crucial for accurate sentiment classification. These research implications signify a significant advancement in the field of sentiment analysis within the scientific literature, empowering researchers and practitioners with a robust and reliable tool for extracting valuable insights from vast repositories of academic texts. The proposed model's ability to handle imbalanced datasets, coupled with its superior performance in classification tasks, represents a meaningful leap forward in sentiment analysis. This advancement has the potential to greatly enhance the quality and depth of insights that can be gleaned from academic texts, ultimately contributing to more informed and nuanced research outcomes.

To address potential biases in citation sentiment analysis, this study adopts a multi-faceted approach. Recognizing the need for diversity and representation, a data supplementation technique was applied using the SCICite dataset, following methods proposed in [41] and [42]. This supplementation introduces an additional 1,000 sentences from SCICite to the original dataset from [4], enhancing diversity in both language and sentiment contexts across various scientific domains. This augmented dataset, illustrated in Table 2 and Figure 2, is designed to increase generalization and reduce potential biases by including more varied expressions of scientific sentiment.

Addressing bias further, we acknowledge that domain-specific language or cultural factors could influence sentiment in scientific text. The model's performance was evaluated to identify if biases may emerge, particularly in fields where sentiment expression might vary by discipline or cultural norms. By broadening the dataset, we aim to create a more balanced corpus that supports fairer sentiment classification across different scientific perspectives. This expanded dataset is intended to minimize biases that may otherwise affect the model's generalization, promoting a fairer analysis of sentiment.

Overall, the Electra-MPNet-BiGRU attention Classifier demonstrates a comprehensive and effective approach to sentiment analysis in scientific texts, significantly outperforming traditional methods and showcasing the potential for sophisticated deep-learning models to handle complex and nuanced language processing tasks in academic research. The use of Electra and MPNet allows the model to leverage the latest advancements in transformer architectures, enabling it to capture subtle semantic and syntactic nuances. The BiGRU component ensures that the model can effectively process sequential data, while the attention mechanism allows for a focused analysis of the most relevant parts of the text.

Furthermore, the ability to handle diverse and large-scale datasets makes this model highly adaptable and scalable for various applications in scientific text mining. Researchers and practitioners can apply this model to other domains within the scientific literature, potentially uncovering new insights and trends that were previously difficult to detect with conventional methods. The comprehensive evaluation of the model across multiple metrics and datasets also provides a robust benchmark for future research in sentiment analysis, setting a high standard for performance and reliability.

The combination of cutting-edge techniques and thorough evaluation underscores the significance of this work in advancing the state of the art in sentiment analysis. By addressing the limitations of existing methods and introducing a novel, high-performing model, this research contributes valuable knowledge and tools to the field, paving

the way for more sophisticated and accurate analysis of scientific texts. The practical implications of this work are far-reaching, offering a powerful solution for extracting meaningful insights from the ever-growing body of academic literature.

VII) CONCLUSION AND FUTURE WORK

The study was carried out in the context of the growing body of scientific literature that is accessible online, which poses a significant difficulty in managing and retrieving information. Sentiment analysis is a crucial method for overcoming this obstacle since it makes it possible to find complex attitudes in this enormous collection of academic writing. Unfortunately, current approaches frequently fail to capture the subtle differences in word meanings within phrases due to context. The proposed Electra-MPNet-BIGRU Attention Classifier a complex architecture that combines the reliable qualities of Electra MPNet with Bidirectional GRU (BiGRU) and linear attention layers, was offered by the study as a solution to this problem. This proposed model uses Electra and MPNet combined strength to recognize complex patterns and contexts in textual inputs to get around the limitation. Furthermore, the BiGRU layer skillfully processes the contextual nuances, improving the model's understanding of sequential data and thereby increasing its predictive accuracy. Considering this, the proposed model's performance was assessed by contrasting it with a range of deep learning benchmarks, including LSTM-GRU (glove), BiGRU, and transformer-based approaches such as MP Net, Electra, Bert-BiLSTM and Bert-LSTM-CNN, in addition to cutting-edge machine learning methodologies. Finding out how well the suggested model handled the problems with sentiment analysis in scientific text was the main objective. The study's findings confirmed that the proposed model is the best at classifying attitudes in scientific writing, demonstrating its dominance across a range of evaluation parameters. Notably, our model exhibits remarkable adaptability to the challenges posed by imbalanced sentiment classes, notably excelling in accurately discerning positive and negative sentiments amidst a heterogeneous dataset. As scholarly discourse continues to proliferate, the ability to accurately parse sentiments within scientific texts holds profound implications for various domains, including academic research, policy formulation, and public discourse. These results add to the continuing efforts to improve sentiment analysis techniques and expand our ability to extract knowledge from the enormous body of online scholarly literature.

For future work, the Electra-MPNet-based BiGRU attention classifier could be extended to handle multilingual datasets. Exploring other transformer models, and integrating knowledge graphs, Additionally, real-time sentiment analysis applications and efforts to improve model explainability and fine-grained sentiment detection offer promising directions for further research.

DATA AVAILABILITY

The data set used in this study can be accessed at
https://awaisathar.com/citation-sentiment-corpus/
https://github.com/allenai/scicite
https://github.com/UFOdestiny/DictSentiBERT/blob/main

REFERENCES

[1] H. Raza, M. Faizan, A. Hamza, A. Mushtaq, and N. Akhtar, "Scientific text sentiment analysis using machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 157-165, 2019.
[2] K. Aristotelis, S. Antonis, D. Konstantinos, and O. Stefanos, "Sentiment dimensions and intentions in scientific analysis: multilevel classification in text and citations," *Electronics*, vol. 13, no. 9, pp. 1753, 2024.
[3] W. Nie and S. Ou, "Micro citation importance identification and its application to literature evaluation," in *Proceedings of the International Conference on Information*, Cham: Springer Nature Switzerland, pp. 356-375, 2024.
[4] A. Athar, "Sentiment analysis of scientific citations," University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-856, 2014.
[5] H. Q. Low, P. Keikhosrokiani, and M. Pourya Asl, "Decoding violence against women: analysing harassment in Middle Eastern literature with machine learning and sentiment analysis," *Humanities and Social Sciences Communications*, vol. 10, no. 11, pp. 1-8, 2024.
[6] C. Xiang, J. Zhang, J. Zhou, F. Li, C. Teng, and D. Ji, "Phrase-aware financial sentiment analysis based on constituent syntax," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
[7] H. Murfi, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis," Applied Soft Computing, vol. 151, pp. 111112, 2024.
[8] R. Cai, B. Qin, Y. Chen, L. Zhang, R. Yang, S. Chen, and W. Wang, "Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM," IEEE Access, vol. 8, pp. 171408-171415, 2020.
[9] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724-1734, 2014.
[10] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.
[11] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," In International Conference on Machine Learning, pp. 5156-5165, 2020.
[12] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857-16867, 2020.
[13] M. S. K. Abadah, P. Keikhosrokiani, and X. Zhao, "Analytics of public reactions to the COVID-19 vaccine on Twitter using sentiment analysis and topic modelling," in *Handbook of Research on Applied Artificial Intelligence and Robotics for Government Processes*, D. Valle-Cruz, N. Plata-Cesar, and J. L. González-Ruíz, Eds., IGI Global, pp. 156-188, 2023.
[14] M. A. Z. B. M. Asri, P. Keikhosrokiani, and M. P. Asl, "Opinion mining using topic modeling: a case study of Firoozeh Dumas's *Funny in Farsi* in Goodreads," in *Advances on Intelligent Informatics and Computing*, F. Saeed, F. Mohammed, and F. Ghaleb, Eds., Springer, Cham, pp. 219-230, 2022.
[15] K. E. Chu, P. Keikhosrokiani, and M. P. Asl, "A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts," *Pertanika Journal of Science & Technology*, vol. 30, no. 4, pp. 2535-2561, 2022.
[16]E. F. B. K. Fasha, P. Keikhosrokiani, and M. P. Asl, "Opinion mining using sentiment analysis: a case study of readers' response on Long Litt Woon's *The Way Through the Woods* in Goodreads," *Advances on Intelligent Informatics and Computing*, Springer, Cham, 2022.
[17] N. N. Jafery, P. Keikhosrokiani, and M. P. Asl, "An artificial intelligence application of theme and space in life writings of Middle

Eastern women: A topic modelling and sentiment analysis approach," in *Handbook of Research on Artificial Intelligence Applications in Literary Works and Social Media*, IGI Global, pp. 19-35, 2023.
[18] S. A. John and P. Keikhosrokiani, "COVID-19 fake news analytics from social media using topic modeling and clustering," in *Big Data Analytics for Healthcare: Datasets, Techniques, Life Cycles, Management, and Applications*, Academic Press, pp. 221-232, 2022.
[19] M. H. Al Mamun, P. Keikhosrokiani, M. P. Asl, N. A. Anuar, N. H. Hadi, and T. Humida, "Sentiment analysis of the Harry Potter series using a lexicon-based approach," in *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media*, P. Keikhosrokiani and M. P. Asl, Eds., IGI Global, pp. 263-291, 2022.
[20] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: a survey," *Information*, vol. 10, no. 4, pp. 150, 2019.
[21] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: a comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1-40, 2021.
[22] J. R. Jim, M. A. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 29, pp. 100059, 2024.
[23] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: convolutional LSTM model for sentiment analysis in social big data," *Information Processing & Management*, vol. 58, no. 1, pp. 102435, 2021.
[24] N. H. Suhendra, P. Keikhosrokiani, M. P. Asl, and X. Zhao, "Opinion mining and text analytics of literary reader responses: a case study of reader responses to KL Noir volumes in Goodreads using sentiment analysis and topic modeling," in *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media*, P. Keikhosrokiani and M. P. Asl, Eds., IGI Global, pp. 191-239, 2022.
[25] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 113-126, 2022.
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017. arXiv preprint arXiv:1810.04805, 2018.
[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[29]T. Brown, B. Mann, N. Ryder et al,"Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
[30]C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020.
[31]O. Araque, L. Gatti, J. Staiano, and M. Guerini, "Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 49-60, 2020.
[32]W. X. Zhao, K. Wang, S. Yan, J. Wang, and X. Li, "Weakly-supervised deep embedding for product review sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 1, pp. 185-197, 2020.
[33]F. Li, Z. Zhou, X. Li, and F. Zhou, "Deep learning with differential privacy," *arXiv preprint arXiv:2105.05526*, 2021.
[34]M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 17283-17297, 2020.
[35]S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342-8360, 2020.
[36]C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, 2019.
[37]E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610-623, 2021.
[38] T. Susnjak, "Applying BERT and ChatGPT for sentiment analysis of Lyme disease in scientific literature," in *Borrelia burgdorferi: Methods and Protocols*, pp. 173-183, 2024.
[39] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M. Y. Kan, and Y. F. Tan, "The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics," 2008.
[40] T. Bienz, R. Cohn, and Adobe Systems, *Portable Document Format Reference Manual*, Boston, MA: Addison-Wesley, p. 214, 1993.
[41]D. Yu and B. Hua, "Sentiment classification of scientific citation based on modified BERT attention by sentiment dictionary," in *Proc. of the EEKE/AII @ JCDL*, pp. 59-64,2023
[42] A. Cohan, W. Ammar, M. Van Zuylen, and F. Cady, "Structural scaffolds for citation intent classification in scientific publications," *arXiv preprint arXiv:1904.01608*, 2019.
[43]Y. A. Nanehkaran, J. Chen, S. Salimi, and D. Zhang, "A pragmatic convolutional bagging ensemble learning for recognition of Farsi handwritten digits," The Journal of Supercomputing, vol. 77, no. 11, pp. 13474-13493, 2021.
[44] H. Zhang, Y.-N. Cheah, O. M. Alyasiri, and J. An, "Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey," *Artificial Intelligence Review*, vol. 57, no. 2, p. 17, 2024.
[45]O. M. Al-Janabi, M. K. Ibrahim, A. Kanaan-Jebna, O. M. Alyasiri, and H. J. Aleqabie, "An improved Bi-LSTM performance using Dt-WE for implicit aspect extraction," in *2022 International Conference on Data Science and Intelligent Computing (ICDSIC)*, pp. 14-19, IEEE, 2022.
[46] Y. A. Nanehkaran, D. Zhang, S. Salimi, J. Chen, Y. Tian, and N. Al-Nabhan, "Analysis and comparison of machine learning classifiers and deep neural networks techniques for recognition of Farsi handwritten digits," *The Journal of Supercomputing*, vol. 77, pp. 3193-3222, 2021.

ACKNOWLEDGMENT

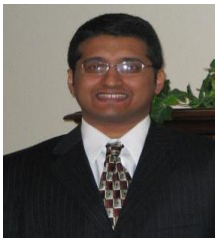
MUHAMMAD INAAM UL HAQ received a B.S. degree in computer science from the National University of Computer and Emerging Sciences (FAST-NUCES) in 2010 and an M.Phil. degree from Lahore Leads University, Pakistan, in 2014. He is pursuing a Ph.D. in computer science with the School of Computer Science and Engineering, Nanjing University of Science and Technology (NJUST), China. His research interests include data mining, text mining, NLP, and machine and deep learning.

KHALID MAHMOOD (Senior Member, IEEE, ACM Professional Member) received a Ph.D. degree in computer science from the International Islamic University, Islamabad, Pakistan, in 2018. He is currently working as an Assistant Professor at the Future Technology Research Center, National Yunlin University of Science and Technology, Yunlin, Taiwan. Earlier, he also served as a faculty member with COMSATS University Islamabad, which is ranked # 1 in the IT category. He is an approved supervisor by the Higher Education Commission of Pakistan. He is the founder of Network Security Research Group (NSRG) since 2017. His research interests include the design and development of lightweight authenticated and key agreement solutions for diverse infrastructures like smart grid, the Internet of Drones (IoD), the Internet of Things (IoT), vehicular ad hoc networks (VANET),

mobile edge computing, and blockchain. In 2017, considering his research, the Pakistan Council for Science and Technology granted him the Prestigious Young Productive Scientist Award, while affirming him among the Top Productive Computer Scientist in Pakistan.

QIANMU LI (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from the Nanjing University of Science and Technology, China, in 2001 and 2005, respectively. He is a Full Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology. He is the author or the coauthor of over 100 high-indexed (SCIE/E-SCI/EI) journal/conference articles and eight books. His research interests include information security, computing system management, and data mining. He received the China Network and Information Security Outstanding Talent Award in 2016 and multiple Education Ministry Science and Technology Awards in 2012.

ASHOK KUMAR DAS (Senior Member, IEEE) received a Ph.D. degree in computer science and engineering, an M.Tech. degree in computer science and data processing, and an M.Sc. degree in mathematics from IIT Kharagpur, India. He is currently a full Professor with the Center for Security, Theory and Algorithmic Research, IIIT, Hyderabad, India. He is an adjunct professor with the Korea University, Seoul, South Korea. He was also a visiting research professor with the Virginia Modeling, Analysis and Simulation Center, Old Dominion University, Suffolk, VA 23435, USA. His current research interests include cryptography, system and network security including security in smart grid, Internet of Things (IoT), Internet of Drones (IoD), Internet of Vehicles (IoV), Cyber-Physical Systems (CPS) and cloud computing, intrusion detection, blockchain, AI/ML security, and post-quantum cryptography. He has authored over 450 papers in international journals and conferences in the above areas, including over 385 reputed journal papers. He was a recipient of the Institute Silver Medal from IIT Kharagpur. He has been listed in the Web of Science (Clarivate™) Highly Cited Researcher 2022 and 2023 in recognition of his exceptional research performance. He was/is on the editorial board of IEEE Transactions on Information Forensics and Security, IEEE Systems Journal, Journal of Network and Computer Applications (Elsevier), Computer Communications (Elsevier), Journal of Cloud Computing (Springer), Cyber Security and Applications (Elsevier), IET Communications, KSII Transactions on Internet and Information Systems, and International Journal of Internet Technology and Secured Transactions (Inderscience). He also served as one of the Technical Program Committee Chairs of the first International Congress on Blockchain and Applications (BLOCKCHAIN'19), Avila, Spain, June 2019, International Conference on Applied Soft Computing and Communication Networks (ACN'20), October 2020, Chennai, India, second International Congress on Blockchain and Applications (BLOCKCHAIN'20), L'Aquila, Italy, October 2020, and International Conference on Applied Soft Computing and Communication Networks (ACN'23), December 2023, Bangalore, India. His Google Scholar h-index is 90 and i10-index is 283 with over 23,900 citations.



SACHIN SHETTY (Senior Member, IEEE) received a Ph.D. degree in modeling and simulation from Old Dominion University, in 2007. He was an Associate Professor at the Electrical and Computer Engineering Department, Tennessee State University, USA. He is currently a Professor with the Virginia Modeling, Analysis and Simulation Center, at Old Dominion University. He holds a joint appointment with the Center for Cybersecurity Education and Research and the Department of Modeling, Simulation, and Visualization Engineering. He has authored or co-authored more than 200 research papers in journals and conference proceedings and two books. His current research interests include the intersection of computer networking, network security, and machine learning.

MAJID HUSSAIN is serving as Dean of, the Faculty of Information Technology and Professor of the Computer Science department at The University of Faisalabad. His research interests include Artificial Intelligence, Blockchain, UAVs, Visual Sensor Networks/Image Processing, and Wireless Sensor Networks, and published his work in well-reputed international Journals. He has published a good number of research articles at national and international conferences and journals. He has won around 10 national and international research projects with considerable funding volume. He has acquired multiple international/national training and certifications at CISCO, Huawei, etc. He has earned multiple senior management level national and international training. As a Senior Member of IEEE, he has hosted many summits, workshops, seminars, and trainings for faculty and research students. His research interests include Blockchain, Artificial Intelligence, UAVs, and Visual Sensor Networks.