IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 July 2024, accepted 27 July 2024, date of publication 9 August 2024, date of current version 28 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441310

RESEARCH ARTICLE

Climate Change Sentiment Analysis Using Domain Specific Bidirectional Encoder Representations From Transformers

V. S. ANOOP[1], T. K. AJAY KRISHNAN[1], ALI DAUD[2], AMEEN BANJAR[3], AND AMAL BUKHARI[3]

[1]NLP for Social Good Laboratory, School of Digital Sciences, Kerala University of Digital Sciences, Innovation and Technology, Thiruvananthapuram 695317, India
[2]Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates
[3]Department of Information Systems and Technology, College of Computer Science and Engineering (CCSE), University of Jeddah, Jeddah 21959, Saudi Arabia

Corresponding author: Ali Daud (alimsdb@gmail.com)

ABSTRACT Climate change's impact on human health poses unprecedented and diverse challenges. Unless proactive measures based on solid evidence are implemented, these threats will likely escalate and continue to endanger human well-being. The escalating advancements in information and communication technologies have facilitated the widespread availability and utilization of social media platforms. Individuals utilize platforms such as Twitter and Facebook to express their opinions, thoughts, and critiques on diverse subjects, encompassing the pressing issue of climate change. The proliferation of climate change-related content on social media necessitates comprehensive analysis to glean meaningful insights. This paper employs natural language processing (NLP) techniques to analyze climate change discourse and quantify the sentiment of climate change-related tweets. We collected a total number of 5506 tweets for the period of January 2022 and February 2023 and manually labeled them to make the dataset for this experiment. ClimateBERT, a pre-trained model fine-tuned specifically on the climate change domain was used to generate the context vectors. Several machine learning algorithms with different feature encoding techniques, such as TF-IDF and BERT, have been implemented to classify user sentiments. When comparing the performance of the classifiers using different evaluation metrics such as precision, recall, accuracy, and f-measure, the ClimateBERT + Random Forest model is found to be outperforming all the other baselines with an accuracy of 90.22%, recall of 85.22%, and an f-measure of 85.47%. The findings from this experiment unearth valuable insights into public sentiment and the entities associated with climate change discourse. Policymakers, researchers, and organizations can leverage such analyses to understand public perceptions, identify influential actors, and devise informed strategies to address climate change challenges.

INDEX TERMS Climate change, sentiment analysis, climateBERT, public discourse, natural language processing, sentiment analysis.

I. INTRODUCTION

According to the World Health Organization (WHO), the greatest health threat to people in the twenty-first century is climate change. Between 2030 and 2050, this risk is expected to cause an additional 250,000 fatalities annually and express itself in various ways.[1] Many of these health concerns can be decreased or avoided with prompt and effective adaptation, but doing so necessitates in-depth research and policies that are multi-sectoral, multi-system, and collaborative at several scales [18]. The scientific communities may agree that human activity is accelerating the effect of climate change, which is

The associate editor coordinating the review of this manuscript and approving it for publication was Asadullah Shaikh.

[1]https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

disastrous for the world and its population. The consequences of climate change are clearer now, with extreme weather conditions like hurricanes, tornadoes, hail, lightning, fires, and floods [42]. As the world's ecosystems change quickly, access to the natural resources and agricultural methods that support humanity is in danger [10]. The problem of climate change is complicated, and there is no quick fix. But it's critical to comprehend the issue using technological advancements to identify solutions.

Large volumes of text data can be analyzed using natural language processing, which may unearth interesting patterns [6], [28]. The natural language processing approaches can also be applied to the climate change domain to find causes and leverage patterns such as public sentiment and discourse towards this global issue [16], [24]. Recent years have witnessed many people using social media to share their views, concerns, and public opinions on any topic under the sky [9], [21], [22]. This has caused a huge amount of unstructured but dynamic data to be generated in such platforms, which are goldmines for social science researchers [8], [23]. Collecting, curating, and analyzing such data is crucial for finding public perceptions and viewpoints on socially relevant discussions [32], [45]. Similarly, understanding public opinions and sentiments on climate change is crucial for policymakers, governments, and other administrators to devise better policies and intervention measures to address the challenges [7], [16].

In this research, natural language processing is used to examine tweets that discuss climate change. We use ClimateBERT [47], a pre-trained language model trained on a large set of climate change-related documents, and fine-tune the same for sentiment classification tasks. The findings may be used better to understand the public's understanding of climate change, and it can also be used to identify the key stakeholders in the climate change debate. The interesting insights from this project will provide a foundation for informed decision-making and policy formulation regarding climate change. Additionally, the findings will contribute to advancing NLP techniques and their application in climate change analysis.

A. EFFECTS OF CLIMATE CHANGE

Climate change is already manifesting globally, with a notable increase in extreme weather events. The frequency and intensity of hurricanes, floods, and droughts have amplified, causing widespread destruction and loss of life [12]. Coastal areas are also seriously threatened by increasing sea levels, which might result in massive population displacement, increased erosion, and flooding. As glaciers continue to melt, water supplies diminish, affecting regions dependent on glacial meltwater for agricultural, industrial, and domestic purposes [36]. Climate change-induced shifts in environmental conditions are causing profound changes in plant and animal life. Species are forced to adapt or face extinction as they grapple with altered ecosystems and changing habitats. This disruption to biodiversity has cascading effects on ecosystem functioning and services, with implications for food security, ecosystem stability, and human well-being [38].

Another consequence of climate change is the heightened risk of diseases [36]. As temperatures rise, disease-carrying organisms like mosquitoes expand their geographic range, exposing previously unaffected regions to vector-borne illnesses. This poses a significant public health challenge, necessitating the development of effective strategies for disease prevention, control, and surveillance [38]. The impacts of climate change extend beyond the natural environment, affecting societies and economies globally. Disruptions to ecosystems and weather patterns have severe social and economic repercussions, disproportionately affecting vulnerable communities [35]. Climate-induced events, such as extreme heatwaves, prolonged droughts, and intense storms, lead to the displacement of populations, loss of livelihoods, and increased socioeconomic inequality. Consequently, countries face significant challenges in managing climate change's economic and social ramifications, including the need for adaptation measures and the transition to sustainable practices [40]. Addressing climate change requires concerted global efforts to mitigate greenhouse gas emissions, enhance resilience, and foster sustainable development practices to safeguard the future of our planet and its inhabitants [2].

B. NLP FOR CLIMATE CHANGE ANALYSIS

Natural language processing (NLP) uses algorithms and models to enable computers to comprehend, analyze, and produce human language [28]. In climate change analysis, NLP techniques have proven invaluable in extracting meaningful insights from vast amounts of textual data [28]. One application of NLP in climate change analysis is the ability to analyze public opinion on the topic. By leveraging sentiment analysis techniques, researchers can gauge the prevailing sentiments, attitudes, and beliefs surrounding climate change [13]. This understanding of public opinion is crucial for policymakers, as it helps them tailor communication strategies, design effective interventions, and foster public engagement in addressing climate change challenges [15].

Furthermore, NLP techniques allow for identifying key stakeholders in the climate change debate. Through the extraction and examination of written information from various outlets, including news articles, social media platforms, and scientific journals, scholars can discern the key individuals, organizations, and institutions influencing the conversation surrounding climate change [14], [44]. This knowledge provides valuable insights into the various perspectives, interests, and motivations shaping climate change discussions, facilitating informed decision-making and targeted engagement with relevant stakeholders [37], [48]. Textual data analysis from scientific reports, environmental assessments, and socio-economic surveys would be instrumental for researchers to gain insights into the specific vulnerabilities, risks, and adaptation strategies associated with climate change in different geographic areas [30]. This information is crucial for policymakers and local communities to prioritize resources, implement targeted interventions, and build resilience against the impacts of climate change [31].

The major contributions of this research may be summarized as follows:

- A detailed study on related literature and approaches reported on public discourse and sentiment analysis using social media data.
- Highlights the potential of ClimateBERT - a pre-trained model on climate data and proposes an approach for fine-tuning ClimateBERT for the sentiment analysis of climate change tweets.
- Extensive experimental comparison of different machine learning approaches implemented and a detailed discussion of the results.
- Publishes the labeled dataset used in this work for other researchers and developers to build advanced models for climate change analysis.

The remaining sections of this paper are organized as follows: Section II discusses some of the recent and prominent approaches related to the area of study, and Section III presents the materials and methods used in this paper. In Section IV, the proposed approach for fine-tuning ClimateBERT for sentiment analysis is presented, followed by the results and discussion in Section V. Section VI discusses the conclusions and future research dimensions.

II. RELATED STUDIES

This section discusses some state-of-the-art approaches in natural language processing and machine learning for climate change analysis. It also discusses relevant studies exploring sentiment text classification approaches pertinent to the proposed method. An approach for examining the polarization and belief systems prevalent in climate change discussions on Twitter [11] was reported [39]. The authors have proposed a framework to identify statements denying climate change and classify tweets into denier or believer stances. The framework focused on two interconnected tasks: stance detection and sentiment analysis. Experimental results demonstrated that the proposed framework enhances stance detection accuracy by leveraging sentiment analysis, outperforming uni-modal and single-task approaches. Lydiri et al. proposed another approach that uses Bidirectional Encoder Representation from Transformers (BERT) and Convolutional Neural Networks (CNN) to analyze public opinions on climate change by examining Twitter data [34]. The results indicated that the proposed model surpasses conventional machine learning methods, accurately identifying climate change believers and deniers. The authors suggested that this model has significant potential for monitoring and governance purposes, particularly in smart city contexts.

Recently, an approach to analyze extensive unstructured data concerning climate change was reported by Ceylan [15]. This study aims to develop an information management system capable of extracting pertinent information from diverse data sources, particularly technical design documentation. By utilizing pre-trained natural language processing models trained on textual data and integrating non-textual graphical data, the researchers showcased the system's effectiveness in swiftly and efficiently retrieving precise information. Sentiment analysis deals with analyzing user-generated text to understand the writer's emotions. Several works are reported in this area with varying degrees of success, but recent advancements in pretrained language models recorded state-of-the-art results [19], [27]. There have been many works reported in the natural language processing and machine learning literature in recent years on the advancements of sentiment analysis for different domains, such as conflicts between countries [41], pandemics [8], [9], and health [6]. Large language models are also being used for developing better sentiment-capturing algorithms to deal with complex unstructured text [1], [4], [17].

The pre-trained language model landscape in natural language processing is ever-evolving, and recent years witnessed the Transformer-based models gaining popularity. Researchers are fine-tuning Transformer-based models such as Bidirectional Encoder Representation from Transformers (BERT) for many domains, such as mental health, finance, and biomedicine.

There are some very recent approaches reported in the literature on using advanced machine learning and deep learning techniques [5], [26] for effective analysis of user sentiments for domains such as education [25] and government [3]. A novel hybrid approach for sentiment analysis was proposed very recently by Islam et al. [20]. The authors have proposed a capsule-based approach with CNN and R.N.N. for better sentiment classification and showed that their proposed methodology could score state-of-the-art results for different datasets. A multi-modal sentiment analysis approach that uses a coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks was reported very recently in the literature [33]. Extensive experimental results on different datasets showed that their approach could score an accuracy of 86.5% and 86.1%, and best F1 of 86.4% and 86.1%. From the literature review, it is evident that sentiment analysis plays a crucial role in understanding the discourse and sentiment of the public towards different topics, which is a well-studied area. So, it is highly relevant to understand the perspectives of the public towards climate change, which is essential for framing regulations and other policies for the benefit of society, as discussed in this paper.

A. ClimateBERT - A PRETRAINED LANGUAGE MODEL FOR CLIMATE-RELATED TEXT

CimateBERT, a transformer-based language model pre-trained on a corpus of climate-related paragraphs extracted from diverse sources such as news articles, research papers, and corporate disclosures, was reported [47]. This aims to overcome the limitations of general language models in effectively representing climate-related texts. The improved performance of CimateBERT contributes to lower error rates in various climate-related downstream tasks. To encourage further research at the intersection of climate change and natural language processing, the authors provided public access to the training code and weights of CLIMATEBERT. Another approach, ClimaText dataset [46], was specifically developed to detect sentence-level climate change topics within textual sources. The authors emphasized the significance of automating the extraction of climate change information from media and other text-based materials to facilitate various applications, including content filtering, sentiment analysis, and fact-checking. Through a comparative study of different approaches for identifying climate change topics, they find that context-based algorithms like BERT outperform simple keyword-based models. However, the authors also identify areas that require improvement, particularly in capturing the discussion surrounding the indirect effects of climate change. The authors anticipated this dataset will be a valuable resource for further research in natural language understanding and climate change communication. Upadhyaya et al., in their approach, underscored the importance of comprehending public perception and acceptance of climate change policies [43]. The study examined diverse data sources, such as social media, scientific papers, and news articles, to perform sentiment analysis. The paper concluded that supervised machine learning techniques exhibit effectiveness in sentiment analysis, highlighting that ensemble and hybrid approaches yield superior outcomes compared to individual classifiers. Fig. 1 shows a basic architecture of the fine-tuned ClimateBERT model used in this study for sentiment classification.

B. RESEARCH GAP

The literature analysis confirms that limited approaches are reported in using natural language processing and machine learning techniques in addressing climate change. A need exists to develop better text-understanding algorithms and models to enhance the quality of existing approaches. In this connection, this proposed work utilizes the ClimateBERT model. It fine-tunes the same for analyzing public discourse and sentiments toward climate change using user-generated data collected from Twitter. There should be a collective effort from the scientific communities to tackle the climate change challenges. Keeping this in mind, the authors of this work are making the labeled dataset publicly available for other researchers to build advanced models.

III. MATERIALS AND METHODS

A. LABEL STUDIO

Label Studio is an open-source data annotation tool (available at https://labelstud.io/) that provides a user-friendly interface for creating labeled datasets by annotating data for machine learning and artificial intelligence tasks. The tool supports various annotation types, including text classification, NER, object detection, image segmentation, and more. Label Studio allows users to import data from multiple sources, such as CSV files, JSON, or databases, and annotate them using a customizable interface. It provides a collaborative environment where various annotators can collaborate on a project, with features like task assignment, annotation review, and inter-annotator agreement measurement. One of the critical features of Label Studio is its extensibility. It provides a flexible architecture that allows users to customize the annotation interfaces and incorporate custom labeling functions using JavaScript and Python. This enables the tool to adapt to annotation requirements and integrate with machine-learning workflows. Label Studio also supports active learning, where the tool can suggest samples to be annotated based on a model's uncertainty, helping to optimize the annotation process and improve model performance.

B. SNSCRAPE

snscrape is a Python library and command-line tool (available at https://github.com/JustAnotherArchivist/snscrape) for scraping social media content. It lets you retrieve public data from various social media platforms, including Twitter, Instagram, YouTube, Reddit, etc. With snscrape, you can fetch posts, comments, likes, followers, and other relevant information from social media platforms. It provides a flexible and customizable way to search for specific keywords, hashtags, usernames, or URLs and extract the desired content. The library supports scraping recent and historical data from social media platforms, enabling you to gather insights, perform analysis, monitor trends, and conduct research based on social media content. snscrape offers a command-line interface that allows you to search for and scrape social media data interactively. To customize your scraping process, you can specify various parameters, such as the number of results, date range, and output format. In addition to the command-line interface, snscrape provides a Python API that allows you to integrate social media scraping into your own Python scripts and applications. The API offers more advanced functionalities, giving you fine-grained control over the scraping process and allowing you to process the scraped data programmatically. One of the key advantages of snscrape is its ability to work with multiple social media platforms, providing a unified interface for scraping different types of content. It handles the intricacies of each platform's APIs and HTML structures, making it easier for developers to extract data without needing to learn the specific details of each platform. It's important to note that snscrape respects the terms of service and usage restrictions of each social media platform. It is primarily intended for scraping publicly available content and should be used responsibly and in compliance with the platform's policies.

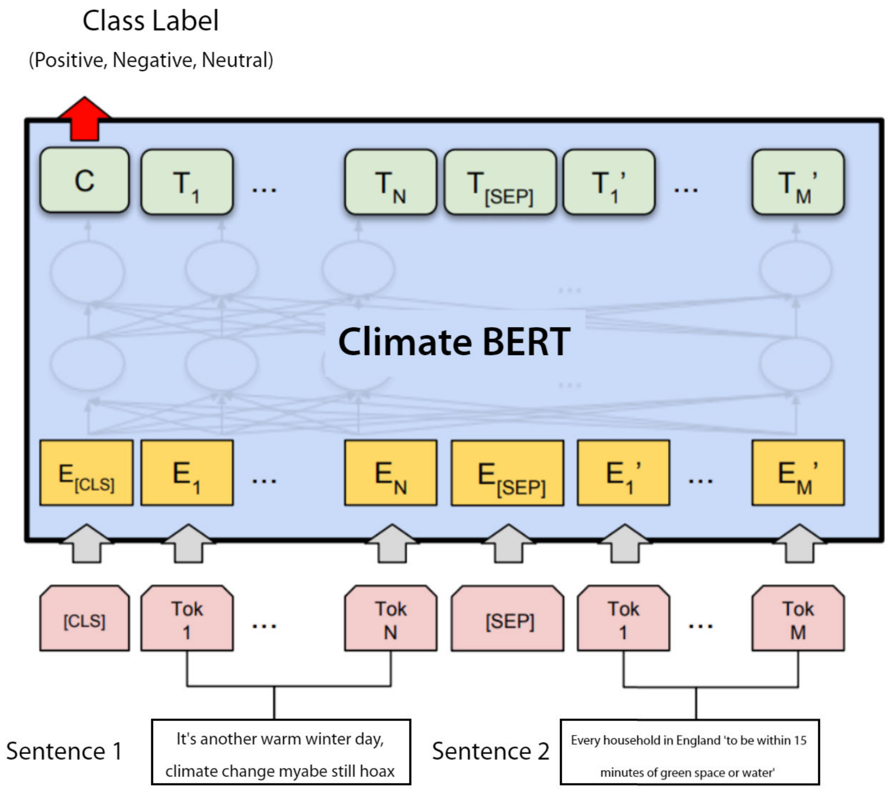


FIGURE 1. A basic architecture of the fine-tuned ClimateBERT model used in this study for sentiment classification.

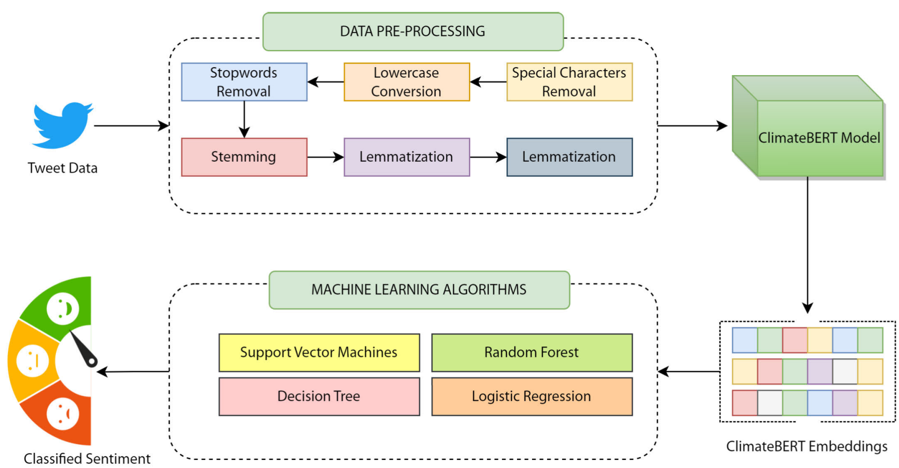


FIGURE 2. Overall workflow of the sentiment analysis on climate change data.

C. NEWSPAPER 3K

Newspaper3k is a Python library and web scraping tool (available at https://newspaper.readthedocs.io/) that allows you to extract and parse information from online news articles. It provides a simple interface to automate the fetching and processing news articles from various online sources. With Newspaper3k, you can retrieve article metadata from news websites, such as the title, author, publish date, and article text. It also supports extracting additional information like keywords, summaries, and article images. The library uses advanced NLP techniques to extract relevant information from the HTML structure of the news articles. Newspaper3k is designed to handle various complexities of news websites, including different article formats, pagination, and content extraction. It has built-in functionality to handle newspaper-specific features like multi-page articles, article pagination, and RSS feeds. One of the advantages of Newspaper3k is its ease of use. It abstracts away the complexities of web scraping and provides a clean and intuitive API. It also handles various encoding and parsing issues that often arise when dealing with news articles from different sources. Newspaper3k is widely used for multiple applications, including content analysis, sentiment analysis, and data mining. It offers a convenient way to gather news data for research, data analysis, and machine learning projects.

D. ClimateBERT

ClimateBERT is a specialized variant of the BERT model specifically trained and tailored for addressing climate change-related language tasks [47]. Building upon the foundation of BERT, ClimateBERT is pre-trained on a large corpus of climate change-related documents and text sources, enabling it to capture the nuances and domain-specific knowledge relevant to climate science. This fine-tuning process equips ClimateBERT with a deep understanding of climate-related concepts, terminology, and contextual dependencies. Integrating domain-specific knowledge into the pre-training process makes ClimateBERT a powerful tool for extracting insights, identifying patterns, and extracting valuable information from climate-related text data. Its application in climate change analysis can aid in improving decision-making, facilitating research, and enhancing our understanding of the complex challenges climate change poses.

The ClimateBERT model was trained on a large set of climate-related documents from three categories: news articles, research abstracts, and corporate climate reports. For the domain-specific adaptive pretraining, the batch size was set to 2016, the learning rate to 5e-4, and the number of epochs to 12. The Adam optimizer was used with a weight decay of 0.01. The proposed sentiment analysis approach discussed in this paper used the default ClimateBERT model with default parameters and then fine-tuned the same for the sentiment analysis task using the dataset prepared. The basic architecture of the ClimateBERT model used in this approach is shown in Fig. 1.

IV. PROPOSED APPROACH

This section deals with the proposed methodology for the sentiment analysis of climate-related tweets from Twitter using ClimateBERT embeddings and Random Forest Classifier. The overall workflow of the proposed approach is given in Fig. 2.

A. DATASET

The methodology begins with data collection from Twitter using the *snscrape* library available at https://github.com/JustAnotherArchivist/snscrape. We have used the keyword and hashtag-based Twitter search for selecting the tweets related to climate change. Some of the keywords used are *#climatechange*, *#climatecrisis*, *#climateaction*, *#climatechangeisreal*, etc. These keywords and hashtags are identified after a careful evaluation of the tweets already reported

Algorithm 1 Algorithm for climate change sentiment analysis

Input: Publicly available Tweets on climate change
Output: Classified sentiment - Pos, Neg, or Neu
for `tweets` t **do**
  $L_t \leftarrow$ Assign_sentiment_label(t)
end for
for `Labeled tweets` L_t **do**
  Remove_special_characters(L_t)
  Remove_stopwords(L_t)
  Stemming(L_t)
  Lemmatization(L_t)
  Senti_Embed ←Generate_ClimateBERT_embeddings(L_t)
end for
while *Senti_Embed* $\neq$ *NULL* **do**
  $C_1 \leftarrow$ SVM(*Senti_Embed*)
  $C_2 \leftarrow$ Random_Forest(*Senti_Embed*)
  $C_3 \leftarrow$ Decision_Tree(*Senti_Embed*)
  $C_4 \leftarrow$ Logistic_Regression(*Senti_Embed*)
end while
$S_L \leftarrow$ Generate_Sentiment_Label()
Return S_L

and also after discussing with the subject matter experts. Inter-annotator agreement quantifies the degree to which annotators of a dataset agree on their ratings. It is essential to ensure that the annotation system is consistent and that multiple raters can give the same emotion label to the same comment. We have employed three human annotators for this work and utilize Krippendorff's alpha coefficient [29] because it applies to any number of annotators. We utilized the NLTK (nltk.metrics.agreement module) and obtained Krippendorff's alpha values of 0.72 for our dataset. The tweets were gathered between January 2022 and February 2023, and the collected data contain class imbalance, where certain sentiment categories are over-represented while others are underrepresented. As this could potentially bias the model's predictions, the initial data points (4410 tweets) have been processed for data augmentation to reduce the class imbalance. The collected data was labeled using the Label Studio tool available at https://labelstud.io/. For the labeling, domain experts and research scholars from ecological informatics and environmental studies have been employed, and they manually annotated the samples assigned to them. To ensure the credibility of the labeled data, some of the senior research staff members and faculties from the same domain have manually verified the data points and made necessary changes. The labeled data is then merged and loaded into a pandas DataFrame for further processing and to create a dataset for sentiment analysis. After the process, the final dataset consisted of 5506 tweets, with three labels - positive, negative, and neutral. The dataset and the code to process the same are available at https://github.com/anoop-vs/nlp-climate-change

TABLE 1. A snapshot of the labeled dataset used for the experiment.

Content	Labels
Researchers use deep learning to simulate chlorophylla & phycocyanin with an internet of things system to detect & quantify #cyanobacteria, to improve #eutrophication management schemes for freshwater reservoirs. #algae #microbiology #environment #iot eeer.org/journal/view.pâ€¦	Positive
Why is our @Conservatives government so evil? #RishiSunak #climateChange #Conservatives #FuckingThieves https://t.co/cCGyylmYlf	Negative
Sierra snowpack 205% of its historical average ǀ Climate Change ... - San Francisco Examiner dlvr.it/ShpGVN #ClimateChange	Neutral

to encourage other researchers to train their advanced natural language processing models. Table 1 shows a snapshot of the labeled dataset containing the tweet and the corresponding labels for each tweet.

B. EXPERIMENT

The proposed approach uses several pre-processing steps to prepare the experiment-ready version of the dataset. This has been done as we are using user-generated content from social media, which may contain noises that may affect the output of the trained model. Firstly, the special characters and unwanted digits are removed, and the text is converted to lowercase to ensure consistency using the Python regular expression module. Then, tokenization is performed, which involves splitting the words into individual units. Stopwords (common words with little contextual meaning) are removed, and stemming and lemmatization techniques are performed to normalize the words. All these processes were carried out using the Natural Language Toolkit (NLTK) library with Python. This pre-processing step ensured the text data was cleaned and ready for further analysis. The oversampled dataset has been split into a train-test split with 80% for training and 20% for testing. The epoch and batch size have been set to 4 and 16 to train the models, respectively. To encode the features for the tweets, this work uses ClimateBERT [47], a domain-specific pre-trained model for the climate change domain. The ClimateBERT embeddings were generated using *distilroberta-base-climate-f* pre-trained model publicly available at https://huggingface.co/climatebert/distilroberta-base-climate-f.

All the experiments were executed on an NVIDIA A100 machine with 80 GB GPU memory and 1935 GB/Second bandwidth. All scripts were written in Python 3.9, and the machine learning algorithms have been implemented using the scikit-learn library available at https://scikit learn.org/stable/. The proposed approach trains the model using four machine learning algorithms, such as random forest, support vector machines, decision trees, and logistic regression. We have specifically used these algorithms to construct the baseline for the initial work on the sentiment analysis and also to quantify the performance of different text encoding techniques such as TF-IDF and BERT. As the proposed approach was one of the earliest approaches in using ClimateBERT for the qualitative analysis of the discourse and sentiment of the general public towards climate change, these easy-to-train machine learning models have been employed. As the results are promising, the proposed approach may be implemented using deep learning-based approaches in future work. The algorithm for sentiment analysis of climate change data is given in Algorithm 1.

V. RESULTS AND DISCUSSIONS

This section presents the results obtained from the experiment using the proposed approach outlined in Section IV. The results and a detailed discussion are given in this section.

TABLE 2. Precision, Recall, Accuracy, and F-Measure values for the TF-IDF feature encoding.

Model	TF-IDF			
	Accuracy	Precision	Recall	F-measure
Random Forest	87.11	86.98	87.11	87.01
Support Vector Machine	84.39	84.34	84.39	84.14
Decision Tree	73.23	72.63	73.23	72.71
Logistic Regression	90.10	90.13	90.10	90.04

TABLE 3. Precision, Recall, Accuracy, and F-Measure values for word2Vec.

Model	word2Vec			
	Accuracy	Precision	Recall	F-measure
Randrom Forest	72.59	72.91	72.59	72.69
Support Vector Machine	40.29	40.59	40.29	36.92
Decision Tree	62.34	61.24	62.34	61.59
Logistic Regression	39.56	40.25	39.56	36.53

TABLE 4. Precision, Recall, Accuracy, and F-Measure values for CountVectorizer.

Model	CountVectorizer			
	Accuracy	Precision	Recall	F-measure
Randrom Forest	89.56	89.53	89.56	89.50
Support Vector Machine	88.11	88.02	88.11	87.98
Decision Tree	79.67	79.50	79.67	79.58
Logistic Regression	79.21	79.09	79.21	78.76

The results obtained from the evaluation metrics are reported. Accuracy provides an overall measure of correctness, precision measures the proportion of correctly predicted positive sentiments, recall captures the ability to identify all positive sentiments, and the F1-score provides a balanced measure between precision and recall. These metrics provide insights into how well the model predicts sentiment on climate change-related tweets. Tables 2 to 8 show the

TABLE 5. Precision, Recall, Accuracy, and F-Measure values for TF-IDF + CountVectorizer.

Model	TF-IDF + CountVectorizer			
	Accuracy	Precision	Recall	F-measure
Random Forest	86.93	86.96	86.93	86.87
Support Vector Machine	81.94	81.92	81.94	81.75
Decision Tree	74.77	74.30	74.77	74.46
Logistic Regression	81.30	81.11	81.30	80.99

TABLE 6. Precision, Recall, Accuracy, and F-Measure values for TF-IDF + word2Vec.

Model	TF-IDF + word2Vec			
	Accuracy	Precision	Recall	F-measure
Randrom Forest	79.21	79.45	79.21	79.29
Support Vector Machine	85.39	85.34	85.39	85.22
Decision Tree	62.43	61.28	62.43	61.60
Logistic Regression	74.77	74.33	74.77	74.28

TABLE 7. Precision, Recall, Accuracy, and F-Measure values for CountVectorizer + word2Vec.

Model	CountVectorizer + word2Vec			
	Accuracy	Precision	Recall	F-measure
Random Forest	79.03	79.45	79.03	79.18
Support Vector Machine	81.30	81.10	81.30	81.09
Decision Tree	62.06	61.01	62.06	61.37
Logistic Regression	81.21	81.04	81.21	80.92

TABLE 8. Performance evaluation of SVM, LR, RF, and DT algorithms using BERT.

Model	BERT			
	Accuracy	Precision	Recall	F-measure
Random Forest	76.78	77.46	76.78	76.93
Support Vector Machine	64.35	63.65	64.35	63.70
Decision Tree	68.89	67.13	68.89	67.89
Logistic Regression	63.81	63.48	63.81	63.60

TABLE 9. Performance evaluation of SVM, LR, RF, Naive Bayes, and DT algorithms using ClimateBERT.

Model	ClimateBERT			
	Accuracy	Precision	Recall	F-measure
Random Forest	**90.22**	85.73	85.22	85.47
Support Vector Machine	88.66	76.20	75.66	75.92
Decision Tree	80.62	79.88	78.62	79.24
Logistic Regression	73.84	72.92	73.84	73.37

accuracy, precision, recall, and f-measure values for RF, SVM, DT, and LR algorithms. For BERT embeddings, RF has 76.78%, 77.46%, 76.78%, and 76.93%, SVM has 64.35%, 63.65%, 64.35%, and 63.70%, DT has 68.89%, 67.13%, 68.89%, and 67.59%, and LR has 63.81%, 63.48%, 63.81%, and 63.60% for the Accuracy (A), Precision (P), Recall (R), and F-measure values. For ClimateBERT embeddings (Table 9), the Random Forest classifier has recorded 90.22%, 85.73%, 85.22%, and 85.47%, Support Vector Machine has recorded 88.66%, 76.20%, 75.66%, and 75.92%, Decision Tree has got 80.62%, 79.88%, 78.62%, and 79.24%, and Logistic Regression has 73.84%, 72.92%, 73.84%, and 73.37% for the A, P, R, and F-measure, respectively. After training, the model's performance is evaluated on the test set to assess its ability to predict sentiments. Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8 shows the accuracy, precision, recall, and f-measure comparison for different text encoding techniques for the selected classifiers. The model is switched to evaluation mode, and predictions are made on the test set. The results show that ClimateBERT, a fine-tuned domain-specific language model, showcases better textual

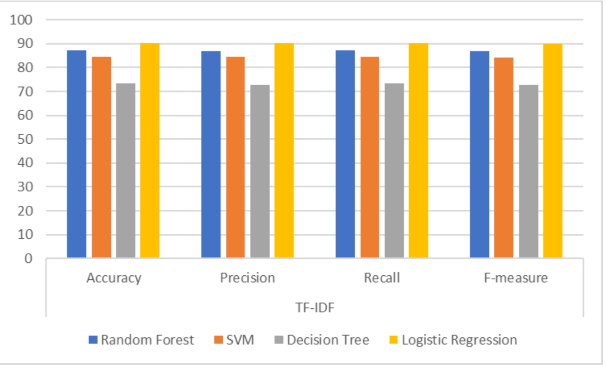


FIGURE 3. Accuracy, precision, recall, f-measure comparison for Support Vector Machine,Logistic Regression, Random Forest, and Decision Tree algorithms with TF-IDF feature encoding.

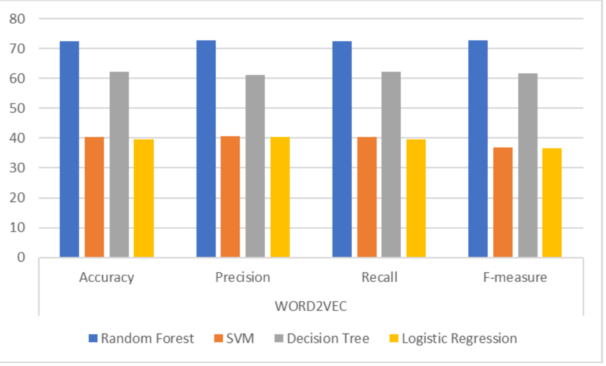


FIGURE 4. Accuracy, precision, recall, f-measure comparison for Support Vector Machine,Logistic Regression, Random Forest, and Decision Tree algorithms with word2Vec feature encoding.

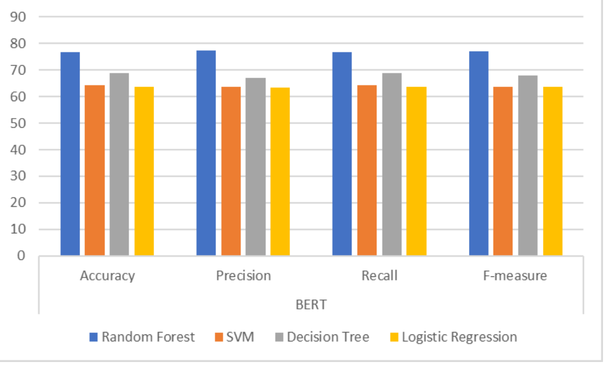


FIGURE 5. Accuracy, precision, recall, f-measure comparison for Support Vector Machine,Logistic Regression, Random Forest, and Decision Tree algorithms with BERT feature encoding.

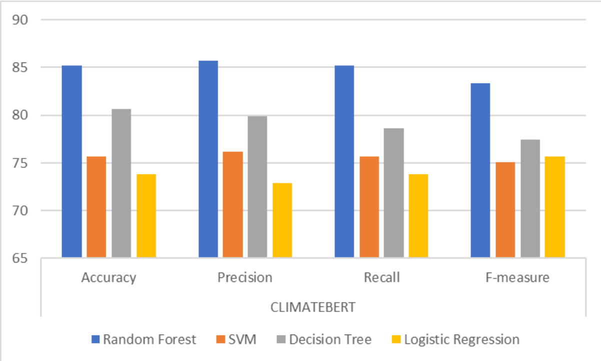


FIGURE 6. Accuracy, precision, recall, f-measure comparison for Support Vector Machine,Logistic Regression, Random Forest, and Decision Tree algorithms with ClimateBERT feature encoding.

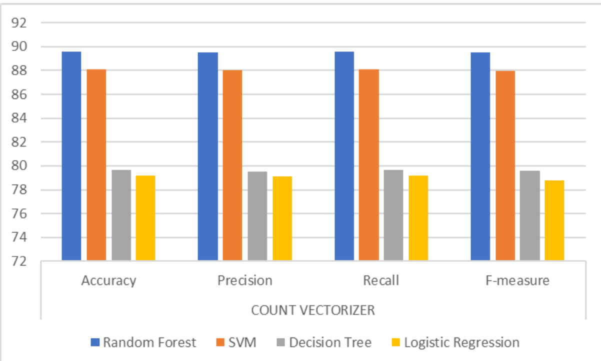


FIGURE 7. Accuracy, precision, recall, f-measure comparison for Support Vector Machine, Logistic Regression, Random Forest, and Decision Tree algorithms with count vector feature encoding.

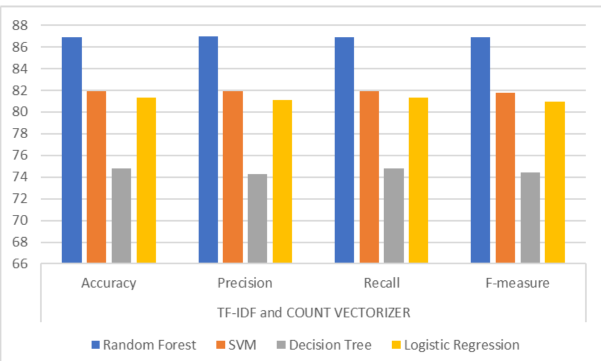


FIGURE 8. Accuracy, precision, recall, f-measure comparison for Support Vector Machine, Logistic Regression, Random Forest, and Decision Tree algorithms with TF-IDF+CV feature encodings.

semantics capturing for downstream applications such as text classification and sentiment analysis. The comparative performance in precision, recall, accuracy, and f-measure shows that ClimateBERT outperforms all the other feature encoding techniques used in this work and shows its potential to leverage better contextual representations that may aid in several natural language processing applications.

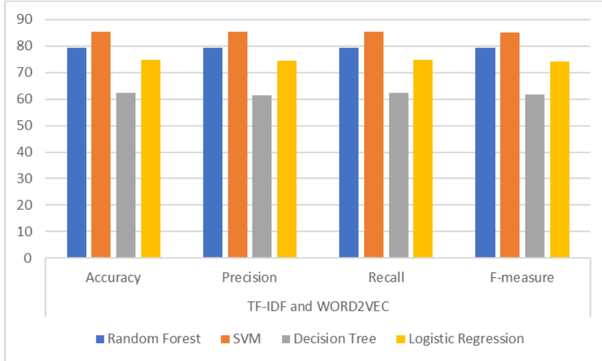


FIGURE 9. Accuracy, precision, recall, f-measure comparison for Support Vector Machine, Logistic Regression, Random Forest, and Decision Tree algorithms with TF-IDF+word2Vec feature encodings.

FIGURE 10. Accuracy, precision, recall, f-measure comparison for Support Vector Machine, Logistic Regression, Random Forest, and Decision Tree algorithms with CV+word2Vec feature encodings.

VI. CONCLUSION AND FUTURE WORK

This work reported the findings of a domain-specific pre-trained transformer-based approach for classifying the public discourse and sentiment towards a pressing global concern, climate change. Rapid developments and paradigm shifts happen with the natural language processing, specifically on pretrained models and large language models trained with the attention mechanism. These models could be used to tackle several global concerns such as healthcare, education, and climate. Comprehending climate change's complexities and nuances through textual data promises efficient document processing and policy formation with informed decisions. This paper utilized the advancements in domain-specific large language models to address the challenges posed by climate change through sentiment analysis. As the results are promising, this work may promise multiple future research diemsions. One immediate future work is to use advanced machine learning models such as deep learning in which sequential data handling algorithms such as recurrent neural networks are implemented to improve the results. The data used in this experiment was sourced from a social media platform and was very limited. Another dimension of future work could be to use extensive scale data for training advanced machine learning and natural language processing models.

REFERENCES

[1] T. Abdullah and A. Ahmet, "Deep learning in sentiment analysis: Recent architectures," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–37, Aug. 2023.

[2] S. Ali, A. Basit, T. A. Makanda, Inamullah, F. U. Khan, M. Sajid, T. Riaz, H. F. Abbasi, Manzoor, and A. Sohail, "Improving drought mitigation strategies and disaster risk reduction through MODIS and TRMM-based data in relation to climate change over Pakistan," *Environ. Sci. Pollut. Res.*, vol. 30, no. 14, pp. 40563–40575, Jan. 2023.

[3] O. Alqaryouti, N. Siyam, A. A. Monem, and K. Shaalan, "Aspect-based sentiment analysis using smart government review data," *Appl. Comput. Informat.*, vol. 20, nos. 1–2, pp. 142–161, Jan. 2024.

[4] A. Alsayat, "Improving sentiment analysis for social media applications using an ensemble deep learning language model," *Arabian J. Sci. Eng.*, vol. 47, no. 2, pp. 2499–2511, Feb. 2022.

[5] A. Alslaity and R. Orji, "Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions," *Behav. Inf. Technol.*, vol. 43, no. 1, pp. 139–164, Jan. 2024.

[6] V. Anoop, "Sentiment classification of diabetes-related tweets using transformer-based deep learning approach," in *Proc. Int. Conf. Adv. Comput. Data Sci.* Cham, Switzerland: Springer, 2023, pp. 203–214.

[7] V. S. Anoop and S. Asharaf, "Extracting conceptual relationships and inducing concept lattices from unstructured text," *J. Intell. Syst.*, vol. 28, no. 4, pp. 669–681, Sep. 2019.

[8] V. S. Anoop and S. Sreelakshmi, "Public discourse and sentiment during Mpox outbreak: An analysis using natural language processing," *Public Health*, vol. 218, pp. 114–120, May 2023.

[9] V. Anoop, J. Thekkiniath, and U. H. Govindarajan, "We chased COVID-19; did we forget measles?-Public discourse and sentiment analysis on spiking measles cases using natural language processing," in *Proc. Int. Conf. Multi-Disciplinary Trends Artif. Intell.* Cham, Switzerland: Springer, 2023, pp. 147–158.

[10] S. Ardabili, A. Mosavi, M. Dehghani, and A. R. Várkonyi-Kóczy, "Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review," in *Proc. Int. Conf. Global Res. Educ.* Cham, Switzerland: Springer, 2020, pp. 52–62.

[11] A. Banjar, Z. Ahmed, A. Daud, R. A. Abbasi, and H. Dawood, "Aspect-based sentiment analysis for polarity estimation of customer reviews on Twitter," *Comput., Mater. Continua*, vol. 67, no. 2, pp. 2203–2225, 2021.

[12] V. Bhola, A. Hertelendy, A. Hart, S. B. Adnan, and G. Ciottone, "Escalating costs of billion-dollar disasters in the U.S.: Climate change necessitates disaster risk reduction," *J. Climate Change Health*, vol. 10, Mar. 2023, Art. no. 100201.

[13] S. S. Biswas, "Potential use of chat GPT in global warming," *Ann. Biomed. Eng.*, vol. 51, no. 6, pp. 1126–1127, Jun. 2023.

[14] B. Brzic, I. Boticki, and M. Bagic Babac, "Detecting deception using natural language processing and machine learning in datasets on COVID-19 and climate change," *Algorithms*, vol. 16, no. 5, p. 221, Apr. 2023.

[15] C. Ceylan, "Application of natural language processing to unstructured data: A case study of climate change," Ph.D. thesis, Massachusetts Inst. Technol., 2022.

[16] K. Chen, A. L. Molder, Z. Duan, S. Boulianne, C. Eckart, P. Mallari, and D. Yang, "How climate movement actors and news media frame climate change and strike: Evidence from analyzing Twitter and news media discourse from 2018 to 2021," *Int. J. Press/Politics*, vol. 28, no. 2, pp. 384–413, Apr. 2023.

[17] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do LLMs know about financial markets? A case study on Reddit market sentiment analysis," in *Proc. Companion ACM Web Conf.*, Apr. 2023, pp. 107–110.

[18] B. J. Fard, S. A. Hasan, and J. E. Bell, "CliMedBERT: A pre-trained language model for climate and health-related text," 2022, *arXiv:2212.00689*.

[19] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood, "Towards deep learning prospects: Insights for social media analytics," *IEEE Access*, vol. 7, pp. 36958–36979, 2019.

[20] M. S. Islam, M. N. Kabir, N. A. Ghani, K. Z. Zamli, N. S. A. Zulkifli, M. M. Rahman, and M. A. Moni, "'Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach,"' *Artif. Intell. Rev.*, vol. 57, no. 3, p. 62, Mar. 2024.

[21] M. A. Jarwar, R. A. Abbasi, M. Mushtaq, O. Maqbool, N. R. Aljohani, A. Daud, J. S. Alowibdi, J. R. Cano, S. García, and I. Chong, "CommuniMents: A framework for detecting community based sentiments for events," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 2, pp. 87–108, Apr. 2017.

[22] S. Jickson, V. Anoop, and S. Asharaf, "Machine learning approaches for detecting signs of depression from social media," in *Proc. Int. Conf. Inf. Technol. Appl. (ICITA)*. Cham, Switzerland: Springer, 2023, pp. 201–214.

[23] R. John, V. Anoop, and S. Asharaf, "Health mention classification from user-generated reviews using machine learning techniques," in *Proc. Int. Conf. Inf. Technol. Appl. (ICITA)* Cham, Switzerland: Springer, 2022, pp. 175–188.

[24] L. H. Kaack, P. L. Donti, E. Strubell, G. Kamiya, F. Creutzig, and D. Rolnick, "Aligning artificial intelligence with climate change mitigation," *Nature Climate Change*, vol. 12, no. 6, pp. 518–527, Jun. 2022.

[25] R. Kasumba and M. Neumman, "Practical sentiment analysis for education: The power of student crowdsourcing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 23110–23118.

[26] S. Kazi, S. Khoja, and A. Daud, "A survey of deep learning techniques for machine reading comprehension," *Artif. Intell. Rev.*, vol. 56, no. S2, pp. 2509–2569, Nov. 2023.

[27] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Lang. Process. J.*, vol. 4, Sep. 2023, Art. no. 100026.

[28] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.

[29] K. Krippendorff, "Bivariate agreement coefficients for reliability of data," *Sociol. Methodol.*, vol. 2, p. 139, Jan. 1970.

[30] Y.-W. Lai and M.-Y. Chen, "Using natural language processing with explainable AI approach to construct a human-centric consumer application for financial climate disclosures," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1112–1121, Feb. 2024.

[31] F. Larosa, S. Hoyas, J. García-Martínez, J. A. Conejero, F. Fuso Nerini, and R. Vinuesa, "Halting generative AI advancements may slow down progress in climate research," *Nature Climate Change*, vol. 13, no. 6, pp. 497–499, Jun. 2023.

[32] S. Lekshmi and V. Anoop, "Sentiment analysis on COVID-19 news videos using machine learning techniques," in *Proc. Int. Conf. Frontiers Comput. Syst. (COMSYS)*. Cham, Switzerland: Springer, 2022, pp. 551–560.

[33] Q. Lu, X. Sun, Z. Gao, Y. Long, J. Feng, and H. Zhang, "Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis," *Inf. Process. Manage.*, vol. 61, no. 1, Jan. 2024, Art. no. 103538.

[34] M. Lydiri, Y. El Mourabit, Y. El Habouz, and M. Fakir, "A performant deep learning model for sentiment analysis of climate change," *Social Netw. Anal. Mining*, vol. 13, no. 1, p. 8, Dec. 2022.

[35] E. Park, H. H. Loc, D. Van Binh, and S. Kantoush, "The worst 2020 saline water intrusion disaster of the past century in the Mekong delta: Impacts, causes, and management implications," *Ambio*, vol. 51, no. 3, pp. 691–699, Mar. 2022.

[36] J. Purohit and H. S. Rout, "Impact of climate change on human health concerning climate-induced natural disaster: Evidence from an eastern Indian state," *Climatic Change*, vol. 176, no. 8, p. 109, Aug. 2023.

[37] B. Reynante, N. M. Ardoin, and R. Pea, "Reducing the cognitive abstractness of climate change through an 'engineering fiction,"' learning experience: A natural language processing study," *J. Environ. Psychol.*, vol. 95, Sep. 2024, Art. no. 102287.

[38] A. V. Sanson and A. S. Masten, "Climate change and resilience: Developmental science perspectives," *Int. J. Behav. Develop.*, vol. 48, no. 2, pp. 93–102, Mar. 2024.

[39] N. M. Sham and A. Mohamed, "Climate change sentiment analysis using lexicon, machine learning and hybrid approaches," *Sustainability*, vol. 14, no. 8, p. 4723, Apr. 2022.

[40] R. Singh, K. Ram, C. Yadav, and A. R. Siddiqui, "Climate change, disaster and adaptations: Human responses to ecological changes," in *Climate Change, Disaster Adaptations: Contextualising Human Responses to Ecological Change*. Cham, Switzerland: Springer, 2022, pp. 121–130.

[41] U. Sirisha and B. S. Chandana, "Aspect based sentiment & emotion analysis with ROBERTa, LSTM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 11, 2022.

[42] S. Sreelakshmi and S. S. Vinod Chandra, "Machine learning for disaster management: Insights from past research and future implications," in *Proc. Int. Conf. Comput., Commun., Secur. Intell. Syst. (IC3SIS)*, Jun. 2022, pp. 1–7.

[43] A. Upadhyaya, M. Fisichella, and W. Nejdl, "A multi-task model for sentiment aided stance detection of climate change tweets," in *Proc. Int. AAAI Conf. Web Social Media*, 2022, pp. 1–12.

[44] S. A. Vaghefi, C. Huggel, V. Muccione, H. Khashehchi, and M. Leippold, "Deep climate change: A dataset and adaptive domain pre-trained language models for climate change related tasks," in *Proc. NeurIPS Workshop Tackling Climate Change With Mach. Learn.*, 2022, pp. 1–11.

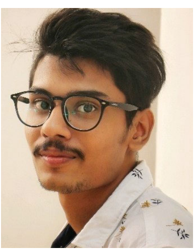
[45] M. Varghese and V. Anoop, "Deep learning-based sentiment analysis on COVID-19 news videos," in *Proc. Int. Conf. Inf. Technol. Appl. (ICITA)*, Cham, Switzerland: Springer, 2022, pp. 229–238.

[46] F. S. Varini, J. Boyd-Graber, M. Ciaramita, and M. Leippold, "ClimaText: A dataset for climate change topic detection," 2020, *arXiv:2012.00483*.

[47] N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold, "ClimateBert: A pretrained language model for climate-related text," 2021, *arXiv:2110.12010*.

[48] Y. Yu, S. Scheidegger, J. Elliott, and Å. Löfgren, "climateBUG: A data-driven framework for analyzing bank reporting through a climate lens," *Expert Syst. Appl.*, vol. 239, Apr. 2024, Art. no. 122162.

V. S. ANOOP received the master's degree in computer applications from the School of Computer and Information Sciences, Indira Gandhi National Open University (IGNOU), in 2011, the M.Phil. and Ph.D. degrees in computer science from the Faculty of Technology, Cochin University of Science and Technology, Kochi, India, in 2014 and 2020, respectively, and the Postgraduate Diploma degree in higher education from the School of Education, IGNOU. He was a Postdoctoral Fellow with the Smith School of Business, Queen's University, Kingston, ON, Canada. During his career, he has worked in various roles as a Senior Research Scientist, an Assistant Professor of computer science, a Data Scientist, and a Visiting Faculty Member. He is currently the Research Officer of the School of Digital Sciences, Kerala University of Digital Sciences, Innovation and Technology, Thiruvananthapuram, India. He has published two books, 15 book chapters, 20 international journals, and 22 international conference papers. His research interests include applied natural language processing and text mining. He was a recipient of the Mozilla Foundation and United States Agency for International Development (USAID) Responsible Computing Challenge Grant in 2023.



T. K. AJAY KRISHNAN received the bachelor's degree in physics from the St. Berchmans College, Kerala, India, and the Master of Computer Science degree in data analytics from Kerala University of Digital Sciences, Innovation and Technology, Thiruvananthapuram, India. He is currently a Junior Data Science Engineer with the Applied AI Innovations and Research Laboratory, School of Digital Sciences, Digital University Kerala. His research interests include machine learning, deep learning, and natural language processing.

ALI DAUD received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in July 2010. He is currently a Full Professor with the Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates. He has 13 years of post-Ph.D. experience in teaching, supervision, and research at B.S., M.S., and Ph.D. levels. He has published over a hundred research papers in reputed international impact factor journals and conferences. He has also taken part in many research projects and has written and acquired much research funding. He has proven and extensive experience in data mining, artificial intelligence (machine learning/deep learning) applications to social networks, data science, and natural language processing.

AMEEN BANJAR received the Ph.D. degree in distributed network functions virtualization from the University of Technology Sydney, Sydney, Australia, in November 2016. He is currently an Associate Professor of information technology and telecommunications with the Department of Information Systems and Technology, College of Computer Science and Engineering (CCSE), University of Jeddah, Saudi Arabia. He has published extensively, including numerous conference papers, journal articles, and book chapters. His research interests include industry 4.0, leveraging intelligent digital technology, machine learning, and deep learning for enhanced manufacturing ecosystems. He also specializes in data science analytics and modeling. He held various academic and research positions, contributing significantly to advancements in his field.

AMAL BUKHARI is currently an Assistant Professor with the Department of Information Systems and Technology, College of Computer Science and Engineering (CCSE), University of Jeddah, Saudi Arabia. She has published many high-quality research papers in reputable international journals and conferences. She has taken part in many research projects as well. Her research interests include data science, data mining, and artificial intelligence.