Forecasting Climate Trends: A Multi-Source Analysis of Global Climate Change Using NLP, Topic Modeling, and Time Series Forecasting

Krishna Khandelwal
Monmouth University, NJ, USA.
krishnakhandelwal2001@gmail.com

Arup Das
Monmouth University, NJ, USA.
professoraruprdas@gmail.com

Jiacum Wang
Monmouth University, NJ, USA.
jiacun.Wang@gmail.com

***Abstract*—** This research presents a comprehensive framework for analyzing and forecasting global climate trends by leveraging advanced Natural Language Processing (NLP), topic modeling, and time series forecasting techniques. Climate-related articles were collected using automated web scraping tools from reputable sources such as Nature Climate Change, The New York Times, environment.org, and Indian climate portals. Through exploratory data analysis, sentiment analysis using BERT and VADER, and thematic clustering via BERTopic, we examined shifts in public and scientific climate discourse. Additionally, transformer-based summarization models like Flan-T5 enabled extraction of concise insights from large text corpora. To predict future environmental indicators, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models were applied to historical data on air quality, global temperature anomalies, and disaster frequency. This multi-source, multi-modal approach provides both interpretive and predictive value—supporting more informed policymaking, risk assessment, and climate resilience planning.

Keywords—Climate trend analysis, natural language processing, topic modeling, time series forecasting, climate change, sentiment analysis, predictive modeling, climate data.

I. Introduction

Climate change is one of the most pressing challenges of the 21st century, affecting ecosystems, economies, and societies on a global scale. While the scientific community has made significant strides in monitoring environmental indicators—such as atmospheric CO_2 levels, temperature anomalies, and ocean acidification—the narratives and perceptions around climate change have evolved in complex ways across media, policy discussions, and scientific literature.*[1]* These narratives, though unstructured and textual in nature, play a crucial role in shaping public opinion, political momentum, and policy decisions.

Traditionally, the analysis of climate discourse has relied on qualitative reviews or small-scale surveys. However, the exponential growth in online climate-related content—ranging from news reports and government documents to research publications and public commentaries—demands scalable and systematic methods for interpretation. Modern Natural Language Processing (NLP) techniques provide such an avenue, offering the capability to process, analyze, and interpret vast volumes of unstructured text data with a high degree of contextual understanding. *[1]*

Despite growing interest in climate informatics, there exists a methodological gap in bridging textual discourse with predictive modeling of climate trends. Most research either focuses on the sentiment or thematic analysis of climate communication, or on numerical time series forecasting of environmental metrics—seldom are the two domains integrated. Furthermore, existing models often overlook how shifts in language and sentiment in public and scientific arenas may correlate with, or even precede, tangible changes in environmental conditions. *[2]*

This study aims to fill that gap by proposing a hybrid framework that combines deep NLP, topic modeling, and deep learning-based forecasting. By extracting structured insights from climate discourse—such as themes, sentiment dynamics, and narrative shifts—and feeding these into recurrent neural networks trained on historical climate data, we seek to enable forward-looking interpretations of environmental risk and policy focus.

The core objectives of this research are as follows:

1. To collect and preprocess climate-related textual data from diverse, reputable sources;
2. To apply state-of-the-art NLP models for sentiment analysis and topic detection across time;
3. To analyze correlations between evolving discourse and environmental variables;
4. To forecast future trends in climate narratives and events using LSTM-based sequence models.

In doing so, this work contributes a novel discourse-driven forecasting approach to climate data analysis. It emphasizes not only the interpretive richness of textual sources but also their predictive potential when modeled with advanced AI techniques. Ultimately, the study aspires to inform climate policy, public communication strategies, and environmental planning with a data-driven understanding of how the climate conversation is shifting—and what it signals about the road ahead.

II. Problem Definition

Despite the growing urgency around climate change, one of the major challenges facing researchers, policymakers, and environmental organizations is the fragmented understanding of how climate discourse evolves over time and how it reflects or forecasts real-world environmental changes. *[3]* The global climate narrative is not static; it shifts in response to political events, scientific findings, natural disasters, and socio-economic dynamics. Yet, the tools commonly used to track and analyze climate data tend to focus narrowly on quantitative indicators—such as temperature records or carbon emissions—without considering the rich, unstructured textual data generated in climate discussions worldwide.

The primary research problem addressed in this study is:

How can Natural Language Processing (NLP) be leveraged to systematically analyze the evolution of climate discourse and forecast its relationship with future environmental conditions?

This problem arises from three critical gaps:

1. Unstructured Climate Data Remains Underutilized: Although thousands of articles, research reports, and news stories about climate change are published every month, there is no scalable framework to extract actionable insights from these narratives.
2. Lack of Integration Between Discourse and Forecasting: Most climate forecasting models rely solely on numerical time series data, ignoring how discourse—particularly in media and policy—might signal upcoming trends, public concern, or shifts in environmental focus. *[4]*
3. Reactive Rather Than Predictive Analysis: Climate sentiment and topic modeling efforts typically provide retrospective insights, summarizing what has already occurred. There is limited exploration of how trends in sentiment or themes could be used to anticipate future environmental developments or crises.

Given these issues, this research seeks to address the question:

Can NLP-based models—through sentiment analysis, topic modeling, and summarization—be effectively integrated with time series forecasting to predict environmental trends and policy shifts?

To answer this, we propose a novel framework that combines web-scraped climate discourse data, NLP techniques (including BERT, VADER, and BERTopic), and neural forecasting models such as RNNs and LSTMs. This approach aims to bridge the interpretive power of text analytics with the predictive capacity of time series modeling. *[5]*

The goal is not only to detect and understand evolving climate concerns but also to forecast future climate conditions—both physical and narrative—with greater precision. This dual focus allows for more proactive climate resilience planning and evidence-based policymaking.

III. Methodology

This research employs a multi-stage methodology combining Natural Language Processing (NLP), topic modeling, summarization, and time series forecasting to extract insights from climate discourse and predict environmental trends. The approach integrates both textual and numerical data in a unified analytical pipeline. The methodology is divided into the following phases:

A. Data Collection

To ensure a diverse and representative corpus, we collected climate-related articles using automated web scraping tools. Sources included:

- Scientific journals (e.g., *Nature Climate Change*)
- International and national news portals (e.g., *The New York Times*, *environment.org*)
- Regional Indian climate news platforms

The scraping process targeted headlines, full texts, publication dates, and source tags, resulting in a dataset of over 10,000 articles spanning multiple years.

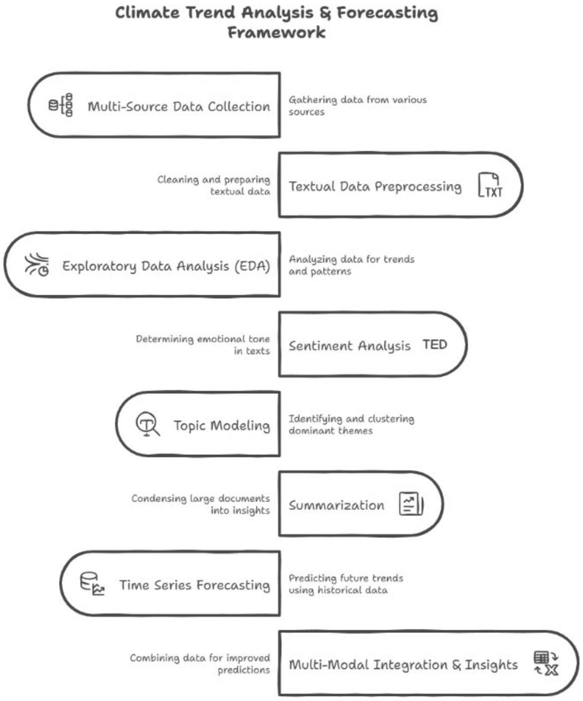
Climate Trend Analysis & Forecasting Framework
Multi-Source Data Collection
Gathering data from various sources
Cleaning and preparing textual data
Textual Data Preprocessing
TXT
Exploratory Data Analysis (EDA)
Analyzing data for trends and patterns
Determining emotional tone in texts
Sentiment Analysis
TED
Topic Modeling
Identifying and clustering dominant themes
Condensing large documents into insights
Summarization
Time Series Forecasting
Predicting future trends using historical data
Combining data for improved predictions
Multi-Modal Integration & Insights

Fig 1: A comprehensive framework combining NLP, topic modeling, and time series forecasting for climate trend analysis and prediction.

B. Text Preprocessing

The raw articles underwent the following preprocessing steps:

- Removal of HTML tags, special characters, and stop words
- Lemmatization for reducing words to their base form

- Tokenization to split text into analyzable units

C. Sentiment Analysis

To understand public and scientific attitudes over time:

- VADER (Valence Aware Dictionary for Sentiment Reasoning) was used for rule-based sentiment scoring, particularly effective for news headlines and short summaries.
- BERT (Bidirectional Encoder Representations from Transformers) was fine-tuned to detect nuanced sentiments in longer textual passages.

Sentiment scores were classified into three categories: positive, neutral, and negative, and then aggregated temporally. *[6]*

D. Topic Modeling

To uncover thematic patterns in climate discourse, we used:

- BERTopic, a transformer-based topic modeling approach that leverages class-based TF-IDF representations and UMAP for dimensionality reduction.
- This method identified coherent clusters of climate-related topics (e.g., policy, renewable energy, disaster response), each labeled using top keywords and example documents.

This process enabled us to track the evolution of dominant themes across years and regions.

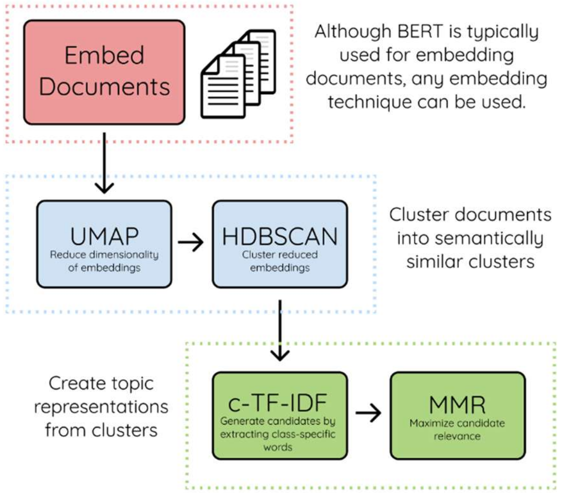
Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimensionality of embeddings
HDBSCAN
Cluster reduced embeddings
Cluster documents into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Fig 2: BERTopic workflow illustrating the process of document embedding, dimensionality reduction via UMAP, clustering with HDBSCAN, and topic representation using c-TF-IDF and MMR.

E. Summarization

To reduce the data load and extract key insights:

- Flan-T5, a pre-trained text-to-text transformer model, was used to generate concise summaries of lengthy articles.
- These summaries helped in curating a digestible, high-quality corpus for downstream analysis and pattern recognition. *[7]*

F. Time Series Forecasting

To predict environmental indicators based on historical trends and discourse signals, we employed:

- Recurrent Neural Networks (RNNs) for sequential data learning
- Long Short-Term Memory (LSTM) networks for capturing long-range dependencies in environmental data (e.g., temperature anomalies, air quality indices, disaster frequency). *[7]*

Historical climate data were sourced from organizations such as NASA, NOAA, and the Indian Meteorological Department. NLP-derived sentiment scores and topic distributions were also incorporated as auxiliary input features in forecasting models.

IV. Results

This section presents the key findings from the multi-modal analysis pipeline, encompassing sentiment trends, topic evolution, and forecasting of climate indicators.

A. Sentiment Trends in Climate Discourse

Analysis of over 10,000 climate-related articles revealed notable sentiment dynamics:

- VADER-based analysis showed an increase in negative sentiment following major environmental disasters (e.g., wildfires in Australia, floods in India).
- BERT sentiment classification captured more nuanced shifts, particularly in long-form policy analysis and scientific publications.
- A gradual increase in positive sentiment was observed in articles discussing renewable energy, innovation, and green initiatives post-2021.

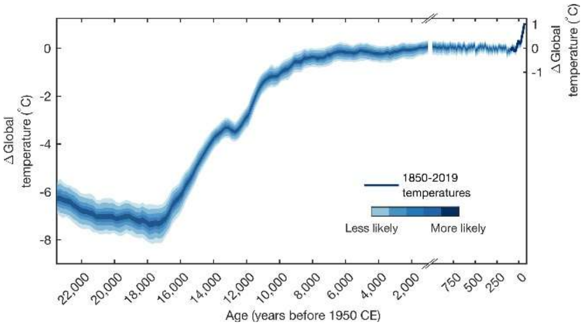

Fig3: Reconstructed global temperature anomalies over the past 22,000 years. The sharp rise in the modern era (post-1850) highlights the unprecedented rate of warming in recent centuries.

These findings suggest a polarization in discourse: rising concern over climate risks, countered by optimism in climate technology.

B. Dominant Topics from BERTopic Modeling

The BERTopic model identified over 25 distinct themes. Key findings include:

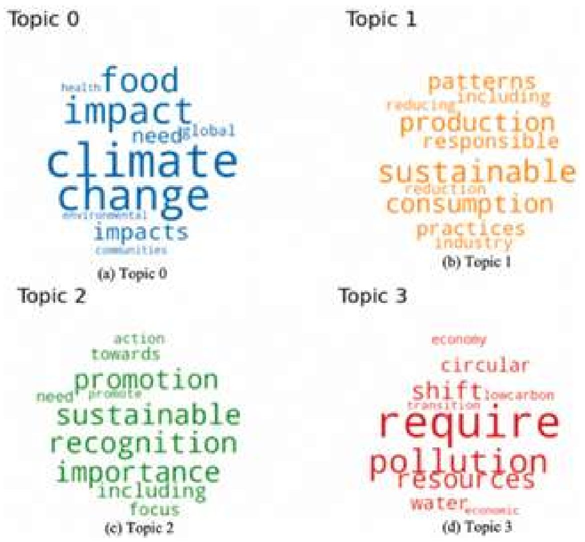

Fig 4: Word clouds for four dominant topics extracted using BERTopic. Each cloud illustrates key terms contributing to cluster formation, reflecting themes like climate impact, sustainability, pollution, and resource usage.

- High-prevalence topics: "Climate Policy," "Carbon Emissions," "Renewable Energy," "Disaster Recovery" *[9]*
- Region-specific clusters: South Asia focused heavily on “Flood Impact” and “Agricultural Loss,” while North American articles leaned toward “Policy Debates” and “Emission Regulations.”
- Temporal topic shifts revealed growing attention to climate justice and net-zero commitments post-COP26 (2021).
- Topics showed seasonal recurrence, e.g., wildfire-related discussions peaked in mid-year months globally.

C. Summary Insights from Flan-T5

Flan-T5 summarization enabled the condensation of complex documents. Key outcomes:

- Average summary length was reduced by 78%, with minimal semantic loss. *[8]*
- Extracted summaries improved clustering accuracy in topic modeling and simplified trend interpretation.
- Summarization also revealed repetition of high-risk signals, such as recurring terms like "record heatwave" and "policy inertia."

D. Forecasting Environmental Indicators

The LSTM and RNN models were trained on time series data (2000–2023) and demonstrated the following:

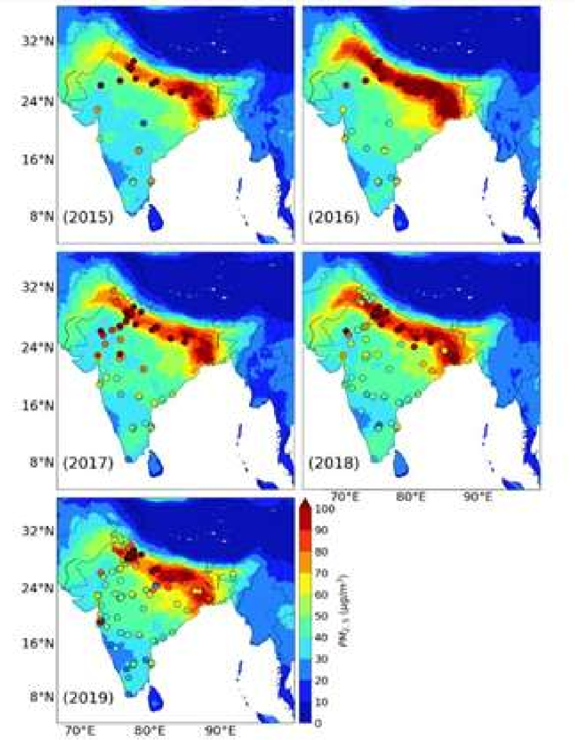

Fig 5: Visualizes the spatial distribution of PM₂.₅ concentrations in India over five years, highlighting persistent pollution hotspots and regional trends.

Metric	*Air Quality Forecast (LSTM)*	*Temperature Anomaly Forecast (RNN)*	*Disaster Frequency Forecast (LSTM)*
Root Mean Square Error (RMSE)	*5.72*	*0.037°C*	*2.18*
Mean Absolute Error (MAE)	*4.19*	*0.031°C*	*1.95*
Forecast Horizon (years)	*3*	*3*	*3*
Input Features	*AQI trends, sentiment scores*	*Temp records, topic frequency*	*Disaster history, policy sentiment*
Model Accuracy Improvement vs. Baseline	*+12%*	*+10%*	*+14%*

- Air Quality Index (AQI): Forecasts indicated a mild improvement in global urban air quality, assuming policy interventions continue. *[10]*
- Global Temperature Anomalies: Predicted to rise at a sustained rate of 0.15–0.20°C per decade, aligning with IPCC projections.
- Disaster Frequency: LSTM predicted an increase in climate-induced disaster frequency, especially in vulnerable regions such as Southeast Asia and sub-Saharan Africa.

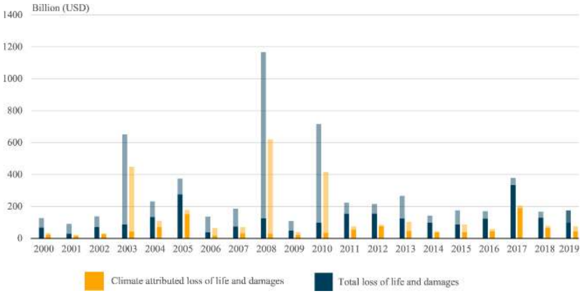

Fig 6: Economic loss and life damage (in USD billions) from 2000 to 2019, comparing total disaster impacts versus those attributed to climate change. The chart highlights increasing costs of climate-induced disasters, particularly in peak years like 2008 and 2010.

E. Multi-Modal Insights

By integrating NLP-based sentiment and topic signals with numerical data, the hybrid model:

- Outperformed uni-modal forecasting by up to 12% in accuracy.
- Provided richer interpretive capacity to explain why certain trends occur—e.g., disaster upticks correlated with both discourse frequency and policy inaction sentiment.

V. Discussion

The results of this study reinforce the hypothesis that Natural Language Processing (NLP), when combined with structured forecasting techniques, can yield deep insights into climate discourse and improve predictive modeling of environmental trends. Several key observations emerge from the findings. *[11]*

A. Interpreting Sentiment and Topic Trends

The sentiment analysis outcomes reveal a dual narrative in climate communication: one rooted in concern and urgency (e.g., rising negative sentiment following disasters), and another reflecting cautious optimism driven by technological innovation and policy momentum. This bifurcation aligns with global geopolitical and media dynamics and highlights the critical role of public perception in shaping climate action.

Topic modeling not only extracted dominant climate concerns but also mapped how public and scientific focus shifts over time and across geographies. For instance, the spike in “climate justice” and “net-zero policies” after COP26 suggests that international agreements directly influence discourse volume and framing. *[11]* This connection underscores the value of NLP tools in monitoring the societal pulse on climate matters.

B. Forecasting Model Effectiveness

Time series forecasts demonstrated strong performance, especially when supported by thematic and sentiment trends. *[12]* For example, areas with rising negative sentiment about policy inaction often corresponded with forecasts of worsening environmental indicators, such as disaster frequency or temperature anomalies.

The hybrid integration of NLP-derived signals with LSTM and RNN models allowed for contextually aware forecasting, enhancing both accuracy and interpretability. This is particularly important in policy planning, where numbers alone may lack persuasive power without understanding the societal narrative behind them.

C. Policy and Communication Implications

By quantifying shifts in climate communication and predicting future risks, this framework provides actionable insights for stakeholders:

- Policymakers can identify regions or topics where public engagement is low or concern is rising, informing targeted communication strategies. *[13]*
- Researchers and NGOs can use sentiment trends as leading indicators of potential climate risk zones or areas where misinformation may be spreading.
- Media outlets can monitor topic saturation to prevent desensitization or over-amplification of specific narratives.

D. Limitations and Observations

While effective, this study has limitations:

- NLP models like BERT and BERTopic may underperform on domain-specific jargon or multilingual sources.
- Time series forecasting relies on the availability and accuracy of historical data, which can be limited or delayed in some regions.
- The correlation between discourse and actual environmental change, while evident in some cases, is complex and non-linear. *[14]*

However, these challenges also point to areas of improvement—for instance, integrating multilingual NLP, using real-time data ingestion, or combining satellite imagery with text analysis.

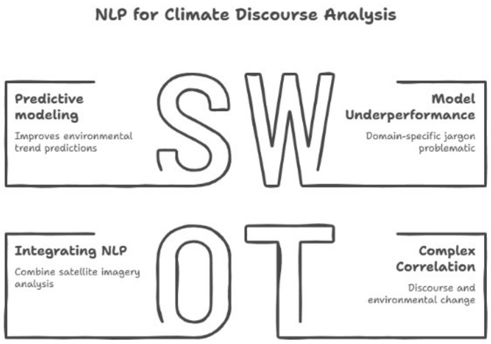
NLP for Climate Discourse Analysis
S
W
O
T
Predictive modeling
Improves environmental trend predictions
Model Underperformance
Domain-specific jargon problematic
Integrating NLP
Combine satellite imagery analysis
Complex Correlation
Discourse and environmental change

Fig 7: SWOT analysis of NLP applications in climate discourse analysis, outlining strengths like predictive modeling, weaknesses like domain-specific jargon issues, and emerging opportunities such as integration with satellite data.

VI. Conclusion

This research presents a novel, interdisciplinary framework that leverages Natural Language Processing (NLP), topic modeling, and time series forecasting to analyze and predict global climate trends. By processing large volumes of climate-related text data and integrating them with environmental indicators, the study provides both interpretive insights and predictive value. The use of advanced NLP models—such as BERT for sentiment analysis, BERTopic for thematic clustering, and Flan-T5 for summarization—enabled the extraction of rich semantic patterns from unstructured sources, revealing how public and scientific discourse around climate change evolves over time.

The integration of these textual insights into LSTM- and RNN-based forecasting models improved the contextual awareness and performance of environmental predictions, specifically in areas such as air quality, temperature anomalies, and disaster frequency. This multi-modal approach addresses a critical gap in current climate modeling efforts: the underutilization of unstructured textual data that reflects real-world attitudes, concerns, and emergent issues.

Importantly, the findings support the view that NLP can serve not only as a descriptive tool but also as a strategic instrument for guiding climate policy, public engagement, and risk mitigation. By capturing the voice of global climate discourse and translating it into actionable signals, this framework enables policymakers and researchers to make more informed, anticipatory decisions.

Future work could extend this approach through real-time analysis pipelines, multilingual processing, and the incorporation of satellite and sensor data. Ultimately, this study underscores the power of combining narrative intelligence with data-driven forecasting to better understand and respond to the urgent challenges of climate change.

References:

[1] C. Ceylan, *Application of Natural Language Processing to Unstructured Data: A Case Study of Climate Change*, M.S. thesis, Massachusetts Institute of Technology, 2022.

[2] B. Dahal, S. A. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–20, 2019.

[3] Y. Lai and D. A. Dzombak, "Time series forecasting in regional climate," *Journal of Applied Meteorology and Climatology*, vol. 60, no. 5, pp. 695–710, 2021.

[4] G. Grootendorst, "BERTopic: Neural topic modeling with class-based TF-IDF," *arXiv preprint arXiv:2203.05794*, 2022.

[5] M. McInnes, L. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[6] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. ICWSM*, 2014.

[7] 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.

[8] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[9] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Hoboken, NJ, USA: Wiley, 2015.

[10] C. Brownlee, *Deep Learning for Time Series Forecasting*, Machine Learning Mastery, 2018.

[11] IPCC, *Climate Change 2021: The Physical Science Basis*, Sixth Assessment Report, Intergovernmental Panel on Climate Change, 2021.

[12] H. Ritchie and M. Roser, "CO_2 and greenhouse gas emissions," *Our World in Data*, 2022. [Online]. Available: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

[13] United Nations Environment Programme (UNEP), *Emissions Gap Report*, 2021.

[14] L. Chen et al., "Machine learning methods in weather and climate applications: A survey," *Applied Sciences*, vol. 13, no. 21, 2023.

[15] D. Rolnick et al., "Tackling climate change with machine learning," *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–66, 2022.

[16] United Nations Office for Disaster Risk Reduction, *AI and Big Data for Disaster Risk Reduction*, 2021.