EnClaim: A Style Augmented Transformer Architecture for Environmental Claim Detection

Diya Saha, Manjira Sinha and **Tirthankar Dasgupta**
TCS Research
India
(diya.saha, sinha.manjira, dasgupta.tirthankar)@tcs.com

Abstract

Across countries, a noteworthy paradigm shift towards a more sustainable and environmentally responsible economy is underway. However, this positive transition is accompanied by an upsurge in greenwashing, where organizations make exaggerated claims about their environmental commitments. To address this challenge and protect consumers, initiatives have emerged to substantiate green claims. With the proliferation of environmental and scientific assertions, a critical need arises for automated methods to detect and validate these claims at scale. In this paper, we introduce EnClaim, a transformer based architecture augmented with stylistic features for automatically detecting claims from open web documents and social media posts. The proposed model considers various linguistic stylistic features in conjunction with language models to predict whether a given statement constitutes a claim. We have rigorously evaluated the model using multiple open datasets. Our initial findings indicate that incorporating stylistic vectors alongside the BERT-based language model enhances the overall effectiveness of environmental claim detection.

1 Introduction

Amid the ongoing climate crisis, a remarkable shift is taking place towards establishing a more sustainable and environmentally responsible economy. This transition is primarily being propelled by evolving regulations, shifting public sentiments, and changing attitudes among investors. However, this promising shift has been accompanied by a surge in greenwashing, with companies making exaggerated claims about their environmental commitments[1]. Such environmental advertisements can also mislead consumers due to vague or false claims, thereby harming brand or product outcomes. To address this challenge and protect consumers, initiatives have emerged to substantiate green claims. With the proliferation of environmental and scientific assertions, there is a pressing need for automated methods to detect and validate these claims at scale. This capability can prove invaluable for policymakers, regulators, journalists, activists, the research community, and an informed public, enabling them to thoroughly assess and scrutinize environmental and scientific claims made by companies and thus advance the transition to a greener company. Consequently, the first step towards claim validation is to first detect the claims from a collection of environment-related textual mentions. For example, Table 1 depicts sample sentences from the environmental domain. However, not all of them are making claims. Thus, we introduce the task of environmental claim detection. This intriguing task involves classifying sentences to discern whether they contain environmental and scientific claims. For the definition of such claims, we follow the definition provided by the European Commission (EC), which is, *Environmental claims refer to the practice of suggesting or otherwise creating the impression (in the context of commercial communication, marketing or advertising) that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or*

Type (0/1)	Texts
1	A total population of 6148 is getting the benefit of safe potable drinking water due to this initiative.
0	Our ambition is to be the preferred energy company for all stakeholders, and we have a disciplined three-phase strategy to meet that ambition.
1	Says GOP primary opponents Glenn Grothman and Joe Leibham cast a com-promise vote that cost $788 million in higher electricity costs.
0	Says the Annies List political group supports third-trimester abortions on demand.

Table 1: Sample textual mentions depicting *claim (1)* and *not a claim (0)*

[1]See, e.g., The Economist, May 22nd, 2021.

is less damaging to the environment than competing goods or services.[2] To be precise, Environmental claims are often clearly and concisely articulated at the sentence level, designed to transparently convey a company or product's environmentally friendly qualities to consumers and stakeholders, and the said property is utilized to differentiate between statements that assert a claim and those that don't.

Recent advancements in large-scale language models such as transformers, and GPT have influenced researchers to apply such techniques for claim detection tasks. The advantage of such techniques lies in the fact that multi-layer neural networks can automatically learn complex language structures. However, such deep neural network architectures are yet to take into account integral linguistic factors present in the text, which play an important role in determining claim-type statements. Moreover, existing approaches to claim detection seldom consider the deep stylistic features embedded within the text that can play an important role in the classification task

Accordingly, in this paper, we propose a stylistically enhanced transformer-based architecture for the automatic classification of statements into "*claim*" or *"not a claim"* statements. Our model considers stylistic feature embeddings along with the standard transformer-based language model. To our knowledge, no prior work in this field has investigated the effectiveness of combining the above factors for environmental claim detection tasks. Our preliminary investigation shows that the incorporation of stylistic feature vectors along with the language model does improve the overall performance of the classification model and it is not only limited to environmental claim detection. Regardless of the dataset, our proposed architecture empowers claim detection leveraging stylistic fingerprints within sentences.

2 Related Work

A plethora of studies have been done on the analysis of environmental fake news, and corporate greenwashing. Recent endeavors on building computational models include ClimateBERT (Webersinke et al., 2022), and ClimateGPT (Vaghefi et al., 2022), two language models pre-trained on climate-related text. NLP tasks and datasets include climate change topic detection (Varini et al., 2020) and detecting media stance on global warming (Luo et al., 2020a). Duong et al. (2022) (Duong et al., 2022) collect climate change opinions at scale from social platforms, Al-Rawi et al.(2021) (Al-Rawi et al., 2021) analyze fake news Tweets around climate change. In a similar direction Coan et al. (2021) (Coan et al., 2021) analyze contrarian claims about climate change and Piskorski et al., 2022 (Piskorski et al., 2022) explore data augmentation techniques for climate change denial classification. Further, there exists work on claim verification of climate change-related claims (Diggelmann et al., 2020 (Diggelmann et al., 2020)), detecting media stance on global warming (Luo et al., 2020 (Luo et al., 2020b)), collecting climate change opinions at scale from social platforms (Duong et al.,2022 (Duong et al., 2022)), and finally, the analysis of regulatory disclosures (Friederich et al., 2021 (Friederich et al., 2021); Kölbel et al., 2022 (Kölbel et al., 2020)). Claim spotting is the task of finding fact-check worthy claims (Arslan et al., 2020 (Arslan et al., 2020); Atanasova et al., 2018 (Atanasova et al., 2018); Barron-Cedeno et al., 2020 (Barron-Cedeno et al., 2020)). Pledge detection aims to detect pledges made in, for example, political campaigns (Subramanian et al., 2019 (Subramanian et al., 2019); Fornaciari et al., 2021 (Fornaciari et al., 2021)). To the best of our knowledge, we have not encountered any approach that extensively makes use of the deep linguistic and stylistic factors for the identification of claim-worthy sentences.

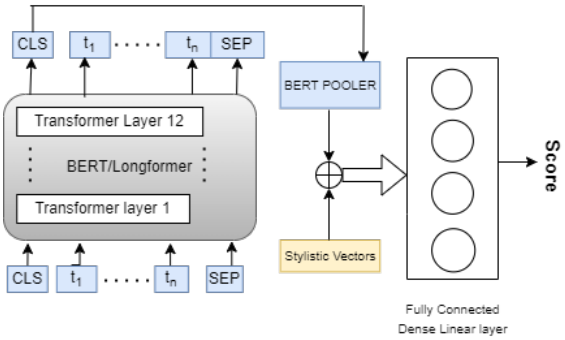

Figure 1: Model architecture depicting augmented stylistic and grammatical error categories with transformer architecture for the claim detection.

3 The Style Aware Transformer Network

In this section, we will present the style-aware transformer network that considers different com-

[2]From the Commission Staff Working Document, Guidance on the implementation/application of Directive 2005/29/EC on Unfair Commercial practices, Brussels, 3 December 2009 SEC(2009) 1666.

plex linguistic, grammatical, and stylistic features associated with a text document connected to a BERT-based language model network. The overall architecture of the model is depicted in Figure 1. We will begin the model architecture by first explaining about generating the grammatical and stylistic feature embeddings that will in turn be used by the transformer architecture.

3.1 Generating Stylistic Vector

Linguistic Inquiry and Word Count is a text analysis program that calculates the percentage of words in a given text that fall into one or more of over 80 linguistic, psychological, and topical categories indicating various social, cognitive, and affective processes. The core of the program is a dictionary containing words that belong to these categories. Dictionaries for many languages are available (Pennebaker et al., 2001). Consequently, we have generated a vector of size 73. The vector represents the frequency of different categories such as Noun, Verb, Adjective, Subject Verb Agreement, etc. in a document. We used the LIWC dictionary published in 2015 (Pennebaker et al., 2015). LIWC reads a given text and compares each word in the text to the list of dictionary words and calculates the percentage of total words in the text that match each of the dictionary categories. Hence, given a text S, we obtain $[x_1, x_2, x_3, x_4, \ldots, x_{73}]$ where each x_{i} represents the frequency in text .

Apart from the LIWC features, we also introduce the following three stylistic factors from a text.

3.1.1 Vagueness

Vagueness or uncertainty refers to the quality of not being clearly expressed, known, described, or decided. Vagueness in the text means that the text lacks explicit details. Instead, there are generalizations and broad judgments used in the text. We used the tree-BiLSTM-based model for vagueness prediction as proposed in the literature (Sinha et al., 2020). To make parity with the other stylistic scores, we took the inverse of the vagueness scores. Therefore, a high score will imply that the document is clearer.

3.1.2 Conviction

Conviction refers to a very strong belief or opinion of a person. Conviction in the text can represent whether the author is not nervous or has questions about his beliefs. A writer with strong convictions would not take sides and the essay flows in a constant direction rather than bouncing back on contrasting sides. We used the Empath library (Fast et al., 2016) that represents conviction as the presence of *pride*, and *trust* and the absence of *timidity*, *nervousness*, and *confusion* in a text. We have followed the tree-BiLSTM-based model as depicted by Sinha et al. in 2020 (Sinha et al., 2020).

3.1.3 Commitment

Commitment refers to the act of binding yourself to a cause because you believe it is right and important. Commitment in a text means whether the text displays the commitment of the writer to a particular point he believes in. Here, commitment is represented as the presence of optimism, zest, gain, and achievement in a text. Like the prior two scores i.e., **Vagueness** and **Conviction**, we have employed the tree-BiLSTM-based model outlined by Sinha et al. in their 2020 publication (Sinha et al., 2020).

These three stylistic measures along with the LIWC scores are then concatenated to obtain a styled vector of size 76 dimension.

3.2 Model Architecture

We use a pre-trained BERT-base-uncased (Devlin et al., 2018) to obtain the BERT pooler output of the text which is the last hidden state of the $[CLS]$ token with predefined transformation $tanh$ activation to give us a 768-size vector. This vector is then concatenated with the LIWC Vector of size 73 and Stylistic Features of size 3. The whole concatenated vector obtained is a vector of size 844. This 844-size vector is then fully connected to a dense linear layer which gives the output score. We use the Environmental Claim Detection dataset (Stammbach et al., 2023), Towards Automatic Green Claim Detection dataset (Woloszyn et al., 2021) and Scientific Claim Detection Dataset (Achakulvisut et al., 2019) to train and test our model. The dataset released by Environmental Claim Detection consists of text from sustainability reports, earning calls, and annual reports of listed companies and annotated 3000 sentences. After discarding tied annotations, our resulting dataset contains 2647 examples. For each Dataset, 70% of the data is used for training and 30% for testing. We train the model for a fixed number of epochs. To conduct further experiments, first we have combined all three datasets and then split the combined dataset so that 70% is used for training and 30% for testing the trained model. To test any

model's ability to detect claims from environmental domain even if it is trained on a different dataset, we have trained BERT and EnClaim models using one dataset and tested them on the other two datasets.

For the given task of predicting the label of a document for continuous labels, our goal is to minimize the root mean squared error (RMSE) rate(Loshchilov and Hutter, 2017). To achieve that, we have used the AdamW (Kingma and Ba, 2017) optimization algorithm while training the BERT model to minimize the root mean squared error (RMSE) over the test data. This is represented as:

$RMSE(s^*, s) = \left(\frac{1}{N} * \sum_{i=1}^{N} (s^*{}_{\mathrm{i}} - s_{\mathrm{i}})^2\right)^{\frac{1}{2}}$

The model computes the predicted labels l_i for all training essays and then updates the network parameters such that the mean squared error is minimized.

We have set the batch size to 32 for BERT. The model is trained for 20 epochs with the learning rate set to $2 * 10^{-5}$. The max tokens are restricted here at 200 since it is the limit of the BERT-base model.

4 Evaluation

4.1 The Dataset

Dataset-ECD: Environmental Claim Detection Dataset (Stammbach et al., 2023) We have collected the environmental claim dataset available from (Stammbach et al., 2023). The dataset contains environmental claims made by listed companies. The authors have collected text from sustainability reports, earning calls, and annual reports of listed companies and annotated 3000 sentences. After discarding tied annotations, the final dataset contains 2647 examples. There are 665 claim statements and 1982 not claim statements.

Dataset-GCC: Green-Claims Corpus (Woloszyn et al., 2021) We choose the Automatic Green Claims Detection corpus consisting of 773 tweets from domains such as cosmetics and electronics. All the tweets are classified into two classes "green-claim" and "not green-claim". For Binary Classification, there are 506 "not green-claim" and 267 "green-claims". In this corpus, only tweets with an agreement more significant than 75% were considered in the final data set.

Dataset-SCDC: Scientific Claim Detection Corpus(Achakulvisut et al., 2019) To test the generalizability of the proposed model, we took a separate dataset outside the environmental domain. The dataset includes text extracts from expertly annotated 11519 claims in biomedical paper abstracts. Here the dataset is labeled into six classes: "False", "barely-false", "half-true", "pants-fire", "barely-true", and "True".While doing Experiment-I, we have assigned [0,0.25), [0.25,0.5), [0.5,0.6), [0.6,0.75), [0.75,1) and 1 respectively for labels. For Experiment II and Experiment III, to transform the dataset into a Binary classification task, claim scores greater than or equal to 0.5 are assigned 1 and rest as 0 (not a claim). Altogether 6500 sentences are marked as 0 and 4000 are marked as 1 (claim).

4.2 Baseline Models

The pre-trained $BERT_{BASE}$ model with 12 layers of self-attention units (Vaswani et al., 2017) is trained over large publicly available data sets. It can be fine-tuned with domain-specific texts to improve downstream processing tasks. In the present paper, we define the downstream tasks as a classification of *green-claim* and *not green-claim*. Accordingly, the transformer-based BERT network is fine-tuned over the given dataset corresponding to the tasks.

Fine-tuning the pre-trained model with training data from different domains is known to improve the performance of language processing tasks. Further, we set the early stopping of fine-tuning to 800 steps to prevent over-fitting. We use a batch size of 32, a maximum sequence length of 200, and a learning rate of $2 * 10^{-5}$ for fine-tuning this model. Finally, post-processing steps are conducted to align the BERT output with the concept gold standard, including handling truncated sentences and word-pieced tokenization.

To compare the performance of the proposed architecture, we have used ClimateBERT (Webersinke et al., 2022) as a baseline model. As Dataset-SCDC consists of scientific claims and is not restricted to Environmental Claims, in Experiment-II (4.3) and Experiment-III (4.3), we use BERT-base-uncased (Devlin et al., 2018) as the baseline model.

4.3 Experiments

Based on the given datasets, we have conducted three different experiments.

In **Experiment-I:** We take each of the individual datasets and divide them into two groups 70% and 30% for training and testing respectively. We have

performed several experiments to identify the best model architecture for our task.

In **Experiment-II:** We have combined all the datasets and formed a combined annotated corpus of 15293 documents. We then divide the entire corpus into 70% and 30% for training and testing respectively. The entire training set is then used to evaluate the proposed models. It is worth mentioning here that in dataset-**SCDC** the training set is prepared in such a way that there are two output classes, unlike the original 6 classes. Accordingly, we have modified our neural network architecture to output the binary classes.

In **Experiment-III:** We have chosen individual datasets, trained our models over the chosen dataset and finally tested them over other datasets. For example, we have trained our models on the **ECD** dataset(Stammbach et al., 2023) and tested the models using the **GCC** (Woloszyn et al., 2021) and the **SCDC** dataset (Achakulvisut et al., 2019).

When we have used Dataset-**SCDC** as a training set, we have also calculated the RMSE Score to calculate the errors as in **SCDC**, the sentences are marked with continuous labels. Therefore, if we do not convert labels into binary the predicted dataset output remains continuous. Consequently, RMSE serves as a superior evaluation metric.

4.4 Fine-tuning Neural Networks on BERT

To implement our proposed architecture, we have fine-tuned the pre-trained BERT models for all three experiments with a fully connected layer on top of the output layer for the classification tasks. We used the SKlearn library to implement the Multilayer Perceptron classifiers(Glorot and Bengio, 2010), setting a learning rate of 10^{-5}, and tanh as activation function, adaptive learning rate (Schaul et al., 2013) and Limited-memory BFGS as optimizer (Zhu et al., 1997), and a maximum number of 80 epochs, after which we follow the standard practice of selecting the best model based on development holdout data.

4.5 Comparison of Proposed Model Architecture With LLMs

According to recent research, LLMs have the potential to outperform numerous transformer designs. After conducting Experiment-I on Dataset-**ECD**, we compared its output to that of LLAMA-2 13B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023). First, we have evaluated the environmental claim detection ability of LLAMA-2 using zero-shot (Wang et al., 2019) and few-shot prompt techniques. Here, we have used the few-shot technique demonstrated by (Min et al., 2022) and given examples of two claim sentences and two not-a-claim sentences as prompt. We have also fine-tuned the pre-trained Mistral-7B Model with Dataset-**ECD** to compare LLM's ability to perform the domain-specific task of environmental claim detection with our proposed architecture. The Mistral-7B outperforms the LLAMA-2 34B despite having only 7.3 billion parameters on various benchmarks (Jiang et al., 2023). Here, we have primarily used transfer learning, with additional modifications such as quantization and the integration of LoRA adapters (Dettmers et al., 2023) to fine-tune Mistral. The training process involves several key steps. The process begins with loading **ECD** and processing it, where each data sample is augmented with a prompt indicating the task context and the statement to be evaluated for environmental claim detection. Quantization reduces Mistral's precision to a lower bit width (from 32-bit to 4-bit), facilitating faster computation and reduced memory usage without significant loss of accuracy. LoRA adapters are attached to specific layers of the model to enable fine-grained control and specialization for the environmental claim detection task. We have configured the training parameters as follows: the batch size is set to 8, the optimizer being used is AdamW (Zhuang et al., 2022), the learning rate is $2 * e^{-4}$, the learning rate scheduler is cosine, the logging steps are set at 50, the number of training epochs is set to 50, and the maximum number of steps is set at 100. These arguments govern the training loop's behavior, optimizing model parameters iteratively to minimize loss and improve performance. The training loop iterates over the dataset for a specified number of epochs or steps, depending on the training argument configuration.

During each iteration, we input a batch of data samples into the model for forward pass computation. The model makes predictions for the environmental claim label for each input statement, and these predicted labels are then compared with the actual labels to calculate the loss. The loss is then used to update the model's parameters through backpropagation, adjusting the model's weights to minimize prediction errors. To reduce memory usage and stabilize training, gradient accumulation steps are used to accumulate gradients over multiple batches before updating the model parameters. Additionally, learning rate scheduling dynamically

adjusts the learning rate during training, typically decreasing it over time to fine-tune the model's convergence. Logging occurs at specific intervals during training to monitor metrics such as loss, training progress, and resource usage. Furthermore, evaluation metrics are calculated periodically to assess the model's performance on validation data, providing insight into its generalization ability and potential overfitting.

5 Results

	Dataset-ECD		
Model	P	R	F1
BiLSTM-att	0.43	0.39	0.46
CNN+BiLSTM-att	0.47	0.40	0.49
BERT-base	0.49	0.74	0.53
ClimateBERT	0.599	0.72	0.65
EnClaim	**0.79**	**0.865**	**0.83**
	Dataset-GCC		
Model	P	R	F1
BiLSTM-att	0.69	0.69	0.75
CNN+BiLSTM-att	0.73	0.75	0.78
BERT-base	0.75	0.71	0.77
ClimateBERT	0.902	0.86	0.88
EnClaim	**0.96**	**0.97**	**0.96**
	Dataset-SCDC		
Model	P	R	F1
BiLSTM-att	0.77	0.78	0.7
CNN+BiLSTM-att	0.77	0.79	0.81
$BERT_{base}$	0.595	0.95	0.73
EnClaim	**0.94**	0.75	**0.84**

Table 2: Results of Experiment-I demonstrating Precision (P), Recall (R), F1 scores (F1) for each model across the different datasets.

In the case of **Experiment-I** (ref: Table 2), we have found that throughout all the target classes the performance of EnClaim i.e., ClimateBERT+Style model is significantly higher than the individual BERT, CNN, and LSTM models. We have also observed that a combination of such embeddings has been very effective in capturing solely contextual information. In most of the cases, the combined representation surpasses the performance of the individual embedding models. Throughout all the models, it is documented that the performance of the ClimateBERT+Style model is higher when AdamW optimizer is used in the training of dense neural networks. During the analysis of the individual datasets we have observed that for Dataset-GCC, we have achieved an F1 score of 96% using the EnClaim model. This is the highest F1 score that we have achieved among all other datasets. For *Dataset-SCDC*, EnClaim i.e., BERT+Style model shows slightly better performance ($F1 = 84\%$) than BERT ($F1 = 73\%$). However, for this dataset, the recall for the EnClaim model decreases significantly from 95% to 75%. In the case of *Dataset-ECD*, the highest F1 score of 83% is achieved in the ClimateBERT+Style model. As discussed earlier, the poor performance of Dataset-SCDC is primarily due to the higher number of output classes.

Model	P	R	F1
BiLSTM-att	0.71	0.73	0.75
CNN+BiLSTM-att	0.73	0.79	0.75
BERT-base	0.81	0.84	0.82
EnClaim	**0.88**	**0.89**	**0.88**

Table 3: Results of Experiment-II demonstrating Precision, Recall, and F1 Scores for each model across the combined dataset.

In **Experiment-II:** Table 3 reports the results obtained after combining all the datasets and testing the individual models. Similar to the observations reported for Experiment-I we can see that the performance of the EnClaim model (BERT+Style) far surpasses the performance of BERT.

Test Dataset	**RMSE Score**
Dataset-ECD	0.25
Dataset-GCC	0.28
Dataset-SCDC	0.15

Table 4: Results of Experiment-III demonstrating RMSE Score for the BERT (baseline model) when trained over dataset $SCDC$ and tested over datasets ECD, GCC and $SCDC$ respectively.

For **Experiment-III:** Here we have trained the models on one single dataset and tested them over other datasets. The results are depicted in Table 4, 5 and 6. For Dataset-ECD and Dataset-SCDC, the multi-class classification is transformed into binary classification. So, the accuracy of the datasets suffers a bit. As labels of Dataset-SCDC are divided into six classes, the results involving this dataset as test data tend to have less accuracy. If we consider all six labels for Dataset-SCDC while using this dataset as a training set then we have to calculate the RMSE score. From Table 4, we can see that when we have trained the model for Dataset-SCDC while testing the model on Dataset-SCDC gives less RMSE score than the other two datasets as Dataset-ECD and Dataset-GCC are based on environmental domain and SCDC is on biomedical domain. If we compare Table 4 and 5 we can see that in both Binary and Multi-class classification if

Training Dataset	Test Dataset Dataset-ECD		
	P	R	F1
Dataset-ECD		X	
Dataset-GCC	0.59	0.89	0.71
Dataset-SCDC	0.34	0.72	0.46
Training Dataset	**Test Dataset** Dataset-GCC		
	P	R	F1
Dataset-ECD	0.84	0.243	0.37
Dataset-GCC		X	
Dataset-SCDC	0.39	0.52	0.45
Training Dataset	**Test Dataset** Dataset-SCDC		
	P	R	F1
Dataset-ECD		No Prediction	
Dataset-GCC	0.41	0.05	0.09
Dataset-SCDC		X	

Table 5: Results of Experiment-III demonstrating Precision, Recall, and F1 score for the BERT-base model when trained over a given dataset D_i (given in rows) and tested over other datasets D_j (given in columns) such that $i \neq j$.

Training Dataset	Test Dataset Dataset-ECD		
	A	P	R
Dataset-ECD		X	
Dataset-GCC	0.93	**0.88**	0.90
Dataset-SCDC	0.95	**0.91**	**0.92**
Training Dataset	**Test Dataset** Dataset-GCC		
	A	P	R
Dataset-ECD	0.93	**0.87**	**0.91**
Dataset-GCC		X	
Dataset-SCDC	0.94	**0.89**	**0.90**
Training Dataset	**Test Dataset** Dataset-SCDC		
	A	P	R
Dataset-ECD	**0.91**	**0.83**	**0.84**
Dataset-GCC	0.81	**0.62**	0.65
Dataset-SCDC		X	

Table 6: Results of Experiment-III demonstrating Accuracy (A), Precision (P), Recall (R) for the BERT+Style (EnClaim Model) when trained over a given dataset D_i and tested over other datasets D_j such that $i \neq j$.

we train the model on dataset-SCDC then dataset-ECD gives better results than dataset-GCC as GCC consists of tweets rather than complete sentences.

Model	P	R	F1
BERT-base	0.49	0.74	0.53
ClimateBERT	0.599	0.72	0.65
EnClaim	**0.79**	**0.865**	**0.83**
LLAMA-2 13B (zero-shot)	0.632	0.534	0.579
LLAMA-2 13B (few-shot)	**0.97**	0.34	0.503

Table 7: Comparison of Precision (P), Recall (R), F1 score (F1) of **EnClaim** generated output with **LLM** generated output for Dataset-ECD

Also from Table 6, we can see that if we use the EnClaim model and use Dataset-SCDC as the training set, the accuracy is higher when Dataset-ECD is used as test data rather than Dataset-GCC but for precision and recall it follows the same pattern as in BERT model. The precision is higher if we use Dataset-ECD as test data instead of Dataset-GCC. So, here accuracy does not express the correct measure of the experiment. From Table 5 we observe that if we use BERT only and train the model on any other data than Dataset-SCDC the results are poor due to the difference of domain whereas in Table 6 the result improves for taking stylistic features into account.

5.1 Comparing Proposed Model with LLAMA-2

In the landscape of large language models (LLMs), there exists an extensive capacity to surpass various transformer architectures. However, empirical evidence which is presented in Table-7, demonstrates that our novel model architecture, EnClaim, achieves superior performance compared to LLAMA-2 13B. This superiority is attributed to EnClaim's deliberate consideration of the syntactic properties inherent within sentences. Such a focus enables EnClaim to leverage syntactic structures effectively, thereby enhancing its ability to comprehend and detect claim sentences with greater accuracy.

As we can see, the performance of LLAMA-2 using the few-shot approach was notably limited. This limitation stemmed from the complexity of defining environmental claims, which necessitates a comprehensive representation beyond the provided examples as prompt. As evidenced in the present table (Ref: Table 7), while LLAMA-2 achieved a high precision score, its recall and F1 scores were significantly lower, primarily due to

its tendency to classify the majority of sentences as not-claims. Consequently, LLAMA-2 exhibited suboptimal classification performance, particularly in the zero-shot scenario. Conversely, although EnClaim emerged as a superior classifier in Precision, Recall, and F1 scores, its superiority can be attributed to its adherence to the intricate definition of environmental claims, thereby underscoring its effectiveness in classification tasks.

5.2 Outcome of fine-tuned Mistral-7B Model for Environmental Claim Detection

Here, we ran an experiment to compare the output of our proposed architecture with our fine-tuned Mistral Model. We gave the fine-tuned Mistral model the test data samples and asked it to categorize them as *claim* or *non-claim* sentences. However, the Large Language Model's hallucinatory property posed a challenge. Out of the text sentences, the trained Mistral Model provided a distinct classification for only 25% cases, while the remaining 25% cases resulted in a rather confusing answer. Among those, it categorized correctly for 22% cases. Therefore, we concluded that while training the large language model on a specific domain can improve its Environmental Claim Detection capacity, the inherent property of the Large Language Model can still pose a challenge. In Table 8, we have provided examples of instances that document the advantages and limitations of the classification capabilities of all the discussed models.

Texts	**EnClaim**	**LLAMA-2**
Article 2 of the agreement also aims to ensure that finance flows are consistent with low carbon impact, climate resilient devlopment.	Correct	Wrong
Can New Jersey continue to afford to pay for a 0 emissions energy?	Wrong	Correct
Historically, we have backed out significant adjustments to the environmental	Wrong	Wrong
Texts	**EnClaim**	**F. Mistral**
And then we're doing as much we can to offset the labor-related expenses.	Correct	Correct
Can New Jersey continue to afford to pay for a 0 emissions energy?	Correct	Hallucination

Table 8: Sample textual mentions depicting the performance of EnClaim, LLAMA-2 13B (zero-shot), fine-tuned Mistral-7B. Here, we have denoted Correct Prediction as Correct and Wrong Prediction as Wrong.

6 Conclusion

In this paper, we present a style-aware transformer architecture for Environmental Claim Detection. While conventional deep neural networks, including CNN and LSTM, have historically struggled to discern the intricate relationships among various grammatical and stylistic elements that play a pivotal role in assessing text quality, our method takes a different approach and considers the distinct stylistic features such as vagueness, conviction and commitment with the power of pre-trained transformer models. These features have proven to be indispensable in the accurate evaluation of environmental claims. We have compared the performance of the proposed model with the state-of-the-art open-source LLMs including finetuned mistral model and LLAMA-2. We have observed across different datasets. EnClaim surpasses most of the state-of-the-art models in terms of precision, recall and F1-Score.

Our future work focuses on expanding the granularity of our environmental claim detection model. We aim to categorize claims into subcategories like pollution or resource use, and further identify specific environmental aspects impacted. This can be achieved through a hierarchical taxonomy, domain-specific knowledge integration, and named entity recognition techniques. Furthermore, multi-label classification and sentiment analysis can offer a richer understanding of claims' complexity and potential impact. By pursuing these directions, we can empower our model to provide more granular and impactful insights into environmental claims, ultimately contributing to informed decision-making and progress toward sustainability.

References

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*.

Ahmed Al-Rawi, Derrick OKeefe, Oumar Kane, and Aimé-Jules Bizimana. 2021. Twitter's fake news discourses around climate change and global warming. *Frontiers in Communication*, 6.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-

worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829.

Pepa Atanasova, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *Preprint*, arXiv:1808.05542.

Alberto Barron-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. *Preprint*, arXiv:2007.07997.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. 2022. Saving earth one tweet at a time through the lens of artificial intelligence. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Tommaso Fornaciari, Dirk Hovy, Elin Naurin, Julia Runeson, Robert Thomson, and Pankaj Adhikari. 2021. "we will reduce taxes" - identifying election pledges with language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3406–3419, Online. Association for Computational Linguistics.

David Friederich, Lynn H Kaack, Alexandra Luccioni, and Bjarne Steffen. 2021. Automated identification of climate risk disclosures in annual corporate reports. *arXiv preprint arXiv:2108.01415*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Julian F Kölbel, Markus Leippold, Jordy Rillaerts, and Qian Wang. 2020. Ask bert: How regulatory disclosure of transition and physical climate risks affects the cds term structure. *Swiss Finance Institute Research Paper*, (21-19).

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020a. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020b. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Preprint*, arXiv:2202.12837.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, and Jens P Linge. 2022. Exploring data augmentation for classification of climate change denial: Preliminary study. In *Text2Story@ ECIR*, pages 97–109.

Tom Schaul, Sixin Zhang, and Yann LeCun. 2013. No more pesky learning rates. *Preprint*, arXiv:1206.1106.

Manjira Sinha, Nilesh Agarwal, and Tirthankar Dasgupta. 2020. Relation aware attention model for uncertainty detection in text. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 437–440.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. *Preprint*, arXiv:2209.00507.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Deep ordinal regression for pledge specificity prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1729–1740, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Saeid A Vaghefi, Christian Huggel, Veruska Muccione, Hamed Khashehchi, and Markus Leippold. 2022. Deep climate change: A dataset and adaptive domain pre-trained language models for climate change related tasks. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*.

Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. *arXiv preprint arXiv:2012.00483*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. *Preprint*, arXiv:2110.12010.

Vinicius Woloszyn, Joseph Kobti, and Vera Schmitt. 2021. Towards automatic green claim detection. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 28–34.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560.

Zhenxun Zhuang, Mingrui Liu, Ashok Cutkosky, and Francesco Orabona. 2022. Understanding adamw through proximal methods and scale-freeness. *Transactions on Machine Learning Research*.