Comparing the performance of Text Classification Models for climate change-related texts

Gorgi Lazarev, Sasho Gramatikov and Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia

gorgi.lazarev@students.finki.ukim.mk
sasho.gramatikov@finki.ukim.mk
dimitar.trajanov@finki.ukim.mk

Abstract—**This study aims to evaluate and compare the performance of two text classification models specifically tailored for classifying climate change-related texts. The models under investigation are ClimateBert Environmental Claims and ClimateBert Fact Checking, both of which are based on the ClimateBert model and available in the HuggingFace Hub. Our analysis focuses on the impact of fine-tuning these models using specific climate change-related datasets, as well as their performance without fine-tuning. We assess the models using various metrics, including accuracy, precision, recall, and F1 score, and identify the areas where they predominantly make classification errors. Through our findings, we highlight the significance of using these methodologies for the evaluation and comparison of climate change-related text classification models and to appropriately fine-tune the models with context-specific data to achieve optimal classification results.**

Index Terms—**Climate Change, Natural Language Processing (NLP), Text Classification, Models, Datasets**

I. Introduction

Climate change is one of the most challenging problems in our world, impacting not only the aspects of our lives today but potentially being a threat to ecosystems and the existence of various animal and plant species in the future [1]. Sustainability has been emerging as an increasingly important factor in the fight against these climate change threats [2]. This involves restructuring the activities we do to get our work done in the present days such that they do not compromise the lives of future generations. In the context of climate change, sustainability strives to preserve natural resources and habitats, reduce gas emissions and overall, maintain the state of the environment as we know it today. As climate change information is getting more and more prominent and the volume of climate change-related data continues to grow [3], it is important to be able to determine whether the information is really related to the topic of climate change or only appears to be. Text classification is one of the Natural Language Processing (NLP) tasks that can provide an effective and efficient solution to this problem.

A limited range of publicly available text classification models has been specifically fine-tuned for climate change-related texts. The ClimateBert [4] model is a pre-trained Language Model for Climate-Related Text which can be used for various downstream tasks such as text classification, sentiment analysis, mask filling, and fact-checking.

The authors of [5] advocate for leveraging social media in climate change awareness campaigns and introduce a fine-grained classification of climate-change-related social media text. They establish baselines using state-of-the-art contextualized word embeddings and a Reddit-based dataset in a semi-supervised setting.

The value of Twitter for mining public opinion on climate change is emphasized in [6], with geotagged tweets allowing for spatiotemporal evaluation. They apply text mining techniques, such as Latent Dirichlet allocation for topic modeling and VADER[1] for sentiment analysis, on a large dataset of tweets. Findings show negative sentiment in climate change discussions, especially in response to political or extreme weather events, and varying prevalence of discussion topics.

Overcoming limitations in real-time evaluation and scalability, a real-time sentiment analysis framework for social media posts related to smart city analytics was introduced in [7]. Utilizing a Bi-directional LSTM classifier on a dataset of 278,000 climate change-related tweets, the framework achieves high accuracies in discriminating between different emotions. The study highlights the role of geographic location, chosen topic, cultural sensitivities, and posting frequency in shaping public reactions to posts.

For the purposes of our research, we selected two fine-tuned classification models, the ClimateBert Environmental Claims [8] and ClimateBert Fact Checking [9] models. Both models are available on the HuggingFace Hub, and both are fine-tuned on the same base ClimateBert [4] model, but with different labels and a number of classes.

This paper is organized into five sections. In Section 2, we present the data used for our analysis, detailing the sources and preprocessing steps employed to obtain the climate change-related text corpora. Section 3 describes the methodology, which encompasses the text classification models and their fine-tuning process, as well as the evaluation metrics employed for comparison. In Section 4, we delve into the discussion,

[1] https://github.com/cjhutto/vaderSentiment

outlining the performance of the models under investigation and analyzing the impact of fine-tuning on their classification abilities. Finally, in Section 5, we draw conclusions from our findings, highlighting the implications of our research and potential for future work.

II. Datasets

This study aimed to compare the performance of selected text classification models on climate change-related texts, both with and without fine-tuning on similar datasets. We utilized datasets from research papers and the data.world website for our analysis.

A. Environmental Claims Dataset

The first dataset that was used for evaluation of the ClimateBert Environmental Claims [8] model is the Environmental Claims dataset which actually carries the same name as the model and represents an expert-annotated dataset for detecting real-world environmental claims made by listed companies. This dataset is specifically made for detecting real-world environmental claims as stated in the official paper for the Environmental Claims dataset. This dataset contains claims and a label indicating whether the claim is environmental or not. The labels are: 0 – IS NOT AN ENVIRONMENTAL CLAIM and 1 – IS AN ENVIRONMENTAL CLAIM. The main problem with this dataset is that is unbalanced, there are a lot more claims labeled with 0 than with 1.

B. Tweets Global Warming Dataset

This dataset is crowd-sourced (available here) and it consists of contributors' evaluated tweets that indicate whether the tweet believes in the existence of global warming or not. Along with the existence label, the dataset contains a label about the existence confidence, which indicates the level of confidence the claim is expressed with. This label contains values in the range from 0 to 1, where higher values indicate a stronger belief and represents the quantified belief of the tweet writer regarding the existence of global warming.

This dataset is unbalanced as well, it has a lot more claims marked as positive rather than negative. There are also inconsistencies in the column containing the Existence data since in some rows 'true' is marked with 'Yes' and in others with 'Y'. The same happens for 'false' which is marked both as 'No' and 'N'.

Additionally, there are many missing values in this column. There are only a few missing values in the confidence column. In the dataset, the rows with missing values for the existence column contain 1.00 as a value for the confidence column. To address this issue, a data preprocessing was done. To fix the inconsistencies, all positive values were changed to 'Y' and all negative to 'N'. The missing values in the existence column were filled in using the "most frequent" strategy and the missing values in the existence confidence column were filled in using the "mean" strategy. Both strategies fall into the category of Mean Imputation. Jadhav et al., 2019 in their paper [10] have compared the performance of different data imputation methods and have defined Mean Imputation as a common method for handling missing data, where missing values are replaced with the sample mean, median, or mode (the most frequent value) depending on the data distribution. However, it has drawbacks, such as changing the distribution shape and decreasing the standard deviation. Stratifying data into subgroups can slightly improve this method.

The labels were encoded such that 0 represented 'Y' and 1 represented 'N' accordingly. For this dataset there is no publicly available model, so we tested it on the ClimateBert Environmental Claims [8] model.

C. ClimaText Dataset

The ClimaText dataset [11] for climate change topic detection is the dataset that was used to fine-tune the base ClimateBert [4] model. It consists of sentences extracted from Wikipedia articles. More precisely, it contains id of a sentence, a sentence that represents an extract from a Wikipedia article, the title of that article, as well as the paragraph in which the sentence is located, and a label that indicates whether the sentence has a connection with climate change or not, assigning values 0 or 1, accordingly. There are no missing values in any of the splits.

D. Climate FEVER

The Climate FEVER dataset [12] is a dataset for verification of real-world climate claims. Data from this dataset was used to fine-tune and pre-train the ClimateBert Fact Checking [12] model. This dataset contains claims related to climate change, id-s of the claims, and labels for the claims that have the values: 0 – SUPPORTS THE CLAIM, 1 – REFUTES THE CLAIM, 2 – NOT ENOUGH INFORMATION, and 3 – DISPUTED. Every claim contains a list of evidence that validates the claim and every individual evidence has its own id, a label just like the claims which expresses the relation of that evidence to the claim, article of the evidence, the evidence sentence itself, entropy, and a list of votes for the evidence in terms of how it relates to the claim.

III. Methodolgy and Results

Fine-tuning the ML model can have a significant impact on its performance [13]. In this study, we employed a two-stage methodology to evaluate the performance of text classification models for climate change-related texts. The first stage involved assessing the models on various datasets without any fine-tuning on those specific datasets. In contrast, the second stage entailed fine-tuning the models on the datasets prior to evaluating their performance on the corresponding test splits. This approach allowed us to examine the impact of fine-tuning on model performance, as well as to identify the most effective strategies for improving text classification in this domain.

A. Evaluating models without fine-tuning

In the first stage, no additional training was done on either model, the datasets were loaded, and the model was evaluated directly on the test splits of those datasets.

After preprocessing the data by encoding the labels, filling missing values, renaming columns, etc., the dataset was split into training and testing sets, with 80% of the data used for training and 20% used for testing. Next, the models were loaded along with their corresponding tokenizers for the purpose of tokenizing and encoding the textual data in the datasets. This step was necessary to ensure that the encoded datasets could be effectively utilized in the models.

For running the model, TrainingArguments and Trainer classes were used. The Trainer as a class, included in the Transformers library, simplifies training HuggingFace transformer [14] models by automating many aspects of the process. This class comes with built-in features like logging and monitoring. TrainingArguments, on the other hand, is defined as a class that contains all of the hyperparameters that we can provide to the Trainer.

We first set up the Training Arguments with the specific parameters, and then we set the model, the training arguments, the tokenizer, and the datasets into the Trainer. In the next step, the training was skipped, and the models were directly used to predict the classes of the testing data. In the final step, the obtained predictions were used to evaluate the performance of each model on each dataset by comparing them with the actual values. This procedure was followed for both models and each of the datasets.

1) ClimateBert – Environmental Claims model tested on the Environmental Claims dataset without fine-tuning: The first dataset that was used to evaluate this model is the Environmental Claims dataset [8]. After running the model with this dataset, we obtained the results shown in Table I and the predicted classes shown in Table II.

TABLE I: Classification report for the Environmental Claims dataset using the ClimateBert Environmental Claims model without fine-tuning

class	precision	recall	f1-score per class
is not an environmental claim	0.96	0.85	0.90
is an environmental claim	0.62	0.88	0.72
accuracy	0.86		
f1-score	0.72258		

TABLE II: Confusion matrix for the Environmental Claims dataset using the ClimateBert Environmental Claims model without fine-tuning

Is it an environmental claim?	No	Yes
No	201	35
Yes	8	56

From the results, we can conclude that the model performs reasonably well on the dataset even without fine-tuning. It exhibits high precision in identifying non-environmental claims but struggles with classifying environmental claims, as evident from lower precision and f1-score in Table I. The overall accuracy is decent, largely influenced by the model's accuracy in predicting the first class.

2) ClimateBert – Environmental Claims model tested on the Tweets Global Warming dataset without fine-tuning: The model was initially evaluated on a familiar dataset, as it was pre-trained on other splits of the same dataset. Next, an unfamiliar dataset was chosen for binary label classification evaluation, with results shown in Table III and predicted classes in Table IV. However, the model's accuracy on this unfamiliar dataset is poor, as it often misclassifies tweets that discuss global warming as unrelated, resulting in very low precision for this class. The perfect precision of the other class is misleading, as it only reflects the accuracy of identifying tweets labeled as such, which was only 15 tweets in this case, and all of them were correctly classified.

TABLE III: Classification report for the Twitter Global Warming dataset using the ClimateBert Environmental Claims model without fine-tuning

class	precision	recall	f1-score per class
tweet doesn't talk about global warming	0.18	1.00	0.30
tweet talks about global warming	1.00	0.01	0.03
accuracy	0.19		
f1-score	0.16636		

TABLE IV: Confusion matrix for the Twitter Global Warming dataset using the ClimateBert Environmental Claims model without fine-tuning

Does the tweet talk about global warming?	No	Yes
No	215	0
Yes	988	15

3) ClimateBert – Environmental Claims model tested on the ClimaText dataset without fine-tuning: The ClimaText dataset [11] was the last dataset used to test the Environmental Claims model for climate change topic detection. The results showed that the model mostly classified the texts as unrelated to climate change, as seen in the confusion matrix VI. While the model performed average in classifying one class, it performed poorly in the other, with only a small portion of the data being classified. The high precision score in the latter class is misleading, as the model made only one incorrect classification of all the texts it classified into this class, but it is not accurate in predicting this class.

TABLE V: Classification report for the ClimaText dataset using the ClimateBert Environmental Claims model without fine-tuning

class	precision	recall	f1-score per class
text is not climate change-related	0.51	1.00	0.68
text is climate change-related	0.99	0.04	0.08
accuracy	0.52		
f1-score	0.37913		

TABLE VI: Confusion matrix for the ClimaText dataset using the ClimateBert Environmental Claims model without fine-tuning

Is the text climate change-related?	No	Yes
No	1912	1
Yes	1831	82

4) ClimateBert – Fact Checking model tested on the Climate FEVER dataset without fine-tuning: The only dataset used to evaluate the second model is the Climate FEVER dataset [12] which is a dataset for verifying real-world climate claims. The following results indicate that the model is not accurate without fine-tuning and predicts only one class 99% of the time, with the precision of the first class being misleading because out of all data that it classified into this class, it correctly predicted this class since data from this class is most frequent in the dataset.

TABLE VII: Classification report for the Climate_FEVER dataset using the ClimateBert Fact Checking model without fine-tuning

class	precision	recall	f1-score per class
Evidence supports the claim	0.43	0.99	0.60
Evidence refutes the claim	0.00	0.00	0.00
Evidence doesn't provide enough information	0.00	0.00	0.00
Disputed	0.00	0.00	0.00
accuracy	0.42		
f1-score	0.14908		

TABLE VIII: Confusion matrix for the Climate_FEVER dataset using the ClimateBert Fact Checking model without fine-tuning

Relation of the evidence to claim	Supports the claim	Refutes the claim	Not enough information	Disputed
Supports the claim	130	0	1	0
Refutes the claim	60	0	0	0
Not enough information	93	0	0	0
Disputed	22	0	1	0

B. Evaluating models with fine-tuning on the datasets

From what we can see from the results in the first stage when models predict on datasets with similar format and context to their pre-training data, they yield good results; otherwise, they tend to classify data into a single class. This outcome is expected, as models perform better with familiar data but may struggle with even similar topics due to the extensive nature of climate change discussions.

In the second stage, we aimed to evaluate the performance improvement achieved through fine-tuning the models on the training portion of the datasets. The methodology remained the same as in the first stage, with the addition of a fine-tuning step before generating predictions.

1) ClimateBert – Environmental Claims model tested on the Environmental Claims dataset with fine-tuning : The fine-tuning evaluation results show a slight improvement in precision for classifying environmental claims, as seen in Table IX. However, the precision for non-environmental claims remains almost perfect. Overall accuracy remains unchanged, with a slight increase in the f1-score due to the improved precision for classifying environmental claims.

TABLE IX: Classification report for the Environmental Claims dataset using the ClimateBert Environmental Claims model with fine-tuning

class	precision	recall	f1-score per class
is not an environmental claim	0.97	0.89	0.93
is an environmental claim	0.68	0.91	0.78
accuracy	0.89		
f1-score	0.77852		

TABLE X: Confusion matrix for the Environmental Claims dataset using the ClimateBert Environmental Claims model with fine-tuning

Is an environmental claim?	No	Yes
No	209	27
Yes	6	58

2) ClimateBert – Environmental Claims model tested on the Tweets Global Warming dataset with fine-tuning: After fine-tuning (confusion matrix XII), there was a significant improvement in accurately classifying tweets as class 1 (related to global warming), as seen in Table XI. Precision and f1-score for this class improved noticeably. There was also some improvement in precision for classifying tweets as class 0 (not related to global warming), but not as drastic. Overall accuracy improved and the f1-score indicates more accurate predictions.

TABLE XI: Classification report for the Twitter Global Warming dataset using the ClimateBert Environmental Claims model with fine-tuning

class	precision	recall	f1-score per class
tweet doesn't talk about global warming	0.74	0.55	0.63
tweet talks about global warming	0.90	0.95	0.93
accuracy	0.88		
f1-score	0.77745		

TABLE XII: Confusion matrix for the Twitter Global Warming dataset using the ClimateBert Environmental Claims model with fine-tuning

Does the tweet talk about global warming?	No	Yes
No	127	105
Yes	45	941

3) ClimateBert – Environmental Claims model tested on the ClimaText dataset with fine-tuning: After training, the model got very precise and accurate in the classification of climate change-related texts. (figures in Table XIII).

TABLE XIII: Classification report for the ClimaText dataset using the ClimateBert Environmental Claims model with fine-tuning

class	precision	recall	f1-score per class
text is not climate change-related	0.92	0.96	0.94
text is climate change-related	0.96	0.92	0.94
accuracy	0.94		
f1-score	0.93829		

TABLE XIV: Confusion matrix for the ClimaText dataset using the ClimateBert Environmental Claims model with fine-tuning

Is the text climate change-related?	No	Yes
No	1836	77
Yes	159	1754

4) ClimateBert – Fact Checking model tested on the Climate FEVER dataset with fine-tuning: After training, the model evenly distributes the data in each of the classes with good accuracy and precision. The less data available from the class, the less accurate is the model in predicting this class.

TABLE XV: Classification report for the Climate_FEVER dataset using the ClimateBert Fact Checking model with fine-tuning

class	precision	recall	f1-score per class
Evidence supports the claim	0.81	0.90	0.86
Evidence refutes the claim	0.75	0.82	0.78
Evidence doesn't provide enough information	0.84	0.60	0.70
Disputed	0.47	0.61	0.53
accuracy	0.77		
f1-score	0.71684		

TABLE XVI: Confusion matrix for the Climate_FEVER dataset using the ClimateBert Fact Checking model with fine-tuning

Relation of the evidence to claim	Supports the claim	Refutes the claim	Not enough information	Disputed
Supports the claim	118	0	6	7
Refutes the claim	0	49	5	6
Not enough information	20	14	56	3
Disputed	7	2	0	14

IV. Discussion

The summary of the results is presented in Table XVII.

A. ClimateBert – Environmental Claims model tested on the Environmental Claims dataset

The model performing decently even without fine-tuning is due to the fact that this model is already pre-trained on this dataset. However, it still does not give a satisfactory level of performance due to the fact that the dataset is unbalanced, and hence, it has learned how to predict one class with high precision, but struggles with the classification of the other class. As the model was already pre-trained on this dataset, further fine-tuning on a small portion of the same data does not notably improve the performance by much, as we can see from the classification report in Table IX and the confusion matrix in Table X. On top of that, our fine-tuning data is also unbalanced, so the model still makes mistakes when classifying into class 1 (the claim is environmental). The performance of the model on this dataset can be further improved by fine-tuning a balanced portion of the dataset so that the model can learn to classify the other class as accurate as the first one.

B. ClimateBert – Environmental Claims model tested on the Tweets Global Warming dataset

The results before fine-tuning seen in Table III are to be expected since this model is pre-trained on recognizing environmental claims, which is not exactly the same as recognizing the existence of a belief about global warming in a text. In the 80% of the cases, the model predicts only one class, i.e., it predicts that there is no existence of global warming in the tweets, and it makes an erroneous prediction because, as we already stated, it is pre-trained to recognize whether a given text is an environmental claim or not, so it classifies the tweets based on whether they are environmental claims or not.

Fine-tuning the model on this dataset wields significant improvement of the performance. Previously, the model was mostly predicting that the writer of the tweet does not believe in the existence of global warming, but now the huge leap in the performance of correctly classifying class 1 (the tweet talks about global warming) (confusion matrix XII) is indicating that the model had learned and grasped the indicators of the presence of global warming topic in the text.

C. ClimateBert – Environmental Claims model tested on the ClimaText dataset

From the classification report in Table V and the confusion matrix in Table VI, we can see that the model correctly predicts the first class half of the time. This is expected behavior since the model was pre-trained to predict whether a given claim is environmental or not, and the chances are that if the text is not about the environment of some kind, then, most likely, it is not about a climate change topic, either. This does not necessarily mean that the opposite statement is also true.

TABLE XVII: Table with summarized results

Model	Dataset	Accuracy		F1 - Score	
		Original	Fine-tuned	Original	Fine-tuned
ClimateBert Environmental Claims	Environmental Claims	0.86	0.89	0.723	0.778
ClimateBert Environmental Claims	Twitter Global Warming	0.19	0.88	0.166	0.777
ClimateBert Environmental Claims	ClimaText	0.52	0.94	0.379	0.938
ClimateBert Fact Checking	Climate_FEVER	0.42	0.77	0.149	0.717

After fine-tuning on this dataset, the model shows remarkable performance improvement (see classification report XIII and confusion matrix XIV). This can be attributed to the larger volume of data used in fine-tuning and the prior pre-training of the base ClimateBert model [4] on this dataset. The model significantly improved its ability to detect climate change topics in the text, beyond just environmental topics.

D. ClimateBert – Fact Checking model tested on the Climate_FEVER dataset

When this model was fine-tuned, the number of classes was reduced from 4 to 3, the DISPUTED class was dropped. Therefore, the model might have a problem with the predictions of this class.

After running the model without fine-tuning, we obtained results (see classification report VII) which show that, this model, when used on this dataset predicts only the first in 99% of the time (see confusion matrix VIII).

After additional fine-tuning of this model with data that consists of four classes, the model performs significantly better and is able to recognize whether the evidence support or refutes the claim, whether they give insufficient information, or whether the claim is disputed. The results (classification report XV and confusion matrix XVI) show that the model makes the least amount of mistakes with data it has seen the most, since the lower the accuracy of prediction in the specific class, the less amount of data is available in the dataset for that class.

V. Conclusion

In this research, we compared the performance of Text Classification models for Climate change-related texts when the models are used directly, only with their pre-training done, and when the models are trained on the dataset that is used to make predictions.

The results of the work showed the benefits of evaluating and comparing different text classification models for climate change-related texts and how they can be repurposed to be applicable not only in the smaller limited contexts in which they are trained but also in the larger whole in which they belong, in this case, the topic of climate change and appropriately fine-tuned to improve their performance.

With climate change as one of today's world's leading issues, more and more text classification models like these will get developed to help us when determining the connection of a text with the topic of climate change. The accuracy of these text classification models is critically important for the understanding of climate change-related text, and with that, the impact that climate change has on our current lives and the future. By accurately classifying these texts, we can understand the impact and problems that arise from climate change better. With that knowledge in hand, we can hopefully develop more effective plans to reduce and adapt to the impacts of climate change.

In future work, we plan to explore additional methods for enhancing the performance of text classification models for climate change-related texts. This may include investigating other pre-processing techniques, exploring different model architectures, and experimenting with ensemble methods. By further refining these models, we can deepen our understanding of climate change's consequences and implications, which in turn can inform the development of effective mitigation and adaptation strategies.

References

[1] W. Nordhaus, "Climate change: The ultimate challenge for economics," *American Economic Review*, vol. 109, no. 6, pp. 1991–2014, 2019.

[2] R. Goodland, "The concept of environmental sustainability," *Annual review of ecology and systematics*, vol. 26, no. 1, pp. 1–24, 1995.

[3] H. Hassani, X. Huang, and E. Silva, "Big data and climate change," *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 12, 2019.

[4] N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold, *ClimateBert: A Pretrained Language Model for Climate-Related Text*, 2022.

[5] R. Vaid, K. Pant, and M. Shrivastava, "Towards fine-grained classification of climate change related social media text," *10.18653/v1/2022.acl-srw.35*, vol. 7, 2022.

[6] B. Dahal, S. A. P. Kumar, and Z. L. T. modeling and, "and sentiment analysis of global climate change tweets," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, vol. 8, 2019.

[7] M. E. Barachi, M. AlKhatib, S. Mathew, and F. Oroumchian, "A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change," *Journal of Cleaner Production, 312, 127820*, vol. 9, 2021.

[8] D. Stammbach, N. Webersinke, J. A. Bingler, M. Kraus, and M. Leippold, *A Dataset for Detecting Real-World Environmental Claims.* Model available at, 2022. [Online]. Available: https://huggingface.co/climatebert/environmental-claims

[9] A. Konet, "Climatebert fact checking model," vol. 2022. [Online]. Available: https://huggingface.co/amandakonet/climatebert-fact-checking

[10] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Applied Artificial Intelligence, 33:10, 913-933*, 2019.

[11] F. S. Varini, J. Boyd-Graber, M. Ciaramita, and M. Leippold, , *title = ClimaText: A Dataset for Climate Change Topic Detection, year = 2021, archivePrefix = arXiv, eprint = 2012.00483.*

[12] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, *CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims*, 2021.

[13] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: from lexicons to transformers," *IEEE access*, vol. 8, pp. 131 662–131 682, 2020.

[14] S. M. Jain, "Fine-tuning pretrained models," in *Apress, Berkeley, CA*. Introduction to Transformers for NLP, 2022.