Course Report

IEOR 4742: Deep Learning for OR&FE
Subject: NLP Analysis on Corporate Climate Risk Disclosures
Lecturer: Ali Hirsa
Author: Dongxuan Wang, Hongyu Ji, Xinyu Zhang
Date: 12/28/2021

Table of Contents

1. Background and introduction

Climate risks are becoming increasingly important for financial institutions and investment banks to analyze and evaluate the performance of businesses over recent years. The reason that people are stressing more importance on those issues is that climate risks could affect their decision making processes via many aspects. People could change their investment decisions or their views to the companies due to something as minor as that the climate-related issues cannot be well addressed or something as significant as that the companies are changing their marketing strategies for the next five years along with their climate risk solutions. In order to be well prepared for future climate-related scenarios and be able to respond to stakeholder pressure, companies need to strategically identify climate-related risks and opportunities unique to their businesses, and then present their ideas and solutions through corporate annual reports, sustainability reports, financial filings, and voluntary reporting frameworks. Over the last couple of years, reporting the climate-related risks and opportunities through corporate annual disclosures has grown in relevance and importance.

While these disclosures are aiming at increasing the transparency of climate influences to guide investment decisions, their quality and materiality remain largely unclear. Compared to numerous studies on the quantitative aspects of these disclosures, there is not enough existing analysis in the quality area yet. Besides, that not all companies express their information in the same manner makes it even harder to compare the quality of company disclosures with each other. Therefore, if we can build some models that can take any piece of a report, process it, and then automatically answer some of the qualitative questions such as how good the climate disclosures are, whether they emphasize more on transition risks or physical risks, at what level they address those issues, then any further step can be easily taken to better understand the the climate risk part in the corporate annual reports without spending much time going through the reports line by line and therefore to make better investment decisions.

We would mainly continue on a similar project from last year. The team from last year mainly focused on replicating Bloomberg's environmental disclosure score for companies via their annual reports and sustainability reports. Firstly, we greatly appreciate them for their contribution to the process and construction of the corporate level corpuses. It always required a lot of effort to clean the data and reshape it in the form as we desired. Then they divided the analysis part into

two steps. They started with using LSA, LDA, Skip-gram and Doc2Vec methods and tuning the hyperparameters to extract features that could potentially better explain the quality level of climate-related corporate disclosures and followed by using TextCNN and XGBoost as their supervised learning models to replicate the disclosure scores. For our project, we would mainly work on identifying and classifying the type of climate risk in the company reports. We decided to shift gears completely to word embedding via BERT tokenizer and unsupervised text classification via similarity scores, which would allow us to identify the type of climate risk that the company reports were disclosing more accurately and to comprehend the climate risk disclosures more easily.

Our report would start with a literature review section to explain what existing relevant works have been done and what we have learned from those reference papers. Then we would discuss our data and assumptions of analysis as well as the methodologies we were using. In the methodology section, we would discuss both the supervised learning model that we were thinking of at first and the unsupervised learning model that we chose to implement in the end. We would discuss the findings and observations in detail followed by our preliminary conclusions as well as recommendations on future works.

2. Literature review

It is increasingly important for investors to assess the extent and quality of climate risk disclosures by companies and organizations over these years, and the first step of assessment would clearly be accurately identifying what type of climate risk the companies are disclosing. There have been many existing works in this field and both supervised and unsupervised learning algorithms could apply. In particular, the methodology in our project is motivated and closely related to some of them.

One example of using a supervised learning algorithm to identify disclosures of different types of climate risks by companies was recently published and developed by Friederich et al., (2021). They created their own labeled dataset by building a corpus of annual reports from websites of large companies and Refinitiv Eikon (Refinitiv Eikon, 2021) and annotated the paragraphs of the reports themselves. They mainly defined three different tasks, identifying whether each paragraph was disclosing climate risk or not, whether each paragraph was disclosing physical

risk or transition risk, and which of the five more fine-grained risk categories each paragraph was disclosing, and tested three different supervised learning models under these tasks, SVM classification model, DistilBERT text classification model, and RoBERTa Large text classification model. By evaluating the F1-score, they believed that the RoBERTa model achieved the best performance in classifying the paragraphs of reports and thus identifying the type of climate risk each report was disclosing.

Another model of identifying climate risk disclosure in company reports via a supervised learning algorithm was called ClimateBERT and was developed by Bingler et al., (2021). The algorithm was claimed to be the first context-based algorithm in this field and it mainly consisted of two steps. Firstly, a RoBERTa text classification model would be used to classify each sentence within each paragraph of the company reports. Then, the predicted classes of sentences within each paragraph would be aggregated to make the paragraph level predictions. The trivial way of aggregation would just be using the most frequently predicted class of sentences within the corresponding paragraph as the predicted class of the paragraph as well; however, Bingler et al., (2021) claimed that the imbalance of class sizes would fail this approach. Instead, they chose to construct a feature matrix with each row representing a paragraph and each column representing a class. Each entry of the feature matrix would just be the proportion of sentences classified into the class within the paragraph. By manually labeling the class of each paragraph, they were able to apply a logistic regression to the feature matrix to refine their paragraph level predictive model. By evaluating the precision, Bingler et al., (2021) claimed that the new algorithm achieved better performance than traditional BERT text classification models.

Although it was not common to identify the types of climate risk disclosure via unsupervised learning algorithms, there also had been existing works in the field of general text classification without using labels. Pietro (2020) proposed a method of using BERT, word embedding, and vector similarity to do text classification when we do not have labels for our texts. The classification process mainly consisted of three steps. Firstly, the text data had to be embedded into the vector space via a BERT model. Each word in the data would be represented by a vector and each document would be represented by the average of the vectors of the words within it. Secondly, a list of keywords would be defined for each class to explain the contexts within the class and embedded into the vector space via the same BERT model as well. Finally, the

similarity scores between every document vector and every keyword vector would be calculated and used to classify each document into the closest cluster.

Motivated by the idea of aggregating the sentence level predictions to make the paragraph level predictions in Bingler et al., (2021) and the idea of creating target clusters and using the similarity scores to classify texts when we do not have labels in Pietro (2020), what we were doing in our case was to combine the ideas and generate a new unsupervised algorithm to classify company reports to identify the types of climate risk each company was disclosing. Since we do not have access to the class labels of the company reports, we chose to use a similar method as what Pietro (2020) proposed. But instead of representing each report by the average of the vectors of the words within it, we decided to make a sentence level prediction within each report firstly via the method of similarity scores and use the proportion of sentences classified into each class as the prediction vector for each report. The methodology would also be discussed further in detail in the following sections.

3. Data and assumptions

The data we were using mainly came from the team working on a similar project last year. It mainly consisted of annual reports of companies, sustainability reports, and investor presentations. It was in the form of texts and there were in total 1420 observations after cleaning the null values. Since the team from last year had already done some preliminary cleaning on the texts, we started from a data frame where each row contained the company ticker and plain texts of the company report after removing punctuation and stemming. We then did some further cleaning on the texts. In particular, we removed the stopwords and lemmatized the words in the texts.

Ideally, each company report should have either a score to represent the quality of its disclosure of climate risk or a class to represent the specific type of climate risk it mainly disclosed. However, we did not have access to such labels of the texts and that made us unable to train our data via any supervised learning algorithms. Therefore, in order to identify the type of climate risk each report was disclosing, we assumed that all the texts could be well separated into three clusters and tried to classify each report into one of them via some unsupervised learning algorithms. We defined the three clusters according to the types of climate risk as ‘General Risk’,

'Physical Risk', and 'Transition Risk'. We were trying to analyze which type of climate risk each company report mainly disclosed and classify the corresponding report to its cluster. We also defined a list of keywords for each cluster to explain the contexts within the cluster and it will be further discussed in the following section.

```
## Define the keyword list for each cluster
general_words = ['climate','risk','carbon','emission','environment','social','governance','dioxide','sustainability',
                 'green','economy','energy','efficiency','clean','waste','recuperation','renewable','wind','hydro',
                 'solar','global','warming','hydrocarbon','methane','oil','coal','gas','greenhouse','recycling',
                 'composting','soil','water','air','pollution','natural','hazard']

physical_words =  ['heat','wave','cold','floods','droughts','wildfires','storms','change','precipitation','pattern',
                   'sea','level','rise','coastal','hazard','fluvial']

transition_words = ['paris','agreement','regulatory','kyoto','protocol','environmental','protection','agency']
```

4. Methodology

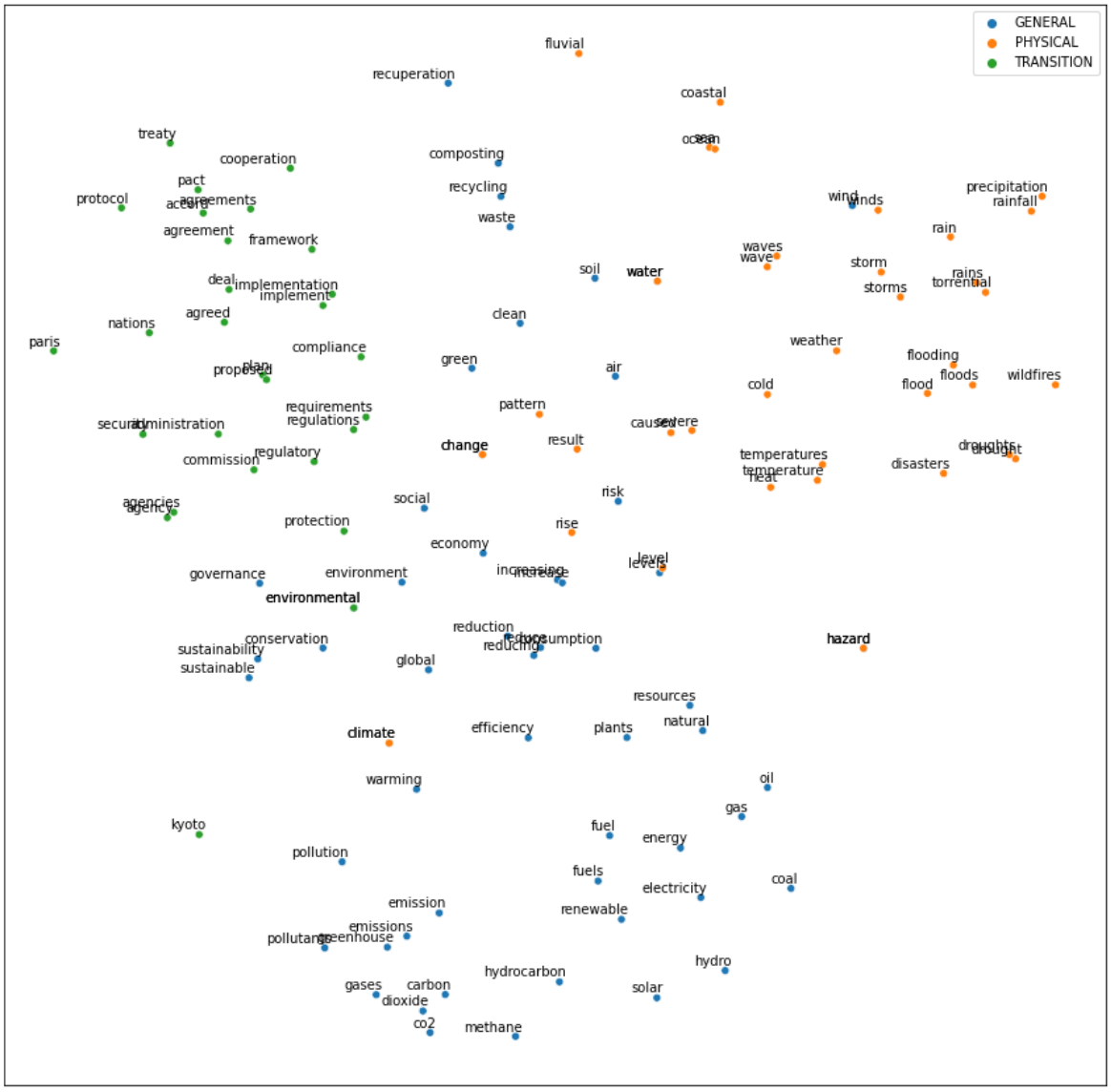
4.1 Supervised text classification via ClimateBERT

At first, we were trying to replicate the recent proposed ClimateBERT algorithm that was discussed in the literature review section since it was claimed to have better performance than traditional BERT methods in the climate-related text classification area. We were planning to use the RoBERTa text classification model to conduct a sentence level analysis and predict the class of risk for each sentence in the reports. Then we would construct a feature matrix as described in the literature with each row representing a paragraph, each column representing a class of risk, and each entry representing the proportion of sentences classified into the class of risk within the paragraph. A logistic regression could then be applied to predict the class of risk for each paragraph in the reports on the constructed feature matrix. However, the whole process required labels of type of climate risk for each sentence and each paragraph within the training set and we were not able to access those labels. Therefore, we decided to turn to an unsupervised learning model instead.

4.2 Unsupervised text classification via similarity scores

A common problem in unsupervised text classification was that the classes that the texts were classified into might not be as interpretable and meaningful as we would expect. For instance, under the setting of an unsupervised clustering problem, we might only come up with classes that the observations within each class were close to each other while outside each class were far from each other, but we might not know exactly what each class represented. Therefore, our first step to build the unsupervised text classification model was to create meaningful target clusters

and try to classify the texts into the created clusters by some similarity measures. As mentioned in the previous section, we assumed that the type of climate risk in all the texts could be well-separated into three classes as ‘General Risk’, ‘Physical Risk’, and ‘Transition Risk’. We then defined a list of keywords for each cluster to explain the context within the cluster and used the most_similar function in the gensim package to expand each list by another twenty similar words. The visualization of the keywords in the 2D space via PCA with two components was shown below. The keywords in the clusters of physical risk and transition risk were well-separated while some of the keywords in the general risk cluster overlapped with the keywords in the other clusters.



After creating the target clusters, we tried to embed the corpus and the target clusters into the same vector space via BERT tokenizer and TFBERT model. The reason that we used BERT here was that BERT could assign vectors to words by looking at the entire sentence of context instead of using a fixed vector embedding for each word. We applied BERT embedding to both the corpus and the target clusters. For the corpus, we applied BERT embedding to each report and generated one vector of 768-dimension for each word in the report. Then we averaged the word vectors in each report to generate one vector to represent each report. For the target clusters, we applied BERT embedding to each cluster and generated one vector of 768-dimension for each keyword in the cluster. Then we averaged the keyword vectors in each cluster to generate one vector to represent each cluster. Hence, we had 1420 word vectors to represent the 1420 reports and 3 keyword vectors to represent the 3 target clusters.

Since all the word vectors and the keyword vectors were in the same vector space, we could then compute the similarity scores between the vectors representing the reports and the vectors representing the target clusters via some similarity measures. We chose to use cosine similarity since it was commonly used as a measure to represent the similarity between vectors. Then we constructed a similarity matrix where each row represented a report, each column represented a cluster, and each entry represented the similarity score between the corresponding vector representing the report and the corresponding vector representing the target cluster. We furthermore normalized the matrix to make the row sums equal to one. Hence, we classified each report to the cluster that had the highest similarity score with it.

4.3 Sliding window approach of encoding

One major problem that we encountered when we were trying to embed the corpus into vectors was that the pre-trained BERT tokenizer could only encode a piece of text with a maximum length of 512 words at one time but most of the reports had more than that. This meant that we could not send each entire report to the tokenizer directly. One natural way to deal with it would be to cut each report in parts and send each part into the tokenizer directly, but it was also hard to determine how and where we should cut the reports. We did some research online and decided to go with the sliding window approach.

The sliding window approach would be automatically performed by the tokenizer if we chose the correct class and parameters to set it up. We chose the fast implementation version of the BERT tokenizer instead of the original BERT tokenizer to encode the texts, and defined the stride, return_overflowing_tokens, and max_length in the encoding function to make it work. Instead of encoding the entire report each time, the tokenizer would first divide the report into sliding windows of max_length and then encode each window one by one. We believed that doing a sliding window approach would be better than merely cutting the reports into normal windows because we could connect all the texts in a report more closely and understand each word in its context in this way.

4.4 Unsupervised text classification on sentence level

We also realized that it might not really make much sense to use the average of the word vectors in each report as the vector representation of the report since a report might contain thousands of words and averaging the word vectors might eliminate the characteristics of the report and make it even harder to identify the correct type of climate risk it was disclosing. Therefore, we decided to apply the idea of ClimateBERT into our model to deal with this issue.

Instead of using the sliding window approach to embed the texts, we firstly cut each report in sentences and embedded each of them. Since the data were plain texts with no mark of sentence, we chose to consider every 20 words as a sentence. We applied the original BERT embedding to every word in every sentence in each report and generated one vector of 768-dimension to represent each sentence by averaging the 20 word vectors in the sentence. Then we computed the similarity scores between the vectors representing the sentences and the vectors representing the target clusters in the form of similarity vectors. Instead of assigning each report to the cluster that had the highest similarity score with it, we decided to classify each sentence to the corresponding cluster firstly and calculated the proportions of sentences classified into each cluster in each report. For example, if a report had 20,000 words in total, then it would be cut into 1000 sentences first, each with 20 words. Then we would apply the original BERT embedding to the 1000 sentences and generate 1000 word vectors of 768-dimension. Then we would compute the similarity scores between the word vectors and the keyword vectors representing the target clusters to generate 1000 similarity vectors. We then would classify each sentence to the cluster that had the highest similarity score with it according to its similarity vector and thus generate

1000 classifications. If 200 of the sentences were classified into ‘General Risk’, 300 sentences were classified into ‘Physical Risk’, and the other 500 were classified into ‘Transition Risk’, the final output vector for this report would just be (0.2, 0.3, 0.5), with numbers representing the ratios. Since we had 1420 reports in data, our final classification would also consist of 1420 output vectors. The ratio vectors were already interpretable enough in the sense that it was easy to tell which type of climate risk a certain report was mainly disclosing by looking at the magnitudes of the ratios.

5. Findings and observations

The first step result we received from using the sliding window approach to encode the texts, averaging the word vectors within the reports to generate one vector to represent each report, and calculating the similarity scores between the vectors representing the reports and the vectors representing the target clusters was strongly imbalanced. A sample output classification for the first 50 reports was shown below.

```
['GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'TRANSITION', 'GENERAL', 'PHYSICAL', 'PHYSICAL', 'GENERAL', 'PHYSICAL',
'GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'TRANSITION', 'GENERAL', 'GENERAL', 'GEN
ERAL', 'GENERAL', 'GENERAL', 'PHYSICAL', 'GENERAL', 'GENERAL', 'GENERAL', 'PHYSICAL', 'TRANSITION', 'GENERAL', 'GENER
AL', 'GENERAL', 'PHYSICAL', 'PHYSICAL', 'GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'PHYSICAL', 'GENERAL', 'GENERAL',
'GENERAL', 'GENERAL', 'GENERAL', 'GENERAL', 'PHYSICAL', 'PHYSICAL', 'GENERAL', 'GENERAL', 'GENERAL']
```

Most of the similarity vectors of the reports had three very similar numbers. It was quite hard to identify the type of climate risk each report was disclosing by merely looking at the similarity vectors. If we assigned each report to the cluster that had the highest similarity score with it, most of the reports would be assigned into the general risk category but only a small portion would be classified into the other two categories.

The second step result we received from applying the idea of ClimateBERT and cutting the reports into sentences of 20 words, making classification for the sentences via similarity scores, and aggregating the classification of the sentences to generate a ratio vector for each report was also strongly imbalanced. A sample output classification for the first 50 reports was shown below.

```
['transition', 'transition', 'transition', 'transition',
 'transition', 'transition', 'physical', 'transition', 'physical',
 'physical', 'transition', 'physical', 'transition', 'general',
 'transition', 'transition', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'general', 'general', 'physical', 'transition', 'transition',
 'transition', 'general', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'general', 'general', 'transition', 'transition', 'transition',
 'transition', 'physical', 'transition', 'transition', 'general',
 'transition']], dtype='<U10')
```

It turned out that the magnitudes of the ratios in the ratio vectors for the reports were more differentiable from each other and it was usually easy to tell which type of climate risk each report was mainly disclosing by looking at its corresponding ratio vector. However, if we took a look at the highest ratio in each of the ratio vectors for the first 50 reports, 7 would be classified into the general risk category, 6 would be classified into the physical risk category, and the other 37 of them would all be classified into the transition risk category. We then expanded the analysis to all of the 1420 reports and it turned out that around 86% of them would be classified into the transition risk category.

The length of each sentence was also an interesting parameter to consider and we also had doubts that the imbalance of the classification result might be explained by the shorter sentence length. Therefore, we decided to expand the sentence length to 50 and did the same procedure for the first 50 reports again. A sample output classification was shown below.

```
['transition', 'transition', 'transition', 'transition',
 'general', 'transition', 'physical', 'transition', 'general',
 'physical', 'transition', 'physical', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'general', 'general', 'transition', 'transition', 'transition',
 'transition', 'general', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'transition', 'transition', 'transition', 'transition',
 'general', 'general', 'transition', 'transition', 'transition',
 'transition', 'physical', 'transition', 'transition', 'general',
 'transition']], dtype='<U10')
```

It turned out that there still existed a strong imbalance in the classification result. If we took a look at the highest ratio in each of the ratio vectors for the 50 reports, 8 would be classified into the general risk category, 4 would be classified into the physical risk category, and the other 38

of them would all be classified into the transition risk category as before. The classification result did not change a lot and the transition risk category still dominated it. There did not seem to be a clear link between the sentence length and the classification result. Moreover, viewing the result from another way, we believed that this method was also quite stable. In other words, when we changed the sentence length, the report-level classification would not change a lot.

6. Conclusion and future works

We initialized our work aiming at continuing on the project from last year to replicate Bloomberg's Environmental Disclosure Score, a measure of how thorough and informative a company was in disclosing climate-related risks, and then studied some literature about BERT and BERT-related algorithms in order to perform more advanced and appropriate regression analysis. However, since this direction would require labels for each report, each paragraph, or even each sentence, and we did not have enough access to those, we decided to switch to use unsupervised text classification to identify the type of climate risk each report was disclosing. We created our own target clusters, mainly used the pre-trained BERT model to encode the texts and embed the corpus, and then utilized either the sliding window approach or the idea of ClimateBERT algorithm to calculate the similarity scores between each report and each target cluster using cosine similarity. Then we classified each report to the cluster that had the highest similarity score with it.

There were several major conclusions that we had reached during the project. Firstly, we believed that the choice of keywords for each target cluster could significantly influence the classification result in the end. Secondly, we believed that instead of using the sliding window approach to encode the texts and averaging the word vectors in the reports to generate a vector to represent each report, intuitively it made much more sense to apply the idea of ClimateBERT algorithm, cut the reports into sentences, make classification for each sentence first, and then aggregate the sentence classification results to generate the classification for each report. Thirdly, although the classification results were strongly imbalanced, the length of sentences might not be a major factor of it.

Our work stopped here due to the lack of labeled data and time, but we did think of some directions of improvement for future references. Firstly, the choice of keywords for the target

clusters and the definition of clusters could be refined by doing frequency analysis on the texts. The assumption that all the texts could be well-separated into general risk, physical risk, and transition risk might not hold and more clusters could potentially be involved. Secondly, it was important to evaluate our model. Since we did not have enough access to the labeled data, we could just use some existing labeled texts such as the transition risk page of wikipedia and send it into our model to see how our model would classify it. Thirdly, if we could have access to the labeled data, then except for evaluating our model, another improvement we could do would be to use a logistic regression with the ratio vectors for the reports as the feature vectors to adjust our classification results. Since we did not have enough access to the labeled data, we could only apply the most straightforward way of using ClimateBERT. With enough labeled data, we might be able to incorporate the whole idea of ClimateBERT and we believed that the classification results could be improved as well.

7. Writer notes

7.1 Best and worst time

As for this project, we enjoyed discussing the methodologies to classify the reports most. The most tedious part was setting the environment on the Google Cloud Platform, which was extremely time consuming and troublesome. We spent 5 hours together on it but did not figure it out. We finally solved the problems and set up the environment using around one week.

7.2 Miscellaneous

It was very time consuming to embed the texts into word vectors and we spent more than 25 hours in total to embed the entire corpus via BERT. In order to save time for the groups that would continue the project based on our methodology, we had saved the embedded vectors for the sentences with lengths of 20 as 2D arrays(number of sentences * 768) for each report. They could be directly loaded into a jupyter notebook for future uses. We had also saved the ratio vectors for the reports for future comparisons.

8. References

Bingler, J. A., Kraus, M., Leippold, M. (2021). *Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures.*

Friederich, D., Kaack, L. H., Luccioni, A., Steffen, B. (2021). *Automated Identification of Climate Risk Disclosures in Annual Corporate Reports.*

Burck, J., Hagen, U., Bals, C., Höhne, N., Nascimento, L. (2020). *CCPI, Climate Change Performance Index. Background and Methodology.*

Pietro, M. D. (2020, September 7). *BERT for Text Classification with NO model training.* towards data science. https://towardsdatascience.com/text-classification-with-no-model-training-935fe0e42180

Alammar, J. (2018, December 3). *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*. https://jalammar.github.io/illustrated-bert/

CodeEmporium. (2020, May 4). *BERT Neural Network - EXPLAINED!* [Video]. YouTube. https://www.youtube.com/watch?v=xI0HHN5XKDo

Codebasics. (2021, July 22). What is BERT? | Deep Learning Tutorial 46 (Tensorflow, Keras & Python) [Video]. YouTube. https://www.youtube.com/watch?v=7kLi8u2dJz0