Online Appendix for "Real-time Transition Risk"

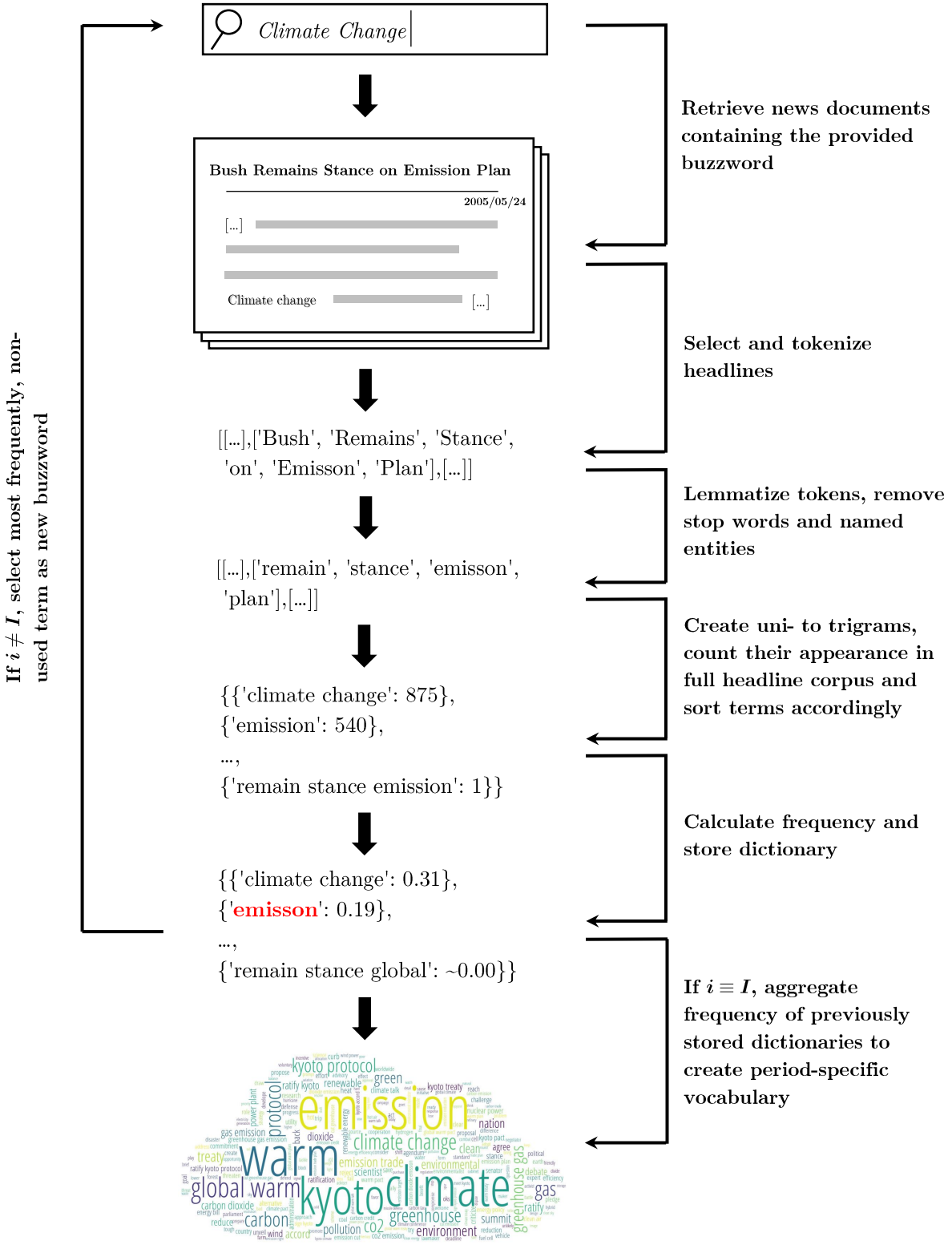
We divide the process of climate news index development into three parts: Domain-specific vocabulary construction (1), topic identification (2), and sentiment classification (3). The decision of applying either full-text news articles or headlines for text analysis is based on the use case and a trade-off between noise reduction (Nassirtoussi et al., 2015) and the risk of incongruency (Ecker et al., 2014).

1. Domain-specific vocabulary

In their paper, Engle et al. (2020) use a topic-specific dictionary to measure the similarity between a "Climate Change Vocabulary" and news articles in the Wall Street Journal (WSJ). To generate their topic-specific dictionary, the authors manually pick white papers and glossaries concerning climate change from 1990 to 2017 and select the most frequent terms from the merged text corpus. While this straightforward approach likely leads to the identification of the most relevant vocabulary related to climate change, it has two disadvantages. First, defining a dictionary containing information from the same sample period that is also used for the subsequent topic identification, results in a look-ahead bias. Time-dependent events and developments shape specific terms and vocabulary (e.g., "Kyoto protocol", "Fukushima", "Paris Agreement", "Green New Deal"). Given their approach, contextual terms will be considered for topic identification before they are mentioned in news media and thus deemed related to climate change. Secondly, time dependence can be a disadvantage itself within a stationary dictionary. By classifying news based on a dictionary that aggregates information from more than 20 years of data, time-dependent terms may be underrepresented during periods in which these terms were actually of increased relevance to the related topic.[1]

[1] We consider this to be an important aspect that could even improve the out-of-sample accuracy of a time dependent approach compared to a stationary approach that does not account for any look-ahead bias.

Figure 1: Iterative approach of vocabulary generation



Used for illustrative purposes only. Numerical values are exemplary and not real observations.

Instead of manually selecting topic-related documents for dictionary construction, we develop an unsupervised algorithm that utilizes the information from millions of unclassified news items. Figure 1 illustrates our approach for automated dictionary construction. The objective of the algorithm is the generation of a dictionary that strongly relates to a domain-specific buzzword. We provide the term "Climate Change" as input parameter.[2] Then, the algorithm selects news (from a given period) that contain the buzzword in the full-text article. The most frequent terms are calculated from the headlines of the selected news articles (in the following, referred to as the headline corpus). However, rigorous text normalization and cleaning raw headline data is required to generate a dictionary with a high degree of pureness in domain-specificity. At first, language-based stop words.[3] and text elements with less than three characters are removed from each headline text. Next, Named Entity Recognition (NER)[4] is performed to identify entity-related terms that should not be considered for dictionary construction. Specifically, names of persons, organizations, and locations are dropped as their inclusion may lead to an undesired bias towards entities and, hence, inaccuracy in the subsequent topic identification task. After lemmatization[5], terms are formed from a contiguous sequence of up to three text elements (unigram to trigram) for each cleaned headline. Next, we calculate the term frequency by counting the appearance of each term, divided by the total number of terms in the cleaned headline corpus. The frequency of terms allows us to compare the relevance of specific terms over dictionaries with different sample sizes. Finally, term frequency calculation is repeated for all unselected news items in the sample (period) to derive a list of frequently used terms in

[2] Generally, the number of provided buzzwords is unrestricted. A single expression refers to the minimum input required.

[3] Stop words are predominantly the most used words of a given language. Generally, stop words are removed because they are not relevant for dictionary construction and distort the word frequency analysis.

[4] Named Entity Recognition (NER) is the process of locating named entities in unstructured text and then classifying them into pre-defined categories. We use the application from the Stanford NLP Group (Stanza) for all NER tasks in this paper.

[5] Lemmatization is the process of reducing inflected forms of a word while ensuring that the reduced form belongs to the language. The initial word is stored to retrieve the original expression from the reduced form. These untrimmed words are needed for the content search process based on the full-text article.

general news. The most commonly used terms (e.g., "stock", "market", "rate") are automatically withdrawn from the topic-related dictionary as specific vocabulary is desired.[6]

This process of buzzword-based vocabulary generation may be repeated multiple times if the initial buzzword has synonyms or strongly related terms. The total number of iterations I defines the amount of most frequent terms that are applied for dictionary construction itself. Re-running the process of vocabulary generation by extending the list of considered, topic-specific buzzwords – by the most frequent and unused terms of the initial search process ($i = 1$) – results in more observations and possibly noise reduction. However, there is a trade-off to be considered as an extending number of iterations also leads to an increased risk of incorporating non-related or less specific buzzwords while diluting the impact of the initial news search process. Hence, the number of iterations may be chosen with respect to the news sample size of the initial search process. Generally, the smaller the number of observations and the more synonyms is available for the initial buzzword, the more iterations may be beneficial. Given multiple iterations, the algorithm stops after the last iteration ($i = I$) and generates the terminal, domain-specific dictionary from the equally weighted term frequencies of the I buzzword-specific dictionaries. The aggregation of multiple dictionaries and the selection of overlapping terms lead to noise reduction.[7]

We run the process of vocabulary generation with five iterations at the end of each year to create a climate-specific dictionary that is used for topic identification in the following year. Each dictionary consists of the period- and domain-specific terms and their respective frequency. The uni- to trigrams are sorted by term frequency. We limit the number of terms in a dictionary to the 500 most frequent expressions.

[6] We set the number of most commonly used terms that are removed from the buzzword-related dictionary to 500. Given our approach this step is expandable and only applied for illustrative purposes. We use *tfidf*-scores for topic identification, consequently unspecific and commonly used terms will be of low relevance anyway.

[7] A threshold of minimum appearances is applied on each term at the dictionary level. By doing so, certain terms that are frequently but almost exclusively used in headlines for only some of the considered buzzwords, get removed from the aggregated dictionary. We set the threshold at 0.5, requiring a term to appear in at least half of the buzzword-specific dictionaries.

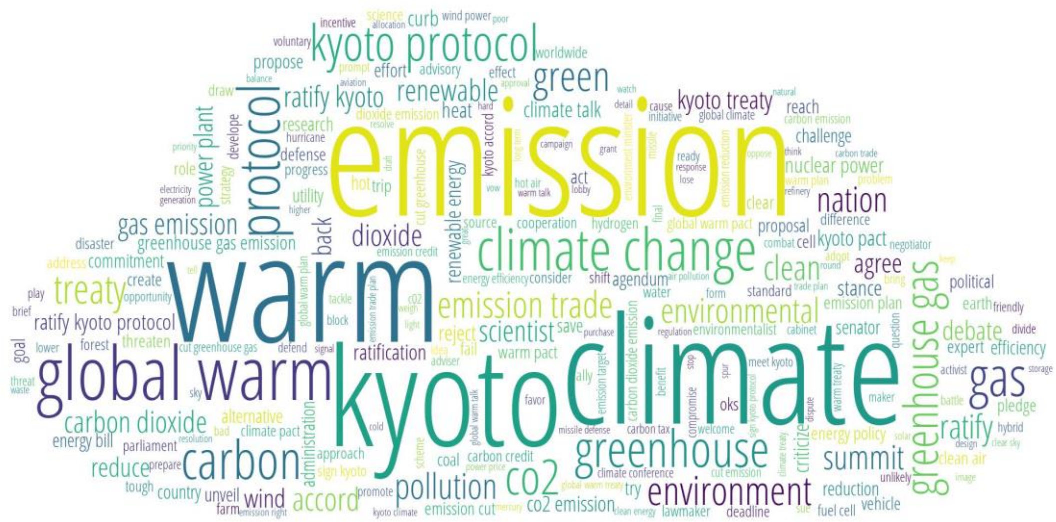
Afterward, the sum of term frequencies is scaled to 1. Finally, the resulting normalized term frequency (tf) represents the relative relevance of each term to the dictionary.[8]

The presented approach tackles the look-ahead bias and the problem of underrepresentation inherent in fixed dictionary construction. With topic-related vocabulary possibly changing over time, we generate period-specific dictionaries to account for the time-dependent relevance of terms. Each period-specific dictionary defines the topic-related vocabulary that is used for next year's news classification.[9] Figure 2 exemplary shows word cloud summaries of period-specific climate change vocabulary. The upper word cloud shows the vocabulary used for topic identification in 2006. The vocabulary of the second word cloud is generated from news data of 2019. The size of each term refers to its respective frequency. The figure illustrates how the relevance of vocabulary is varying over time for "Climate Change". While some terms seem to be of continued relevance over time (e.g., "Emission", "Environment"), other frequent terms of one period are not to be found in the dictionary of the other period. For example, "Kyoto" is a widespread expression in the early 2000's referring to the ratification and adoption of the Kyoto protocol at the beginning of the century. However, since its ratification, more than 15 years have passed. With other climate treaties having replaced the Kyoto protocol, it is of little surprise that the term is not found in the vocabulary based on climate-specific news data from 2019. On the other hand, expressions like "Green New Deal", "Extinction Rebellion" or "Sustainability" that strongly relate to the current climate debate were hardly used in the context of climate change-related discussions in the early 2000's. These findings indicate how topic identification and news index construction may benefit from a time dependent dictionary generation.

[8] The presented results are not sensible to the specific number of iterations in the vocabulary generation process.

[9] Available data history starts in the beginning of the year 2000. For the first period only, we use the same (in-sample) data for dictionary generation and for topic identification.

Figure 2: Word cloud summary of period-specific Climate Change Vocabulary



2005

2019

The figure shows word cloud summaries of period-specific climate change vocabulary. The upper word cloud utilizes news of 2005. The vocabulary of the bottom word cloud is generated from news in 2019. Term sizes are proportional to their frequency in the text corpus.

The main advantages of the presented approach are the opportunity to derive a domain-specific dictionary from news by providing as little input as a single term. Hardly any human intervention or supervision is necessary (with only a few parameters adjustable). This opens the possibility to extend this technique to different languages.

2. Topic identification

Based on the domain-specific vocabulary, we want to score unseen news articles by their relation to the topic of climate change. Therefore, we apply the domain-specific dictionaries to approximate the similarity between the climate-related vocabulary of year t and any news article text of year $t+1$, for all years $T-1$ in the data sample. We follow the approach of Engle et al. (2020) and use a score based on the "term frequency-inverse document frequency" (*tf-idf*), which is often applied in information retrieval and text mining. The *tf-idf* is composed by two functions: (i) the normalized term frequency (*tf*), which we derive directly from the domain-specific dictionary (ii) the inverse document frequency (*idf*), computed as the logarithm of the number of articles in the corpus (of year t) divided by the number of articles in which the considered term appears.

Certain terms, such as "energy" or "climate", are commonly used in news articles but convey no specific information. Independent of their relevance to the topic-specific dictionary (as measured by *tf*), these highly frequent expressions are penalized by multiplying with a low *idf*. Hence, the *tf-idf* identifies the most representative terms that appear infrequently overall but frequently in domain-related documents. We calculate the *tf-idf* for each term in the dictionary of year t. The *tf-idf* scores are stored in a $n \times 1$ vector $\mathbf{v}$, where n refers to the number of terms in the dictionary:

$$\mathbf{v}_{t+1} = \begin{pmatrix} tfidf_1 \\ tfidf_2 \\ \vdots \\ tfidf_n \end{pmatrix}_t$$

We normalize vector $\mathbf{v}$ so that $\mathbf{v} = {}^{\mathbf{v}}/_{\|\mathbf{v}\|}$ and $\|\mathbf{v}\| = 1$.

Next, we want to score news articles of the following period for their usage of climate-specific vocabulary. Therefore, we construct a $m \times n$ matrix $\mathbf{A}$, where n refers to the number of terms in the dictionary (of year t) and m is the number of news articles in the corpus of year $t+1$:[10]

$$\mathbf{A}_{t+1} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ddots & a_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

The matrix stores binary information on whether an article text contains a given term, with any $a \in \{0, 1\}$. Afterward, we perform multiplication of matrix $\mathbf{A}$ and vector $\mathbf{v}$ to calculate the sum of *tf-idf* scores over all terms for each news article in year $t+1$, resulting in a $m \times 1$ vector $\mathbf{w}$:

$$\mathbf{w}_{t+1} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \mathbf{A}_{t+1} \cdot \mathbf{v}_{t+1}$$

We define vector $\mathbf{w}$ to contain the article-based relevance scores with $0 \leq w \leq 1$. The relevance score w measures the weighted intersection in vocabulary between the article text and the domain-specific dictionary. A value close to 1 refers to the inclusion of the most representative terms in the article text. We implement two conditions for noise reduction in the process of topic identification. First, for a given news article to be considered (at least to some extent) climate-related we set a threshold of $w \geq 0.05$. The threshold for topic identification is selected by the inspection of the training sample and is based on a trade-off between sensitivity and specificity. Second, we require the article headlines to include one of the 20% most representative terms. By doing so, we want to exclude less specific article types like news summaries, highlights, and market roundups. Additionally, we reduce the risk of considering incongruent headlines in the following sentiment classification task. Ultimately, we derive a $k \times 1$ vector $\widehat{\mathbf{w}}$ that contains the

[10] The dictionary used for the first period is based on news data from the same period and part of the training sample.

relevance scores for all news articles in the data sample, defined to be climate-related by satisfying the mentioned conditions. The identified news articles will be subject of the upcoming sentiment analysis.

3. Sentiment classification

Sentiment analysis refers to the identification of the tonality of a given text document. Generally, sentiment analysis is performed by the application of either rule-based or machine learning methods. Rule-based approaches commonly use a pre-defined lexicon of emotionally charged (or domain-specific) terms to approximate sentiment by measuring the polarity towards "positive" or "negative" expressions. In contrast, machine learning approaches allow for the adaption and creation of language models for specific purposes and contexts by utilizing pre-labeled data samples that relate text elements to (human) perceived sentiment. Contextual, domain-specific sentiment classification significantly differs in terms of the objective. For example, sentiment analytics of financial news are usually constructed to measure a news event's expected financial (or subsequent return) effect on a mentioned stock.[11] Given our task at hand, we particularly want the sentiment to approximate the impact of a given news event on transition risk. Therefore, the sentiment model needs to be trained on a specific dataset that allows the model to infer the intended classification methodology concerning the objective from the provided text samples.[12] For this purpose, we choose a machine learning model based on Bidirectional Encoder Representations from Transformers (BERT), which is a popular choice for a wide range of domain-specific natural language processing (NLP) applications in

[11] Consequently, news analytics providers, such as Bloomberg, Thomson Reuters, RavenPack or Alexandria, provide clients with a variety of stock- and market-specific news sentiment measures related to different target variables like expected subsequent return impact, change in short-term volatility or investor perception. See Coqueret (2020) for an overview of existing research on stock-specific sentiment.

[12] We will particularly focus the differences of tonality-based and domain-specific sentiment classifiers to approximate transition risk when we evaluate our transition risk index in comparison with other existing approaches of news-based climate risk measurements in the following section.

economics.[13] BERT is a well-established, deep neural language model capable of learning word representations from large volumes of unannotated text (Devlin et al., 2018). Compared to earlier approaches that forward text input sequentially, BERT embeddings are highly contextual due to its bidirectional training. We use the off-the-shelf, base version of BERT, which was trained using English Wikipedia and the BookCorpus (Zhu et al., 2015), accounting for approximately 3,300M words.[14]

The model needs to be trained, i.e., fine-tuned on a labeled dataset to perform a sentiment classification task. For data sample creation, we select all news identified to be climate-related from 2000 to the end of 2008, accounting for more than 25,000 news headlines.[15] Each of the three annotators labels the extracted news items with respect to their implied impact on transition risk. To provide profound guidance on how to manually label climate-related news events for their inferred change in transition risk, we align our approach to widely accepted climate risk frameworks and existing academic research.

In 2017, the Task Force on Climate-related Financial Disclosures (TCFD) established a common framework for climate-related financial disclosures to support stakeholders in assessing the potential financial impacts of climate change on business activities. Since its release, the TCFD recommendations have been strongly supported as industry-standard on climate-related risk disclosure and incorporated into

[13] An early domain-specific adoption of BERT in the scientific field is "SciBERT". Based on large-scale labeled scientific data, Beltagy et al. (2019) create a pre-trained model based on BERT that leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve performance in sequence tagging, sentence classification and dependency parsing. To support computational analysis of financial language, Araci (2019) presents "FinBERT", a variant of BERT trained on the purely financial corpus of Reuters TRC2 to achieve domain language adaptation by exposing the model to financial jargon. Afterwards, the model was fine-tuned on the dataset of Financial Phrasebank for a sentence-based sentiment classification task, achieving higher test set accuracy than previous state-of-the-art models. To our knowledge, the application that comes closest to ours in terms of domain language adaption is "ClimateBERT". Bingler et al. (2021) design a contextual-based pre-trained model variant of BERT on thousands of sentences related to climate-risk disclosures aligned with the TCFD recommendations. By analyzing the disclosures of TCFD-supporting firms, the authors conclude that the firms' TCFD support is mostly cheap talk and that firms tend to cherry pick.

[14] The base version of BERT consists of 12 encoder layers, 768 hidden units, 12 attention heads, and a total of 110M parameters.

[15] We utilize a dictionary of climate-specific vocabulary. As a result, news about the physical effects of climate change is likely to be identified as climate-related. However, any news about physical risks in the training sample is labeled neutral with regard to the implied change in transition risk. Given a sufficient accuracy in out-of-sample sentiment classification, we expect no significant impact from news about physical risk on transition risk approximation.

various sustainability-related disclosure frameworks.[16] An essential element of these frameworks is the consistent categorization of climate-related risks and opportunities. We combine the insights from various climate frameworks (e.g., TCFD, 2017; NGFS, 2020; BCBS, 2021) and identify three different drivers of transition risk that organizations should consider: Policy and legislation (e.g., environmental and emission standards), technology (e.g., decrease of production costs for renewable energy), and climate awareness (e.g., shifts in consumer preferences). These drivers represent climate-related adjustments that could generate, increase or reduce transition risks via different transmission channels. We consider external pressure to be instrumental for a successful shift to a low-carbon economy and for the attempt to reach climate goals. Consequently, news events that infer an increasing transition risk are defined to be positive on climate. News events that imply a decrease in external pressure and respectively transition risk are deemed to be negative for achieving emission targets.

News samples are labeled accordingly. Each headline is assigned with the corresponding sentiment label resulting from the majority vote of the annotators. News data of the year 2008 is reserved as an out-of-sample test set, totaling almost 10,000 observations. The remaining data from 2000 to 2007 is used for model training and validation. Before starting the training process, two methods for model performance enhancement are applied to the initial data sample – *data augmentation* and *entity masking*.

3.1. Data augmentation

With annotated data for supervised learning tasks that remain generally scarce, *data augmentation* originated in computer vision to artificially increase the variety of data for model training without additional observations. Data augmentation for textual data is of particular interest when language from a different subject domain as the pre-trained model is used. A common approach to applying data augmentation on

[16] https://www.fsb-tcfd.org/supporters/

textual information is back-translation (Edunov et al., 2018). Given an input text in some source language A, the text is translated temporarily to a second language B before it is translated back into source language A. This process enables diverse samples to be generated that preserve the semantic meaning of the input text.

Table 1: Examples of back-translated headlines for different softmax settings

Value	Text sample
-	Kyoto protocol creates the climate for new ideas to cool down warming. Quest to lift efficiency, cut emissions sparks a shift in technological thinking.
0.5	Kyoto Protocol creates the climate for new ideas for cooling warming, and the pursuit of efficiency and emission reductions is transforming technological thinking.
0.6	Kyoto Protocol creates the climate for new ideas to slow warming. The drive to increase efficiency and reduce emissions leads to a change in technological thinking.
0.7	Kyoto Protocol creates the climate for new ideas on cooling warming. The pursuit of efficiency and emission reductions leads to a change in thinking about technology.

The table provides examples of back-translated headlines generated using a range of softmax temperature settings. The first example is the original text input.

We use the "fairseq" algorithm (released by Facebook AI Research)[17] for the English-German and German-English models from WMT'19[18] to perform back-translation on each headline in our training set (Ng et al., 2019). Synthetic texts are created by applying a random sampling strategy. We control the likelihood of low probability words being included in the generated sample with the so-called temperature of the softmax (Holtzman et al., 2019). A parameter value close to zero will likely result in samples identical to the original text, while a value of 1 results in highly diverse samples that risk altering the semantic meaning. We generate one augmented headline for each considered softmax setting (0.5, 0.6, 0.7) to yield enhanced variety without sacrificing fluency and coherence. Examples of back-translated headlines for different parameter values are provided in table 1. Duplicates resulting from exact back-translating are

[17] https://github.com/pytorch/fairseq
[18] http://www.statmt.org/wmt19/translation-task.html

dropped. Adding the remaining artificial data to the initial training set more than doubles the training sample size to almost 50,000 different news headlines.

3.2. Entity masking

Next, we confront the risk of overfitting in model training with *entity masking*. As for any classification task, the risk of overfitting arises from overestimating the polarity of specific terms in a text document for sentiment calculation. With a limited dataset of diverse samples, the generalization of particular terms that are expected to have no (standalone) impact on sentiment classification will reduce complexity. Terms or so-called entities we want to be neutral concerning their sentiment contribution are categorized as person names [PER], organizations [ORG], or locations [LOC].[19] To extract the sequences of words in the text that relate to one of the considered entity categories, we use Named Entity Recognition. The identified named entities are substituted with the entities' category label. By doing so, the complexity of non-essential information will decrease while retaining the semantics of the original text. Most importantly, an undesired sentiment bias towards specific entities is prevented. While the entity category (e.g., "organization") is of relevance for the sequential learning process, the entity name (e.g., "European Union" or "G20") should not. A given news event should be consistently classified irrespective of the specific entity value. For example, in the training sample news headlines containing the entity name "US" are negative, whereas headlines with the term "California" are significantly positive. These findings most likely result from the respective political agenda during the considered period. While the US administration under President George W. Bush was somewhat reluctant to adopt climate-friendly policies and refused to join the Kyoto Protocol, California under Senator Arnold Schwarzenegger followed a progressive environmental policy in the early 2000s. By labeling "US" and "California" with the same entity type (i.e., "LOC"), we hinder the

[19] Quantities, dates, monetary values, or percentages are also subject of entity masking. We do however differentiate between the relationship of values if there are multiple entities of the same type.

emergence of biases that may reduce out-of-sample performance - especially when previous biases switch (e.g., due to a change of political direction). Entity masking is performed for the whole data sample. With humans being as (unwittingly) prone to potential biases as trained models, only masked sentences are subject to the manual labeling process.

3.3. Predictive performance

We train different model specifications with respect to the adjustments applied to the input data. To evaluate the potential performance enhancement by the described data optimization techniques, we use the different model specifications to predict the sentiment label on the test set. Generally, the output of the classification model is the log odds for the different class labels: negative, neutral, and positive. We apply a softmax activation function to normalize the log odds into a probability distribution. Finally, any news item is assigned to the sentiment label with the highest predicted probability. We report results for the different model specifications in table 2.

Table 2: Performance of different model specifications on the test dataset

Model	Precision	Recall	F1 score
$BERT_{BASE}$	0.75	0.74	0.74
$BERT_{AUG}$	0.79	0.78	0.79
$BERT_{ENT}$	0.79	0.78	0.78
$BERT_{AUG-ENT}$	**0.82**	**0.82**	**0.82**

The table reports performance metrics of the different model specifications that are evaluated on the test set data. The presented model performance results from equally weighting the scores, calculated for each class individually.

All model specifications are run using identical parameter settings, based on the recommendations given for model configuration in the initial paper of Devlin et al. (2018).[20] We deliberately optimize classification performance solely on input data adjustments not on model configuration. By doing so, we want to choose

[20] For the sentiment classification task, we use a dropout probability of p = 0.3, a maximum sequence length of 64 tokens, a learning rate of 2e-5 and a batch size of 16. We train the model for 6 epochs and choose the iteration with the highest accuracy for the validation set.

the most straightforward approach, confront concerns with respect to design choices and limit complexity. Three performance metrics are considered for the comparison of model specifications: (i) Recall, or sensitivity, is the proportion of actual positives that are correctly predicted positive, (ii) precision, also referred to as positive predictive value, denotes the proportion of predicted positives that are true positives, and (iii) the F1 score, which is defined as the harmonic mean of precision and recall. For a classification task with more than two labels, model performance is calculated by averaging the scores of the individual classes to account for imbalanced classes (Koyejo et al., 2015). All three metrics are commonly used to evaluate the accuracy of machine learning-based classification models.[21]

Table 3: Confusion matrix

		Predicted sentiment		
		Positive	Neutral	Negative
Actual sentiment	Positive	7.4	1.8	0.2
	Neutral	2.0	76.6	2.3
	Negative	0.4	2.1	7.1

The table shows the confusion matrix for the fine-tuned $BERT_{AUG-ENT}$ model and the test dataset.

We define the pre-trained base version of BERT ($BERT_{BASE}$) as a benchmark for performance evaluation. $BERT_{BASE}$ achieves a F1 score of 0.74, which is a considerable performance for a baseline approach, probably due to the already high number of actual training observations. By increasing the size of the training data set with augmented data, model accuracy ($BERT_{AUG}$) improves to 0.79. The

[21] See James et al. (2013), pp. 145ff.

performance is comparable to the one of the fine-tuned models ($BERT_{ENT}$) that is based on the data set with entity masking. Both approaches are outclassed by the fine-tuned model ($BERT_{AUG-ENT}$) that applies both methods for performance enhancement. $BERT_{AUG-ENT}$ achieves an impressive F1 score of 0.82. Overall, our results align with previous findings for textual data augmentation (e.g., Nugent et al., 2020).[22] With an increased diversity of examples using domain-specific language, generalization improves in line with model accuracy.

For further evaluation, we report the confusion matrix for $BERT_{AUG-ENT}$ in table 3. By inspecting the off-diagonal values, we see that only a small fraction of misclassified observations results from mistaking positive for negative sentiment and vice versa. Approximately 95% of the failures happen between labels positive and negative. These findings make intuitive sense as it is easier to differentiate between positive and negative than between positive and neutral or neutral and negative. Just as for human decision-making, the difference in boundaries can be marginal for specific observations.

Ultimately, we use $BERT_{AUG-ENT}$ as the fine-tuned model for the out-of-sample sentiment classification.

3.4. Sentiment score

Sentiment classification is performed on all news that are identified to be climate-related (see section 2). This results in a k-vector **SENT** that contains the sentiment score for each considered news item, where k refers to the number of climate-related news articles in the data sample:

$$\mathbf{SENT} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{pmatrix},$$

[22] Nugent et al. (2020) apply back-translation to improve the accuracy of a fine-tuned BERT model for ESG controversy classification.

with any $s \in \{-1, 0, 1\}$ for either negative, neutral or positive sentiment. The vector **SENT** and the vector $\widehat{\mathbf{w}}$ (containing the relevance scores) have the same dimension. We calculate the element-wise product of vector $\widehat{\mathbf{w}}$ and vector **SENT** to calculate the weighted sentiment score for all climate-related news articles:

$$\mathbf{wSENT} = \begin{pmatrix} w_1 s_1 \\ w_2 s_2 \\ \vdots \\ w_k s_k \end{pmatrix} = \widehat{\mathbf{w}} \odot \mathbf{SENT},$$

with $-1 \leq ws \leq 1$. We use the weighted scores to emphasize the article's relevance to the topic of climate change. The weighted sentiment scores lay the foundation for the calculation of our Transition Risk Index (TRI).

3.5. Index calculation

We run index calculations on a daily, weekly, and monthly frequency. The start and end times of each period are based on the NYSE trading dates and hours. News that is released after closing are considered in the next period. This includes news on weekends and holidays that are ascribed to the next trading day. The index score results from the simple aggregation of sentiment for climate-related news articles, divided by the total number of news articles n. Given the weighted sentiment score $wSENT$ for climate-related news articles $k = 1, \ldots, K$, the index *score* of a given period p is defined as:

$$score_p = \frac{\sum_{k=1}^{K_p} wSENT_{k,p}}{n_p}$$

References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. Working Paper.

BCBS (2021). Climate-related risk drivers and their transmission channels. Basel Committee on Banking Supervision.

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. Working paper.

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

Coqueret, G. (2020). Stock-specific sentiment and return predictability. Quantitative Finance, 20(9), 1531-1551.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Working paper.

Ecker, U. K., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. Journal of Experimental Psychology: Applied, 20(4), 323.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. Working paper.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3), 1184-1216.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. Working paper.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. New York: Springer.

Koyejo, O., Natarajan, N., Ravikumar, P., & Dhillon, I. S. (2015). Consistent Multilabel Classification. Advances in Neural Information Processing Systems, 29, 3321-3329.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems, 104, 38-48.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications, 42(1), 306-324.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 News Translation Task Submission. Working paper.

NGFS (2020). Guide for supervisors: integrating climate-related and environmental risks into prudential supervision. Network for Greening the Financial System.

Nugent, T., Stelea, N., & Leidner, J. L. (2020). Detecting ESG topics using domain-specific language models and data augmentation approaches. Working paper.

Task Force on Climate-related Financial Disclosures (2017). Recommendations of the Task Force on Climate-related Financial Disclosures (TCFD).

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE international conference on computer vision 2015, 19-27.