climateBUG: A data-driven framework for analyzing bank reporting through a climate lens

Citation for the original published paper (version of record):
Yu, Y., Scheidegger, S., Elliott, J. et al (2024). climateBUG: A data-driven framework for analyzing bank reporting through a climate lens. Expert Systems with Applications, 239.
http://dx.doi.org/10.1016/j.eswa.2023.122162

(article starts on next page)

Expert Systems With Applications 239 (2024) 122162

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

climateBUG: A data-driven framework for analyzing bank reporting through a climate lens

Yinan Yu [a,b], Samuel Scheidegger [b], Jasmine Elliott [c], Åsa Löfgren [d,*]

[a] Department of Computer Science and Engineering, Chalmers University of Technology, Sweden
[b] Asymptotic AI, Sweden
[c] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden
[d] Department of Economics, University of Gothenburg, Sweden

ARTICLE INFO

Dataset link: https://www.climatebug.se/

Keywords:
Natural language processing
Annual reporting
Climate change
Sustainability
Finance & accounting

ABSTRACT

This paper applies computational linguistics learning methods to the banking industry and climate change fields. We introduce our data-driven framework, climateBUG, with the aim of detecting latent information about how banks discuss their activities related to climate change using natural language processing (NLP). This framework consists of an ingestion pipeline, a configurable database, and a set of API's. In addition, climateBUG offers two standalone components, namely a unique annotated corpus of approximately 1.1M statements from EU banks' annual and sustainability reporting and a deep learning model adapted to the semantics of the corpus. When benchmarking on classification performance, our model outperforms other models with similar scopes due to its stronger domain relevance. We also provide examples of how the framework can be applied from a user perspective.

1. Introduction

This article bridges the domains of the banking industry and climate change through the lens of computational linguistics. We introduce climateBUG (climate model for Bank reporting analysis from the University of Gothenburg), a comprehensive data-driven framework equipped with an ingestion pipeline, a configurable database, and a suite of APIs. The framework stems from an interdisciplinary approach, drawing on pertinent domain knowledge from annual reporting, climate economics, and advanced computational linguistics in Natural Language Processing (NLP). Uniquely, climateBUG offers two standalone components: a unique annotated corpus of approximately 1.1 million data points drawn from the annual reports of EU banks, with a focus on climate change and finance, and a sophisticated deep learning model tailored to the semantics of this corpus. Designed with versatility in mind, climateBUG and its components can be readily employed by researchers and practitioners to uncover latent information on how banks articulate their climate-related activities. This is achieved through the examination of unstructured data extracted from banks' annual reports. To showcase the utility of this framework, we also provide some examples of potential applications of the system.

Developing a comprehensive data-driven framework with a focus on finance and climate change is important for several reasons. Firstly, the banking industry is critical for providing a significant part of the finance for climate investments necessary to reach the climate net zero target by 2050. The International Energy Agency together with the International Monetary Fund estimate that a global annual energy investment of USD 5 trillion (more than a tripling of current levels) is needed by 2030 to reach the climate net zero target (Bouckaert et al., 2021). In addition, climate change can also affect the stability of the banking system and jeopardize global financial stability if banks do not correctly assess the climate-related risks (both physical and transition) of assets and exposure to carbon-intensive industries (Lamperti, Bosetti, Roventini, & Tavoni, 2019). Still, the availability of transparent quantitative data to evaluate the vulnerability of banks to climate-related risks and to track their efforts in mitigating their exposure over time is limited. This limitation of data availability has created a supply-side effect from regulators that demand banks (as well as public companies) to disclose key financial information but also climate-related risks. One such example is the Sustainable Finance Disclosure Regulation (SFDR) (2019/2088) in the European Union, which came into effect in March 2021 (European Parliament, 2019). We anticipate that regulations like the EU SFDR will have an impact on how banks discuss sustainability and climate change. To address parts of the information gap, it will be important to monitor banks' disclosures under these regulations.

* Corresponding author.
E-mail addresses: yinan@chalmers.se, yinan.yu@asymptotic.ai (Y. Yu), samuel.scheidegger@asymptotic.ai (S. Scheidegger), jasmine.christine.elliott@gu.se (J. Elliott), asa.lofgren@economics.gu.se (Å. Löfgren).

https://doi.org/10.1016/j.eswa.2023.122162
Received 6 December 2022; Received in revised form 13 October 2023; Accepted 13 October 2023
Available online 29 October 2023

This leads us to the second argument as to why the application in this paper is important — the increase in unstructured data from more financial reporting over time driven by both a supply-side as well as a demand-side effect (Lewis & Young, 2019). This increase in unstructured data has spurred researchers within the finance and accounting literature to use computational linguistics learning methods to analyze financial reports and other text sources to detect latent information.[1] In general, the field of NLP and computational linguistics has evolved from early methods that relied on simple rule-based systems and keyword matching, to statistical models, such as Naïve Bayes classifiers (Rish et al., 2001), bag-of-words analysis (Harris, 1954; Zhang, Jin, & Zhou, 2010), Hidden Markov Models (Conroy & O'Leary, 2001), and probabilistic context-free grammars (Jelinek, Lafferty, & Mercer, 1992). Later developments have evolved towards machine learning approaches, where algorithms learn patterns from data rather than relying on hard coded rules. This transition has led to the introduction and wide adoption of neural networks and deep learning techniques (Brown et al., 2020; Devlin, Chang, Lee, & Toutanova, 2018; Vaswani et al., 2017), which now play a central role in many state-of-the-art solutions. Overall, deep learning models are more capable in extracting complex semantic information from texts compared to classical techniques.

When considering the accounting and finance literature and its application of NLP and computational linguistics methods, El-Haj, Rayson, Walker, Young, and Simaki (2019) conclude that the main limitation of studies within this field is the common use of basic techniques like bag-of-words content analysis. They argue that this does not capture the intricacies of language, particularly context and the multiple meanings a word can have. Also, they note that studies often lack transparent evaluations of methodologies. Finally, the authors emphasize the complementarity between computational linguistics methods and high-quality manual analysis when analyzing financial research questions. The research within finance and accounting using computational linguistics methods is hence still under development and in light of this, there has also emerged a literature that review text analysis methodologies with the aim of facilitating researchers in adopting best practices when utilizing these methods (see e.g. Benchimol, Kazinnik, & Saadon, 2022). While enhanced expertise in applying these methods is required, there are other challenges that can hinder effective use of NLP methods as applied to finance and accounting.

One of the more prominent challenges is a lack of interdisciplinary research teams (Lewis & Young, 2019). The authors recognize that "applying NLP to financial reporting output is an inherently interdisciplinary process requiring the marriage of domain expertise from financial reporting with advanced NLP skills from computational linguistics. Neither discipline is capable of delivering step-change on its own" (Lewis & Young, 2019, p. 605). Another challenge we identify is the limited availability of annotated domain-specific data sets, which in turn relates to the absence of well-defined financial lexicon lists (Gupta, Dengre, Kheruwala, & Shah, 2020). To our knowledge there are currently no open access domain (finance and climate) specific annotated data sets available.

With this study we aim to respond to several of the challenges outlined above. Based on previous manual analysis of how banks discuss climate change and sustainability (Elliott & Löfgren, 2022), we are able to offer a data-driven framework – climateBUG – with the overall objective of using NLP techniques to search, classify, and summarize bank reports using modern deep learning techniques to better understand banks' narratives around climate change and their related activities by extracting relevant information from annual bank reports in a structured and scalable way. The framework climateBUG is *human-centric* in the sense that despite having automation as its main functionality, the construction is heavily influenced by the knowledge of domain experts. The primary objective of its outcome is to be interpretable by a broad range of end users including academia, government representatives, journalists, and commercial banks' sustainability managers. We contribute to the research community by providing a trained NLP model open access, a finance and climate corpus based on annual and sustainability reports from EU commercial banks, a database with annotated data that can be used as training data, and expert-driven keywords offering a domain specific dictionary (see Appendix C). Most importantly, climateBUG is optimized based on an iterative process between manual annotation and model optimization, focusing specifically on banks' annual reporting language related to sustainability and climate change. Finally, the performance of the framework and its components is evaluated and transparently reported.

The outline of the article is as follows; we provide an overview of the climateBUG framework and its component details in Section 2; Section 3 presents the development of the data derived from our corpus (climateBUG-Data); Section 4 presents the development of the trained NLP Model (climateBUG-LM); Section 5 introduces applications of the framework and additionally provides a simple step-by-step instruction on how to use the framework for more tailored analysis; and Section 6 concludes.

2. The climateBUG framework

In this section, we provide an overview of the framework and discuss each component and module in detail.

2.1. System overview

The climateBUG framework consists of three main modules that users can interact with (see illustration in Fig. 1):

(1) An **ingestion pipeline** to extract information from a corpus data consisting of statements from annual reports and sustainability reports published by commercial (EU) banks;
(2) A configurable database, **climateBUG-DB**, containing advanced statistics extracted using the ingestion pipeline;
(3) A set of APIs, **climateBUG-API**, to query and visualize information from climateBUG-DB.

In addition, climateBUG provides two standalone components that users can utilize for customized analyses:

- **climateBUG-Data**: an annotated corpus that focuses on both finance and climate. The corpus data consists of statements from annual reports and sustainability reports published by commercial EU banks from 2015- 2020 (discussed further in Section 3);
- **climateBUG-LM**: a deep learning model adapted to the domains of finance and climate (discussed further in Section 4).

For these two standalone deliverables of climateBUG, users can access and query information with commonly used deep learning programming interfaces such as the Huggingface API.

These components and their respective deliverables are summarized in Table 1. Note that although climateBUG-Data is part of climateBUG, it is not marked as a system component since its primary contribution is to train climateBUG-LM.

[1] It is also worth pointing out that there are important applications in finance beyond looking at financial reports using language modeling and deep learning. Prominent examples include constructs of an index based on the frequency of specific words in news coverage of major US newspapers to study the effect of policy uncertainty on firm-level and aggregate economic outcomes (Baker, Bloom, & Davis, 2016). Other examples include Chen, Wu, and Wu (2022) in which the authors use a deep learning approach to predict banks' stock prices, Hilal, Gadsden, and Yawney (2022) who provides a survey of computational linguistic methods to detect financial frauds, and studies focusing on central banks' communications and reports such as Benchimol, Caspi, and Kazinnik (2023) and Correa, Garud, Londono, and Mislang (2021).

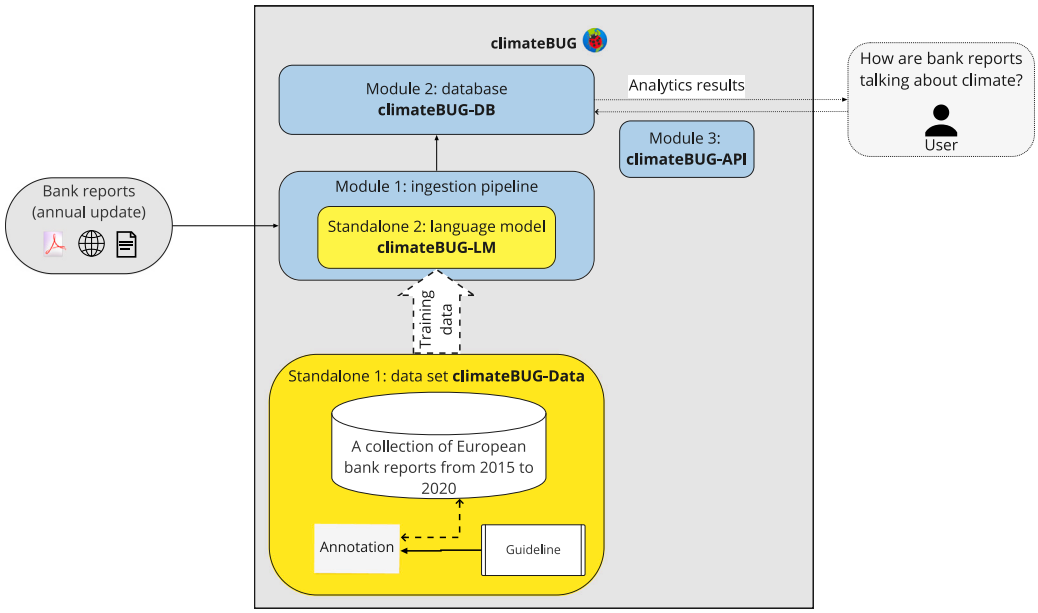

Fig. 1. System overview of climateBUG.

Table 1
Deliverables of this project.

Component	Description	Publicly available deliverables	Type
climateBUG-LM	A language model adapted to climate and finance	A deep learning trained model	Standalone, system component
climateBUG-Data	An annotated corpus of sampled European bank reports from 2015 to 2020	1115M annotated statements (917K automated annotations and 198K manual annotations)	Standalone
climateBUG-DB	A database that contains advanced statistics about how bank reports talk about certain expert-driven keywords	A database	System component
climateBUG-API	A set of functions used to query data from climateBUG-DB	API functions	System component

2.2. Ingestion pipeline

In order to populate climateBUG-DB, a pipeline is executed to collect, analyze and extract information from bank reports. This pipeline, shown in Fig. 2, is called the *ingestion pipeline*, where *ingestion* refers to the action of importing information to a persistent database for future downstream analyses. Within this pipeline, there are two ingestion steps: the *data ingestion step* and the *statistics ingestion step*.

2.2.1. Step 1: data ingestion

Bank reports (typically in PDF format) are manually downloaded and fed to the data ingestion step. First, these reports are parsed into machine-readable strings of characters. These strings are then split into *statements*, where each statement is essentially a sentence. Statements are then passed into a tailor-made deep learning language model, **climateBUG-LM**, that is adapted to the domains of finance and climate. The purpose of climateBUG-LM is to automatically identify and remove a statement if it is not about climate-related subjects.

In the scope of this paper, the annotated corpus, climateBUG-Data contains annual financial and non-financial reporting from 2015–2020. However, in practice, the data set and the corresponding language model can be updated if new bank reports and annotations are added to climateBUG-Data. The development of climateBUG-Data and climateBUG-LM are discussed in Sections 3 and 4, respectively.

2.2.2. Step 2: statistics ingestion

In this step, climateBUG applies text analysis tools (e.g. full-text search, word count, word cloud generation, etc.) and deep learning based semantic analysis methods (e.g. keyword extraction, latent feature vector analysis, etc.) to extract and analyze statistics to populate the climateBUG-DB.

Statistics ingestion is configurable, meaning that users can modify the focus of the analysis given their specific context of interest. A number of possible user configurations in this step are described below.

A. User configurations. There are two input parameters that the user needs to provide before executing the statistics ingestion step: the *partitioning strategy* and *expert-driven keywords*.

Partitioning refers to the action of dividing a set of statements into disjoint subsets. Each subset is called a *partition*. More specifically, given a set of statements $\mathcal{S}$, a partition $\mathcal{P}_i$ is a subset of $\mathcal{S}$ with the following properties:

- $\mathcal{P}_i \neq \emptyset, \forall i$
- $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset, \forall i \neq j$
- $\mathcal{S} = \bigcup_{\forall i} \mathcal{P}_i$

A partition provides a semantic context within a subset of statements $\mathcal{S}$. For instance, one partitioning could be to divide $\mathcal{S}$ into an environmental, social and governance, (commonly referred to as ESG) related context (partition 1) and an "other" context (partition 2). Each of these partitions then provides a semantically meaningful context for the analysis.

The partitioning is produced by a *partitioning strategy*. A partitioning strategy is essentially a function that takes a statement as its input and produces a categorical value that represents to which partition this statement belongs. Note that the partitioning strategy will only be applied to climate-related statements since climateBUG has filtered out the non-climate statements before this step. Therefore, the partitioning strategy will be applied to the output of the see Section 3.3.

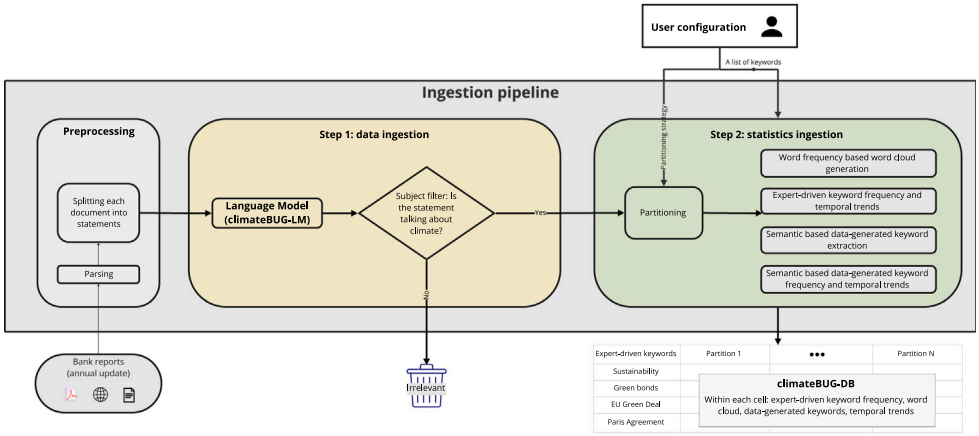

Fig. 2. This figure illustrates the ingestion pipeline proposed by this work.

In practice, this strategy can be either supervised (e.g. classification with user-defined class labels) or unsupervised (e.g. clustering). For instance, a supervised partitioning strategy could be to split statements into the two aforementioned classes (i.e. ESG vs. non-ESG), where a classifier can be trained to place unlabeled statements into one of these two classes. A partitioning strategy can also be unsupervised. For example, statements can be grouped into two mutually exclusive clusters based on their semantics using clustering techniques.

Note that the model used for the partitioning strategy is not part of the climateBUG framework. Users are responsible for choosing a partitioning strategy that is suitable for their use case. If the user does not provide a partitioning strategy, climateBUG will by default see all statements as one partition. Once it is configured, climateBUG will automatically divide the input statements according to the partitioning strategy.

Besides the partitioning strategy, a second user configurable parameter is a set of *expert-driven keywords.* An expert-driven keyword is a phrase of interest provided by the user. Expert-driven keywords are used in the following way. For each expert-driven keyword, first, cliamteBUG searches for statements that include that keyword within each partition. This step results in a subset of statements for each partition. Next, climateBUG extracts a collection of *data-generated keywords* from each of these subsets to gain an understanding of how each expert-driven keyword is being discussed. The data-generated keywords can be extracted using any state-of-the-art NLP techniques. In climateBUG, we have three alternatives: simple word frequency, Term Frequency-Inverse Document Frequency (TF-IDF) and deep learning embedding similarity (Reimers & Gurevych, 2019). The choice depends on the application: if the application requires full transparency and interpretability, word frequency or TF-IDF may provide a sufficient and satisfying outcome, whereas if the application is rather exploratory, a deep learning based approach would be more appropriate. A list of expert-driven keywords can also help to provide a check of the data-generated keywords and focus the analysis from the model on specific research questions. The relation between *expert-driven keywords*, *data-generated keywords* and *partitions* can be visualized in Fig. 3.

Keyword search. Statistics ingestion depends on (both expert-driven and data-generated) keyword search, meaning that we need means to determine if a keyword is present in a sentence. Let us use the symbol $\in_*$ to denote *inclusion*, where "keyword $k \in_*$ statement s" is read as "statement s contains keyword k". When searching for statements that contains a keyword, we apply different searching criteria depending on the type of the keyword. There are three types of keywords.

- Literal key terms: These are keywords with specific meanings (e.g. Paris Agreement, Green Deal, etc.). For these keywords, we use string search ($\in_* := \in$) to determine their presence in a statement.
- Literal key phrases: These refer to words which the basic root form is of interest. For example, for the keyword *sustainable*, if the word "sustainability" is contained within a statement, we consider the word as present. For this type of word, we first apply stemming to both the keyword and the statement, and we use string search on the stemmed version of the text (denoted as $\in_* := \in_{\text{stem}}$).
- Fuzzy key subjects: This category contains keywords and phrases that have a more free format, where permutations and transformations of words are also of our interest. For instance, the key subject "sustainable finance" and its variations such as "financial sustainability" and "financial system that supports sustainable growth" are all considered relevant for this key subject. For this category, we use Sentence-BERT, deep learning based embedding algorithms, and cosine similarity (Reimers & Gurevych, 2019) to identify the relation between the key phrase and the statement (denoted as $\in_* := \in_{\cos}$). More precisely, if the similarity between a fuzzy key subject and a sentence is larger than a certain threshold (which is a design choice for the user), then the statement is considered to contain the key subject.

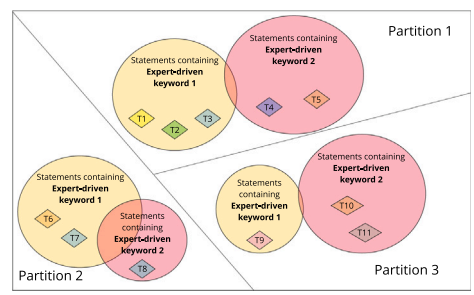

Fig. 3. Visualization of all climate statements and the relation between partitions, expert-driven keywords and data-generated keywords (denoted as T in the figure).

The resulting information is then used to populate climateBUG-DB.

B. Statistics ingestion steps. climateBUG groups statements from bank reports based on partitioning and a keyword search. Information extracted from each group constitutes one entry in climateBUG-DB.

Step 0: As a pre-processing step, statements are partitioned according to the partitioning strategy. In practice, this step can be implemented by a classification step in a supervised manner or clustering without predefined categories.

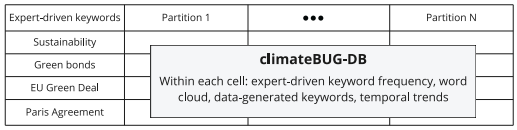

Fig. 4. A high level description of the climateBUG-DB structure.

In the subsequent steps, let $\mathcal{S}^y$ denote the set of statements from year y and a given partition. For the sake of simplicity, we neglect the partition index in the notation since the partitions are mutually exclusive and statements from each partition are analyzed independently.

Step 1: For each year y and expert-driven keyword k, search for all statements that contain k; denote the set as $\mathcal{S}_k^y = \{s \mid k \in_* s, s \in \mathcal{S}^y\}$, where $\in_*$ depends on the aforementioned type of the keyword.

Step 2: Create a count-based word cloud from $\mathcal{S}_k^y$ for each year and show it as a **word cloud**.

Step 3: Estimate word frequency of k for each year y

$$f_k^y = \frac{|\mathcal{S}_k^y|}{|\mathcal{S}^y|} \tag{1}$$

This information is stored as the **word frequency**.

Step 4: Extract top $N_T = 10$ data-generated keywords from each set $\mathcal{S}_k^y$ and denote the set of data-generated keywords as $\mathcal{T}_k^y$.

Step 5: For each data-generated keyword t,

$$t \in \bigcup_{\forall y} \mathcal{T}_k^y,$$

select all statements that contain t. Denote the set as $\mathcal{S}_{k,t}^y = \{s \mid t \in_{\cos} s, s \in \mathcal{S}_k^y\}$, where $\in_{\cos}$ indicates that the statements are selected using the fuzzy semantic search algorithm.

Step 6: Estimate the conditional word frequency (in percentage)

$$g_{k,t}^y = \frac{|\mathcal{S}_{k,t}^y|}{|\mathcal{S}_k^y|}$$

This information is stored as the **data-generated keyword frequency**

Step 7: Sort $g_{k,t}^y$ within each year y and select the top $N_{T_{\text{top}}}^y = 10$ data-generated keywords for each year.

The full ingestion pipeline then populates the database, climateBUG-DB.

2.3. climateBUG-DB

Following from the ingestion pipeline, climateBUG-DB, shown in Fig. 4, is a relational database that stores the statistical information, which can then be queried using a set of API functions provided by climateBUG.

2.4. climateBUG-API

Once climateBUG-DB is populated, users can utilize a set of interface functions, **climateBUG-API**, to extract information from climateBUG-DB for different purposes. Details of these functions can be found in Appendix A.

3. climateBUG-data

Our data set focuses on the intersection between climate and finance. With approximately 1.1M (1,070,070) data points, we believe that this data set is one of the largest data sets to reflect this scope.

3.1. Corpus

The corpus of this current data set comes from annual reports and sustainability reports published by commercial EU banks on their websites from 2015–2020. An annual report is a public document that corporations publish annually to highlight what shareholders should know about the company. Updates may depend on the relevant jurisdiction's legislation, but a report generally includes a financial disclosure and a discussion of both financial and non-financial updates and strategies from the company. Sustainability reports are also public reports that a company publishes annually which focus on various non-financial aspects of the company. These reports may also, for example, be called corporate social responsibility (CSR) reports, and they may be published along with the annual report in an integrated report. Discussions in a sustainability report generally include the bank's activities related to promoting human rights, the environment, sustainable development, and other relevant social issues.

While annual and sustainability reports respectively have aligned broadly to cover the same topics, especially when focusing on the annual reports of one industry like the financial sector, there is no consistent format for these reports where the information disclosed can be easily parsed into relevant categories. Companies, including banks, can interject their own narratives into these reports and have subtly different ways of discussing the same issue (for example, discussions of investment related to climate can be included in a number of headline topics like "responsible" investment, "sustainable" investment, "green" investment, and "ethical" investment). Therefore, while there may be a lot of key information about how a company says they are responding to an issue like climate change within these reports, it is difficult to comprehensively and systematically analyze the relevant statements using only simple keyword search or qualitatively analyze this amount of data by reading the reports.

The analysis is on a statement level, meaning that bank reports are parsed into a collection of sentences, where each sentence is referred to as a statement. The statistics of the corpus can be found in Fig. 5.

As a first step, climateBUG classifies statements from bank reports as *relevant* or *irrelevant* to climate issues and sustainability. If a statement is marked as relevant, it will be passed to the next step in the analysis climateBUG pipeline. Otherwise, it will be discarded. This classification process partially depends on the presence of certain climate-related keywords, but it also relies upon the context of the statement.

Deep learning is among the most efficient approaches when it comes to this type of complex and context-dependent classification task (Deng & Liu, 2018), especially given the recent advances in pre-trained NLP models (Devlin et al., 2018; Liu et al., 2019a; Sanh, Debut, Chaumond, & Wolf, 2019a). A deep learning based classifier is trained to filter out irrelevant statements.

3.2. Data annotation guidelines and manual process

To build an annotated data set for training the classifier, four annotators (master's students with an academic interest in climate and sustainability) were hired from January to August 2022 to help in the manual annotation process. The annotators worked for 8 h each week, which included 7 h annotation on their own and a one-hour group check-in meeting almost every week. The annotators manually reviewed a large subset of the data (sentence/statement level) to confirm whether the statement was about climate change and sustainability or not. The annotators were provided with an annotation guide which included a list of key terms that related to our focus of climate change

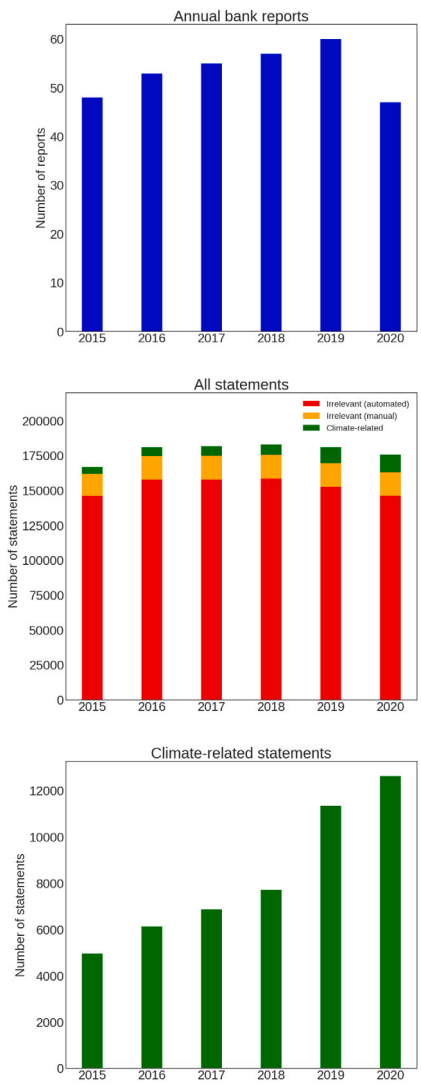

Fig. 5. Statistics of the annotated corpus.

mitigation, climate change adaptation, and sustainable finance. The annotators were provided guidance on what to include and exclude (see Appendix B).

That being said, the annotators could not blindly rely on the key terms and to a certain extent had to critically judge the context of a statement to say if it was relevant to discussions about climate change or sustainability related to climate change. For example, a phrase like "working towards a sustainable future", depending on the context of the sentence and potentially the surrounding sentences, could be about working towards a more green future or could be about the long-term profitability of the business. Therefore, the annotators had to at times use their judgment to decide if a statement was relevant to the data set.

To minimize manipulation the annotators were provided the statements randomly by bank and year. The statements before and after were presented together with the statement they were about to annotate (hence providing potential helpful context). Also, as a part of the annotation process, the annotators and project leads met once a week to review statements and further collectively hone the focus of what would be considered a part of the data set. Initially, these meetings were used to review the annotation guidelines and discuss general questions about the process. The meetings then quickly developed to reviewing tricky statements or themes of statements that the annotators sent in beforehand. When reviewing tricky statements, the annotators were asked to provide their instinctual response on whether to classify the statement as relevant or not and why. The project lead would then facilitate a discussion about the statement and ultimately decide if the statement should be considered relevant. These discussions proved extremely useful to ensure that the project leads and the annotators were consistent in annotation and aware of any problems arising from potentially difficult groups of statements as well as to build the annotators' knowledge of the subject matter. The outcomes of the discussions were noted in reviews that were sent to the group for future reference and used to further develop the annotation guidance.

3.3. Ensemble learning and automated annotation strategy

To speed up the annotation process, we developed a simple and interpretable semi-supervised active learning strategy. The objective of this automation was twofold:

1. To speed up the annotation process;
2. To minimize the chance of missing out on positive examples and potentially "tricky" negative examples.

In order to achieve these objectives, the process was implemented in three steps. We started with 1M unlabeled statements.

Step 1 Manual bootstrapping: As described above, human annotators manually annotated approximately 80K statements into relevant or irrelevant class, among which, around 5% are annotated as relevant. It is important to highlight that we assume the statements from the manually annotated dataset and the unlabeled dataset are identically and independently distributed (i.i.d.). If there is a shift in distribution, it becomes challenging to produce reliable automated annotations for the unlabeled data. In our research, the 80K manually labeled statements are uniformly drawn from all the bank reports to ensure the validity of this assumption.

Step 2 Active learning using meta pseudo-labels: Since our data
set was highly imbalanced with substantially fewer positive examples (i.e., relevant statements), our approach was to capture all unlabeled (not yet classified) statements that were potentially relevant or borderline irrelevant and subject these statements to manual annotation given the budget and time scope of the manual process. The borderline cases were useful for triggering interesting discussions between annotators and project leads so as to iterate and align on the manual annotation guideline and strategy for further improving the manual annotation quality.
In order to produce reasonable pseudo-labels, three classifiers (uncased Bert Devlin et al., 2018, FinBert Araci, 2019 and ClimateBert Bingler, Kraus, Leippold, & Webersinke, 2021; Webersinke, Kraus, Bingler, & Leippold, 2021) were fine-tuned on the 80K manually annotated statements based on a 80%–20% training-validation split. The reason for training three models and using the ensemble to produce automated annotations is to reduce potential bias caused by training a classifier using its own predictions. After the fine-tuning step, these three classifiers were applied to the 80K labeled statements to predict the labels. The final prediction was determined by the ensemble of the three classifiers based on the majority vote. Each prediction was then compared to their corresponding manual label and annotated as True Positives (TPs), True Negatives (TNs), False Positives (FPs) or False Negatives (FNs), which is referred to as the *meta label*. A meta classifier was then based on the nearest centroid algorithm was trained to predict the meta label. The feature space of this classifier was constructed by stacking the softmax outputs (i.e. the output of the deep learning model before the categorical classification output Goodfellow, Bengio, & Courville, 2016) from the three deep learning models into a three dimensional vector.

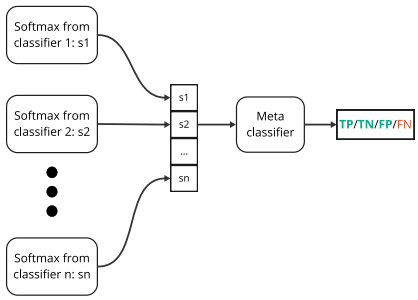

Fig. 6. Meta classifier constructed to curate interesting statements for manual annotation. In particular, the statements predicted to be potential TP, FP and FN are selected and sent for manual annotation.

Table 2
Summary of the annotated statements.

Manual/automatic	Relevant/irrelevant	Statements
Manual	Relevant	49,695
Manual	Irrelevant	100,921
Automatic	Irrelevant	919,454

An illustration of the meta classifier can be found in Fig. 6. In particular, FNs were expected to contain both relevant statements and borderline negative statements. The training data for this step was the 80K manually annotated statements. This meta classifier was applied to all the 990K unlabeled statements. The outcome of this step was a meta label for each unlabeled statement that indicated if the statement was potentially a TP, TN, FP or FN. After a manual plausibility check step, statements predicted to be TP, FP and FN were sent for manual annotation. TNs were discarded to improve efficiency.
The outcome of this step was 70K curated statements in total.
Step 3 Final manual annotation:

In the final step, the 70K statements curated by step 2 were manually annotated following the annotation steps described in Section 3.2.
A summary of the annotated statements can be found in Table 2.

3.4. Evaluation of climateBUG-Data

In this section, we evaluate our methodology for the development of climateBUG-Data.

3.4.1. Semi-automated annotation strategy

In Section 3.3, we discussed the semi-automated annotation strategy for speeding up the manual process.
To characterize the effectiveness of this strategy, we define the following terminology:

- **Hit**: if a statement is relevant and it is selected for manual annotation;
- **Miss**: if a statement is relevant but it is not selected for manual annotation;
- **Cost**: if a statement is irrelevant but it is selected for manual annotation.
- **Recall**: $recall = \frac{Hit}{Hit+Miss}$
- **Precision**: $precision = \frac{Hit}{Hit+Cost}$

The objective is to maximize the recall within the available annotation budget and time scope. In addition, we are also interested in borderline irrelevant statements to be sent for manual analysis and annotation.
To evaluate this strategy, we used 56,253 statements for training the meta-classifier and evaluated the prediction results on the 14,064 validation statements. The confusion matrix illustrating the classification performance can be found in Eq. (2). The result is further summarized in Table 3. A sensible alternative would be to subject only the statements that were predicted positive (i.e. TPs and FNs) to the ensemble model (i.e. majority vote by ClimateBert, FinBert and Uncased Bert) for annotation. Our strategy is compared with this alternative. As a result, we increase the recall by 3% (20 statements) with a cost of 72 out of 14,064 statements in total. This is considered beneficial for our use case since (1) relevant statements are rare, so we do not miss any potentially relevant ones, and (2) the irrelevant statements in the cost category are considered tricky ones and hence interesting to go through manually.

Table 3
Statistics of sending only relevant statements for manual annotation. Note that for our use case, we want to capture as many potentially climate-related data points as possible, which is reflected by the recall.

Method	Total sent	Hit	Miss	Cost	Recall	Precision
Ensemble classifier	726	604	160	122	0.79	0.83
Meta classifier	818	624	140	194	**0.82**	0.76

Table 4
Selected statements sent for the final manual annotation process. We evaluated hit and cost after the statements were annotated.

	Total	Hit	Cost
FN	12,806	4454	8352
FP	9040	5458	3582
TP	47,027	39,660	7367
Total	68,873	49,572	19,301

$$\begin{bmatrix} & \text{FN} & \text{FP} & \text{TN} & \text{TP} \\ \text{FN}(sent) & 20(\text{hit}) & 0(\text{cost}) & 72(\text{cost}) & 0(\text{hit}) \\ \text{FP}(sent) & 0(\text{hit}) & 61(\text{cost}) & 0(\text{cost}) & 122(\text{hit}) \\ \text{TN} & 140(\text{miss}) & 0(-) & 13106(-) & 0(\text{miss}) \\ \text{TP}(sent) & 0(\text{hit}) & 61(\text{cost}) & 0(\text{cost}) & 482(\text{hit}) \end{bmatrix} \tag{2}$$

The meta classifier was applied to all unlabeled statements, where predicted TP, FP and FN were sent for manual annotation. After we got back the annotation results, we evaluated the meta classifier again based on the number of statements that are hit and cost respectively (see Table 4).

4. climateBUG-LM

climateBUG-LM is a deep learning based language model fine-tuned on climateBUG-Data. Its primary task is to classify each statement into a relevant or irrelevant class depending on if the statement is talking about climate-related subjects or not. For recent advances of general purposed NLP systems, see Sakshi and Kukreja (2023) and Shao, Zhao, Yuan, Ding, and Wang (2022).
A common practice is to adapt an off-the-shelf language model to the domains of interest by fine-tuning a pre-trained backbone on the domain-specific corpus (Howard & Ruder, 2018). There are several reasons for this, among them that for a given task, training a language model from scratch requires significant resources (Strubell, Ganesh, & McCallum, 2019). Given these considerations, the development of climateBUG-LM consisted of three steps: expansion of the vocabulary, masked-language model fine-tuning, and downstream classifier fine-tuning.

4.1. Vocabulary expansion

To apply deep learning models to specific domains, it is important to expand the vocabulary due to the domain-specific terminology (Gururangan et al., 2020). To this end, we chose the vocabulary provided by the RoBERTa model (Liu et al., 2019b) as the base vocabulary, which was then modified to include relevant words and phrases from

Table 5
Training hyperparameters for climateBUG-LM backbone fine-tuning.

Hyperparameter	Value
Interface	Huggingface
Learning rate	2e−6
Epochs	30
Warmup steps	1000
Batch size	24
Early stopping	3 steps
MLM probability	0.2
Weight decay	0.01
FP16	True

Table 6
Training hyperparameters for climateBUG-LM downstream classifier fine-tuning.

Hyperparameter	Value
Interface	Huggingface
Learning rate	1e−6
Epochs	100
Warmup steps	1000
Batch size	32
Early stopping	6 steps
Weight decay	0.01
FP16	True

the domains of climate and finance. In particular, our curated list includes phrases related to business acronyms, EU legislation, bank's name, bank's products and services, renewable energy, key commitments, types of climate-related risks, sustainability, policy proposal, and initiatives. A full list of added vocabulary and the reasoning behind the vocabulary can be found in Appendix C.

4.2. Masked-language model

The second step for domain adaptation was to fine-tune the backbone language model given the expanded vocabulary and corpus. We chose the popular and efficient backbone language model DistilRoBERTa (Sanh, Debut, Chaumond, & Wolf, 2019b) and fine-tuned it on the masked-language modeling task (MLM). The corpus for fine-tuning the MLM consisted of approximately 150K manually annotated statements as shown in Table 2. The annotations were not needed in this language model fine-tuning step. However, the annotations were provided in order to construct a class-balanced (i.e. relevant vs irrelevant) corpus for the fine-tuning step. Among these 150K statements, 90% were randomly selected for fine-tuning and 10% for validation. A list of configurations and hyperparameters can be found in Table 5.

4.3. Downstream classification model

The primary task of climateBUG-LM is to analyze the semantic of a statement in EU bank reports such as classifying a statement into a relevant or irrelevant class. In order to achieve this, the language model needs to be fine-tuned for the downstream classification task. We adopted the same data setup as in Section 4.2, where the exact same training-validation split is used. Some key hyperparameters for this fine-tuning step are described in Table 6.

4.4. Evaluation of climateBUG-LM

The primary task of the language model climateBUG-LM is to automatically identify if a statement is relevant to climate related subjects or not. To this end, climateBUG-LM is first fine-tuned on the corpus described in Section 3.1 and then fine-tuned on the downstream classification task based on the annotations in climateBUG-Data.

To evaluate the performance of climateBUG-LM, we used bank statements from year 2015 to 2019 for training and 2020 for validation.

Standard classification metrics are calculated to compare the performance. Each has its own unique strengths.

- **F1 (weighted):** This metric is the harmonic mean of precision and recall.
- **F1 (macro):** The macro version of the F1 metric computes the F1 score independently for each class and then takes the average. This treats all classes equally, regardless of their prevalence in the dataset. This metric was added to address the imbalanced nature of the dataset.
- **Accuracy:** This metric is a straightforward measure of the proportion of correct predictions made by our model. It gives a high-level view of the model's performance but is complemented by the F1 scores, which offer a more nuanced understanding in the context of class imbalance.

The result in terms of accuracy and F1 score are reported in Table 7. In particular, the classification performance was compared to a baseline model Bert (uncased) that is based on a generic vocabulary and corpus, and two other domain-specific language models, ClimateBert and FinBert. Bert and its variants are widely used in various application domains, and they are proven to be among the most effective language models in terms of training efficiency and performance (Khurana, Koli, Khatter, & Singh, 2023; Koroteev, 2021; Wolf et al., 2020; Zhou et al., 2023). Therefore, we primarily focus on comparing our model to this family of models. All models were fine-tuned on the downstream classification task, while only the backbone of climateBUG-LM was fine-tuned using Masked Language Modeling (Devlin et al., 2018) on the domain-specific corpus.

Results suggest that climateBert outperforms FinBert and uncased Bert due to its stronger domain relevance. However, the best performance was achieved by climateBUG-LM thanks to domain adaptation.

As for the choice of our training and testing periods, it was determined to focus on temporal validation, a strategy commonly adopted in predictive modeling, especially when predicting future events is the goal. By training on report statements from 2015–2019 and testing on 2020 data, we ensure that our model is able to predict "future" events based on "past" data. This reflects a realistic use case scenario where the model would be deployed to analyze upcoming annual reports based on learnings from past reports. The high performance of the model on this unseen data, which represents a different time period, validates our hypothesis that our model is reliable in that it is capable of handling potential minor domain shifts given an updated time period.

When considering other potential domains, such as applying the climateBUG framework to bank reports outside Europe, the inherent adaptability of the semi-automated annotation strategy used in our model comes into play, where users can use climateBUG-Data to bootstrap annotations on the new data. This approach allows the model to adjust to new data distributions efficiently, ensuring its ability to accommodate new domain shifts with replicable results and with lesser need for new annotated data. It can also be noted that the language used in annual reporting related to climate change domain has a certain universal character due to the global nature of climate change discussions within the banking industry and common financial terminologies (Elliott & Löfgren, 2022). However, if a larger domain shift is observed, for example by introducing drastically different time periods or industries or seeing a major change in the discussion on climate change, the model should be updated using the methodology (including new annotations) as outlined in Section 5.3. Whether or not a domain shift is large enough to warrant an updating of the model will necessarily be a subjective decision based on the evaluation of the model in relation to the specific use case. In such cases, the annotation guidelines and annotated data as part of the climateBUG framework are helpful as a basis for researchers but would also potentially need to be updated to fit the specific research interests.

Table 7
Evaluation of the classification performance. The masked language model and the downstream classifier are both fine-tuned on the corpus of EU bank reports from 2015 to 2019. The classifier is then evaluated on unseen statements from the year 2020.

Backbone	Domain	Number of parameters	Vocabulary size	F1 (weighted)	F1 (macro)	Accuracy
Bert (uncased)	Generic	109,484,547	30 522	90.92%	90.81%	90.88%
FinBert	Finance	109,484,547	30 522	90.93%	90.82%	90.88%
ClimateBert	Climate	82,301,187	50 500	91.18%	91.07%	91.13%
climateBUG-LM	Climate + finance	82,300,418	50 421	**91.47**%	**91.36**%	**91.42**%

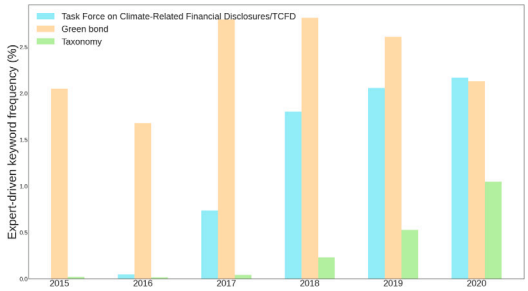

Fig. 7. Annual expert-driven keyword frequency extracted from a single partition.

5. Application of the climateBUG

First, we note that the current climateBUG-Data has already been through an iteration of the model to classify the data based on statements' relevance to climate or sustainability, as discussed in Section 3. Therefore, the data we analyze has already been classified as climate-related and non-climate statements (which were discarded from the set). This process itself is an example of the application of climateBUG in creating a data set that is focused on the nexus of finance and climate change. In this section we present two applications of how the climateBUG-data can be processed and analyzed, as well as providing a step-by-step instruction of how to use the climateBUG framework for more tailored analysis.

5.1. Expert-driven keywords

To show the various ways climateBUG can be used to analyze the current dataset, we highlight three example keywords to show how keywords can be used to provide analytical insight in our analysis: green bond, Taskforce for Climate-Related Financial Disclosures (TCFD), and taxonomy.

Defining each keyword in turn, *green bond* is a keyword because it is an example of how banks are moving investments into more environmentally friendly projects. This can provide insights into how this investment mechanism is discussed in relation to other green mechanisms or respond to the question of if recent legislation has had an effect on the discussion of green bonds.

Taskforce for Climate-Related Financial Disclosures (and its acronym, TCFD) is a keyword because it is an example of an international initiative to promote a standardized climate-related risk disclosure for businesses (Taskforce on Climate-related Financial Disclosures, 2017). This can provide insights into how this initiative has been incorporated into other reporting or disclosure strategies or respond to the question of to what extent this initiative has been adopted since its introduction in 2017.

Taxonomy is a keyword as it relates to a key aspect of the EU Sustainable Finance legislation for defining and categorizing financial products as sustainable (European Parliament, 2020). This can provide insights into how this legislation is being discussed in its upcoming implementation and respond to the question of how the taxonomy has affected the discussion of a bank's financial products.

Looking at Fig. 7 the word frequency of the three keywords "green bond", "taxonomy", and "TCFD", highlight some key observations from the data. While we do not provide a thorough analysis of these observations to show causality in our analysis, we want to point to some interesting potential correlations and trends which exemplify some research questions that can be further pursued.

Starting with "TCFD", we would expect that the keyword would sharply increase with the introduction of the TCFD recommendations in 2017. The graph indeed shows an increase in both 2017 and 2018 as a lot of the banks included in this data set (20 out of the 35 banks included) sign up to supporting the TCFD recommendations by the end of 2018. We also see that the word frequency remains fairly consistent from 2018 to 2020. We suggest this is because the TCFD is an annual disclosure process, so banks would consistently discuss their TCFD disclosure yearly.

Moving to the keyword "taxonomy", we would also expect a sudden increase in frequency in 2018 when the EU taxonomy concept was introduced in the 2018 EU Action Plan on Financing Sustainable Growth (European Commission, 2018). The keyword continues to rise in frequency through 2020 as the taxonomy potentially is becoming more prevalent in discussion as it gets closer to its expected application date starting in 2022. We note that "taxonomy" is used substantially less than the other keywords, and we suggest this is because "taxonomy" is an example of a keyword of developing, upcoming legislation without clear disclosure standards. We expect "taxonomy" to continue to increase in the near future as banks start to apply the taxonomy to relevant financial products subsequent to the mandatory application timeline.

Finally, we look at the keyword "green bond". We would expect that the frequency of "green bond" would increase from 2018 to 2020 as the EU Action Plan also highlighted a focus on promoting and standardizing green bonds through upcoming legislation. Instead, we find that "green bond" rises in 2017 and then remains fairly consistent with minor fluctuations up and down through 2020. We note that, in comparison to the other keywords "green bond" has been a term used before the data set begins and is related to a financial product rather than a developing international or legislative initiative. We suggest that this trend in "green bond" frequency may reflect more the market of green bonds, which has notable increases in amounts issued between 2016–2019 (Climate Bonds Initiative, 2022). Further research could follow whether the market or upcoming legislature or initiatives affect the use of "green bond".

5.2. Clustering as a partitioning strategy

Clustering is a form of unsupervised machine learning where the goal is to group similar data points together for more targeted analysis (Benchimol et al., 2022; Reimers & Gurevych, 2019; Xu & Tian, 2015). Clustering is widely used in many fields, including image recognition, customer segmentation, and recommendation systems. In text analysis, clustering can be used to automatically group texts with similar topics together and enable the user to extract meaningful information from large, unstructured datasets. More specifically, in the context of climateBUG-Data, bank statements in the same cluster are expected to be more semantically similar to each other than to statements in other clusters. This unsupervised technique can be used as a partitioning strategy for exploring the climateBUG-Data.

This type of analysis provides a more powerful way that NLP can be used to draw observations beyond simple word frequency or keyword analysis.

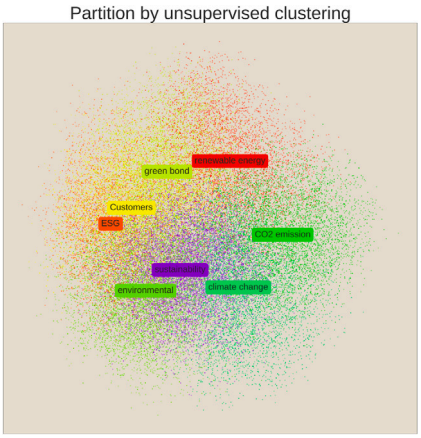

Fig. 8. Unsupervised multi-cluster partitioning and data-generated keyword frequency.

In Fig. 8, we offer an example of unsupervised multi-cluster partitioning, where the model partitioned climateBUG-Data into clusters based on data-generated keywords. We can see that the clusters created by the model include the keywords ESG, environmental, sustainability, climate change, CO_2 emissions, bank, green bond, and renewable energy. These clusters point to key themes within banks' narratives related to their actions regarding climate change and sustainability. For example, green bonds and renewable energy are initiatives that the banks can actively promote through their products. CO_2 emissions are a key indicator used to measure a company's impact on climate, and therefore it is interesting that it is highlighted as a cluster in the analysis. Finally, the model has picked up semantic differences in the clusters of ESG, environmental, sustainability, and climate change; while these terms may be used interchangeably in some ways (as shown by the ways the data points overlap substantially in these categories), it is notable that the model still was able to identify clear clusters for each keyword. Further analysis within each cluster would be needed to make more substantial claims about banks' discussion of their activities against climate change, but we argue that the partitioning itself provides a much-needed first step in analyzing this amount of data in a feasible, more systematic, and comprehensive way.

5.3. Step-by-step instruction for using climateBUG for more tailored analysis

The examples above are potential ways in which the model can be used to look at the intersecting discussion between finance and climate change. Beyond these examples, there are multiple other applications where the model can be used, such as understanding how policy can potentially change how finance and climate change are discussed, detecting references to key policy instruments, and identifying novel topics (such as in relation to offsetting) over time.

In practice, the intended use case is to combine both user configurations, the partitioning strategy, and keywords to generate comprehensive statistics by analyzing statements from EU bank reports. Below follows a succinct step-by-step instruction on how to use the framework for more tailored analysis on how to understand how EU banks talk about climate change.

Step 1 Choosing a partitioning strategy: This strategy can be any algorithm that splits the statements into mutually exclusive subsets (cf. Section 2) (e.g. a simple partitioning strategy could be K-means clustering with 8 clusters as shown in Fig. 8 presented in Section 5.2.

Step 2 Choosing the keywords of interest: The user can choose any keywords of interest, e.g. *Green Bond*, *Taskforce for Climate Related Financial Disclosures*, and *Taxonomy* presented in Section 5.1).

Step 3 Executing the statistics ingestion pipeline to populate the climate-DB database: Relevant statistics will be cached in the database for the user to query.

Step 4 Visualization: After the statistics ingestion step, the user is able to visualize the statistics queried from the database.

6. Concluding remarks

Based on a corpus of European banks' annual reports, climateBUG offers a framework to detect latent information about how banks discuss their activities related to climate change. The framework is built on an ingestion pipeline to extract information from the corpus data, a configurable database, and a set of API's. The framework has been developed using an interdisciplinary approach consisting of domain knowledge from financial reporting and climate economics analyzed through the lens of advanced computational linguistics natural language processing.

In addition, climateBUG provides two standalone components that can be used for customized analyses; climateBUG-Data and climateBUG-LM. The climateBUG-data is a unique annotated corpus with the scope of climate change and finance with approximately 1.1M data points available open access that can be used as training data for further model optimization. The climateBUG-LM is a deep-learning model adapted to the corpus. When benchmarking on classification performance, climateBUG-LM outperforms other models currently available (Bert uncased, FinBert, and ClimateBert). An important feature of the framework climateBUG is that it is *human-centric* in the sense that, despite having automation as its main functionality, the construction is heavily influenced by the knowledge of domain experts. Based on the inherent adaptability of the semi-supervised learning approach used in our model, this approach allows the model to adjust to new data distributions, ensuring its ability to accommodate new domain shifts with little need for new annotated data. We anticipate that adding new yearly reports from banks or adding new bank reports from different jurisdictions would be similar to the currently trained domain. For a significant domain shift, like adding annual and sustainability reports from different industries or a drastic change in sustainability discussion, a phase of manual annotation can be used to further fine-tune the model.

We also provide examples of how the framework can be applied by users by looking at trends in how banks talk about green bonds, TCFD, and the EU Taxonomy, as it relates to climate. We additionally include a simple step-by-step introduction on how to use the framework, tailored for readers with less experience in utilizing NLP and deep learning models. The primary objective of its outcome is to be interpretable by a broad range of end users including academia, government representatives, journalists, and commercial banks' sustainability managers.

CRediT authorship contribution statement

Yinan Yu: Designed the analysis, Contributed data and analysis tools, Performed the analysis, Wrote the paper. **Samuel Scheidegger:** Designed the analysis, Contributed data and analysis tools, Performed the analysis. **Jasmine Elliott:** Conceived and designed the analysis, Contributed and collected data, Wrote the paper. **Åsa Löfgren:** Conceived and designed the analysis, Contributed and collected data, Wrote the paper, Project leader, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Datasets, models, code, and ongoing work are available through the authors' project page at https://www.climatebug.se/.

Acknowledgments

During the process of this project a number of assistants have contributed to the work. In particular we would like to thank Linus von Ekensteen for initial AI support, including model and web interface development; Johan Hammarstedt for AI support; Erik Brinde, Gabriel Nordblom, Ulrika Winter, and Lauren Yehle for annotation assistance. We would also like to thank two anonymous referees for valuable comments. The authors gratefully acknowledge financial support from the Mistra (the Swedish Foundation for Strategic Environmental Research) Carbon Exit Research Program, the UGOT Centre for Collective Action Research and the Vinnova (Sweden's innovation agency) for the Sustainable Finance Lab. The funding sources had no involvement in the research process.

Appendix A. Climatebug-api functions

Function name	Arguments	Return value (data type)	Description
get_statements	Partition, expert-driven keywords, years, data-generated keywords	A set of statements (a list of strings)	Statements given the search criteria on the arguments
get_wordcloud	Partition, expert-driven keywords, years, data-generated keywords	Words and their frequencies (a hashmap with words as its keys and frequency as its values)	Words with top frequencies from statements given by the search criteria on the arguments
get_keyword_freq	partition, expert-driven keyword, year	Frequency (a real value with range $[0, 1]$)	Keyword frequency for a given expert-driven keyword within a partition
get_dgkeys	partition, expert-driven keyword, year	A set of data-generated keywords (a list of strings)	Within a partition, this function returns data-generated keywords extracted from sentences that contain a given expert-driven keyword.
get_dgkey_freq	partition, expert-driven keyword, data-generated keyword, year	Frequency (a real value with range $[0, 1]$)	This function computes the frequency of an extracted data-generated keyword from statements given by the search criteria on the arguments

Appendix B. Annotation guidelines

Generally accept	Generally reject
Climate change Green	Gibberish Languages not in English
Non-financial risks (if related to environment/climate change)	Sustainability not related to the environment (i.e. sustainable profits)
Green bond	Circular economy
Carbon	Recycling not related to banks own operations
Global warming	Nature conservation efforts
Paris Agreement	Prompts from disclosure standards (i.e. GRI or TCFD)
Scope 1, 2, or 3 emissions	Section headings or titles
Sustainable finance	Statements that are part of a table of contents
ESG (environmental, social, governance)	Table figures
Fossil fuels (if related to environment)	
Energy industry (if related to environment)	
Sustainable development goals (if related to environment)	
Energy efficiency (in operations of bank or related to lending)	
Operational recycling efforts by bank	
Reference to disclosures (if related to environment/climate change)	
Non-financial risks (if related to sustainability related to environment)	
Equator Principles	
Task Force on Climate Disclosures (TCFD)	
RE100	
UN Global Compact	
UN Environmental Programme (UNEP)	
UNEP Finance Initiative	
Principles for Responsible Banking (PRB)	
Principles for Responsible Investing (PRI)	
Carbon Disclosure Project (CDP)	
2 Degrees Investing Initiative EU Green New Deal EU Taxonomy EU 2019/2088 EU 2019/2089 EU 202/852	

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.122162.

References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Benchimol, J., Caspi, I., & Kazinnik, S. (2023). Measuring communication quality of interest rate announcements. *The Economists' Voice*.

Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. *Machine Learning with Applications*, *8*, Article 100286.

Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2021). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Corporate Finance: Governance*.

Bouckaert, S., Pales, A. F., McGlade, C., Remme, U., Wanner, B., Varro, L., et al. (2021). Net zero by 2050: A roadmap for the global energy sector.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, Y., Wu, J., & Wu, Z. (2022). China's commercial bank stock price prediction using a novel K-means-LSTM hybrid approach. *Expert Systems with Applications*, *202*, Article 117370.

Climate Bonds Initiative (2022). Explaining green bonds. Available at https://www.climatebonds.net/market/explaining-green-bonds.

Conroy, J. M., & O'Leary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 406–407).

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2021). Sentiment in central banks' financial stability reports. *Review of Finance*, *25*(1), 85–120.

Deng, L., & Liu, Y. (2018). *Deep learning in natural language processing*. Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

El-Haj, M., Rayson, P., Walker, M., Young, S., & Simaki, V. (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, *46*(3–4), 265–306.

Elliott, J., & Löfgren, Å. (2022). If money talks, what is the banking industry saying about climate change? *Climate Policy*, 1–11.

European Commission (2018). Action plan: Financing sustainable growth. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0097.

European Parliament (2020). Regulation EU 2020/852 of the European parliament and of the council of 18 june 2020 on the establishment of a framework to facilitate sustainable investment, and amending regulation (EU) 2019/2088. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32020R0852.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, *6*(1), 1–25.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., et al. (2020). Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, *193*, Article 116429.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Jelinek, F., Lafferty, J. D., & Mercer, R. L. (1992). *Basic methods of probabilistic context free grammars*. Springer.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744.

Koroteev, M. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

Lamperti, F., Bosetti, V., Roventini, A., & Tavoni, M. (2019). The public costs of climate-induced financial instability. *Nature Climate Change*, *9*(11), 829–833.

Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, *49*(5), 587–615.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019a). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019b). Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692. arXiv:1907.11692. URL http://arxiv.org/abs/1907.11692.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Rish, I., et al. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3* (pp. 41–46).

Sakshi, & Kukreja, V. (2023). Recent trends in mathematical expressions recognition: An LDA-based analysis. *Expert Systems with Applications*, *213*, Article 119028.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019a). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019b). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv arXiv:1910.01108.

Shao, Z., Zhao, R., Yuan, S., Ding, M., & Wang, Y. (2022). Tracing the evolution of AI in the past decade and forecasting the emerging trends. *Expert Systems with Applications*, *209*, Article 118221.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3645–3650). Florence, Italy: Association for Computational Linguistics.

Taskforce on Climate-related Financial Disclosures (2017). Final report - recommendations of the task force on climate-related financial disclosures. Available at https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). Climatebert: A pretrained language model for climate-related text. arXiv preprint arXiv:2110.12010.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165–193.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*, 43–52.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., et al. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419.