Knowledge informed sustainability detection from short financial texts

Boshko Koloski[1,2] , **Syrielle Montariol**[1] , **Matthew Purver**[1,3] , **Senja Pollak**[1]
[1]Jožef Stefan Institute
[2]International Postgraduate School Jožef Stefan
[3]Queen Mary University of London
{boshko.koloski, senja.pollak}@ijs.com, m.purver@qmul.ac.uk

Abstract

Nowadays in the finance world, there is a global trend for responsible investing, linked with a growing need for developing automated methods for analysing Environmental, Social and Governance (ESG) related elements in financial texts. In this work we propose a solution to the FinSim4-ESG task, consisting in classifying sentences from financial reports as sustainable or unsustainable. We propose a novel knowledge-based latent heterogeneous representation that relies on knowledge from taxonomies, knowledge graphs and multiple contemporary document representations. We hypothesize that an approach based on a combination of knowledge and document representations can introduce significant improvement over conventional document representation approaches. We perform ensembling, both at the classifier level and at the representation level (late-fusion and early-fusion). The proposed approaches achieve competitive accuracy of 89% and are 5.85% behind the best score in the shared task.

1 Introduction

In this work we develop a *knowledge-backed* approach for the detection of **sustainability** on premises of a given short textual document (i.e. a sentence). More specifically, we propose a solution to the shared task of the FinSim4-ESG workshop, where the task is to classify a given sentence extracted from a company financial report as either sustainable or unsustainable.

Investors have ever-increased interest in the assessment of the **Environmental, Social and Governance** (ESG) criteria, as non-financial factors describing the company's position on social well-being [Nagy *et al.*, 2016]. ESG criteria cover a company's environmental impact (Environmental), their relationships with their community including employees, suppliers and customers (Social), and their leadership structures including executive pay, shareholder rights, audits and controls (Governance). These *ESG* factors are usually reported as a structured output in the companies annual reports. The companies reporting these factors have moved from only a dozen in the 1990s to more than 6000 in 2014 [Serafeim and Yoon, 2022]. In a study by [Amel-Zadeh and Serafeim, 2018] using survey data from mainstream investment organizations, the authors provide insights into why and how investors use reported ESG information and highlight that relevance to investment performance is the most frequent motivation, followed by client demand, product strategy, and then ethical considerations. Also social media impacted the way of interaction between companies, employees and potential customers. Publishing posts that reveal a certain way of operation that leads to company's miss-behaviour can have a wild response from the customers. Aula [2010] studied the impact of social media on the reputation risk and the ambient publicity. An instance of such event is the H&M's *thrash-gate* scandal - where a company producing and selling clothes was charged of damaging and disposing them as waste instead of reusing them. The story sparked a public outrage and a sever impact on the company. Recently [Guo *et al.*, 2020] highlighted the high correlation of the company's volatility on the market based on the ESG factors. These works showcase how social monitoring of posts can be a powerful asset in the way of achieving a more sustainable and environmentally friendly companies and market.

The field of natural language processing (NLP) has seen increased interest in ESG-related automated analysis. For example, [Mehra *et al.*, 2022] propose fine-tuning a generic BERT model [Devlin *et al.*, 2018] on ESG corpus (ESG-BERT), and use this model for detecting positive or negative change in companies stock values based on the related sections of their 10-Q filings. [Serafeim and Yoon, 2022] showed that ESG ratings can be used to predict market reactions to ESG news, particularly when there is disagreement amongst raters.

The remainder of this work is structured as follows: Section 2 describes the related work, Section 3 presents the dataset. Section 4 presents the proposed method, followed by the results in Section 5 and the final remarks and conclusions in Section 6.

2 Related work

In the FinSim4-ESG shared task, the goal is to classify a given sentence. as either sustainable or unsustainable. A sentence is defined as sustainable if it mentions any ESG factor from a dedicated ESG taxonomy, unsustainable otherwise. We treat this problem as **binary document classification**.

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 228 - 234

To be able to learn to classify the documents, initial approaches focused on lexicons or used machine learning techniques. For the financial domain, the collection of financial dictionaries by [Loughran and McDonald, 2011] has been widely used. Later, machine learning approaches have been proposed where models need a numerical representation of documents as input. Initial methods for document classification relied on hand-crafted features or on word frequency counts using various weighting schemes (e.g. TF-iDF [Sammut and Webb, 2010]). For example, [Qiu *et al.*, 2006] represent past annual reports with TF-IDF weighted word stems and various feature selection methods in order to predict Return On Equity (ROE) ratio classes with a linear SVM classifier. Weighted (TF, TF-IDF and logarithm damp weighting) *unigrams and bigrams* are used as features in a study by [Kogan *et al.*, 2009], where a support vector machine for regression (SVR) with linear kernel is trained to predict volatility of stock returns. [Balakrishnan *et al.*, 2010] use a linear SVM classifier to predict subsequent performance based on narrative parts of 10-K reports, based on both word-level and document-level features.

The democratisation of neural-networks introduced denser and more robust document representations, where the models from this paradigm are tasked to predict the next word or the missing word in the sequence. Contemporary state-of-the-art models such as BERT [Devlin *et al.*, 2018] are based on the transformer architecture [Vaswani *et al.*, 2017]. This model learns to generate document representations by being pre-trained on a big corpora from a general domain on the task of Masked Language Modeling, where a portion of the corpora is masked and the model is tasked to predict the words missing. The pre-trained model is then fine-tuned on data from a downstream task such as document classification; this is the transfer learning setting. In this study, we utilize two different variants of the BERT model family: FinBERT [Yang *et al.*, 2020], a model pre-trained on financial data, and LinkBERT [Yasunaga *et al.*, 2022], a model that modifies the initial BERT learning paradigm by taking into account background knowledge.

Taxonomies and ontologies are increasingly used for machine reasoning over the last few years. In our study we use Tax2Vec [Škrlj *et al.*, 2021] which is based on knowledge derived from *taxonomies*, aiming at improving short documents classification. Recently [Koloski *et al.*, 2022] studied the inclusion of *knowledge graphs* as banks of large factual knowledge. In their work, they have proposed heterogeneous representation ensembles that are based on knowledge graphs and contextual and non-contextual document representations. These proposed representations achieve nearly state-of-the art results on various tasks such as classification of short texts in the scope of the depression detection from short documents (social media posts) [Tavchioski *et al.*, 2022]. In the financial domain, automatic classification of a given list of financial terms against a domain ontology was proposed in the scope of FinSim2 [Mansar *et al.*, 2021].

In terms of ESG-related NLP, in addition to ESG-BERT by [Mehra *et al.*, 2022], [Armbrust *et al.*, 2020] studied the effect of the environmental performance of a company on the relationship between the company's disclosures and financial performance, [Sokolov *et al.*, 2021] focused on automated ESG scoring, while [Purver *et al.*, 2022 accepted] performed a diachronic analysis of ESG terms in UK annual reports.

3 Data

The shared task consisted of two phases: development of methods and official evaluation. In the first phase the organizers released *2265* training documents. For our internal evaluation purposes we created custom splits of the data into *1812 (80%)* documents for training, *226 (10%)* for development and *227 (10%)* for testing. We give description of the data in Table 1.

	Training data	Development data	Test data
sustainable	978 (54 %)	122 (54%)	123 (59 %)
unsustainable	834 (46 %)	104 (46 %)	104(41 %)
All	1812	226	227

Table 1: Data distribution in our training set.

In the second phase the organizers released a test set consisting of *205* documents.

4 Methodology

In this section, we present the different methods we used to generate sentence representations. We classify them into 3 categories: standalone, which are either knowledge or text-based, high-level, which are ensembles of representations and models learned on top of the standalone, and fine-tuned BERT models.

4.1 Standalone representations

We derive standalone representations via two different paradigms: textual-driven and knowledge-driven. The former rely only on either contextual or non-contextual word features while the latter is based on features obtained from some knowledge base or taxonomy.

Non-contextual textual features

Following [Koloski *et al.*, 2021], we extract *stylometric* and *latent semantic analysis* based features.

Stylometric features were built on top of word and character frequencies statistic descriptions - maximum and minimum word size, number of characters, number of words, number of vowels, etc.

Latent Semantic Analysis [Dumais *et al.*, 1988] was built on top of top-*n* word and n-grams features, TF-IDF weighted and represented in a latent space of d dimensions. We generate multiple combinations of n-gram features n and final dimension space d:

- LSA - *n=2500*, *d = 512*
- LSA_1 - *n=5000*, *d = 256*
- LSA_2 - *n=5000*, *d = 128*
- LSA_3 - *n=10000*, *d = 512*

Contextual textual features

For the contextual features we use *sentence-transformers* [Reimers and Gurevych, 2019] representations. The method is constructed on top of a *BERT* model, using BERT representations as input to a Siamese network that learns sentence representation as an intermediate task while it predicts sentence similarity.

Taxonomy-based representation

Leveraging background data in form of taxonomy has proven successful for classification of short documents. Here, we use the *Tax2Vec* model [Škrlj *et al.*, 2021][1] where the words from a given document are mapped to the terms of the WordNet taxonomy [Fellbaum, 1998]; then, a term-weighting heuristic is applied for the construction of the final taxonomy-enriched feature space. We use the default parameters *max-features = 10*, *heuristic = "pagerank"*, *disambiguation-window = 2* and *start-term-depth = 3*.

Knowledge graph based representation

Factual knowledge about concepts and relations linking those concepts together are stored in large knowledge bases. We consider a knowledge-backed document representation from the Wikidata5m [Vrandečić and Krötzsch, 2014] knowledge graph. We follow the approach proposed in [Koloski *et al.*, 2022] to extract and generate knowledge graph based document representations. To obtain the representations of the entities, we utilize three different embedding methods:

- TransE [Bordes *et al.*, 2013] - embedding method based on simple tensor factorization, capable of capturing the *antisymmetry*, *inversion*, *transitivity* and *composition* property of relations.
- DistMult [Yang *et al.*, 2014] - embedding method based on neural tensor factorization, capable of capturing the *symmetry* property of relations.
- RotatE [Sun *et al.*, 2019] - embedding method based on complex-space tensor factorization, capable of capturing the *symmetry*, *antisymmetry*, *inversion*, *transitivity*, and *composition* property of relations.

Classifier learning for the standalone representations

For the above representations, we consider learning Stochastic Gradient Descent classifier with a search on the hyper-parameters space proposed in the autoBOT, an auto-ML model [Škrlj *et al.*, 2021]: [2]

- loss : *hinge*, *log* or *modified-huber*
- class-weight : *balanced*
- penalty: *elasticnet*
- power-t $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$
- alpha $\in \{0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005\}$
- l1-ratio $\in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1\}$
- Early-stopping criteria $\in \{8, 32\}$

[1]https://github.com/SkBlaz/tax2vec

[2]https://github.com/SkBlaz/autobot/blob/master/autoBOTLib/learning/hyperparameter_configurations.py

4.2 Fine-tuned BERT variants

We use several *state-of-the-art* BERT variants[3] and fine-tune them for our task.

- **FinSim** [Yang *et al.*, 2020] A contextually pre-trained BERT model on a large scale financial corpora with more than 4.9 billion tokens from corporate reports, conference call transcripts and financial analysts reports.
- **LinkBERT** [Yasunaga *et al.*, 2022] A knowledge-informed BERT model pre-trained on two joint self-supervised objectives: *MLM* (masked language modeling) and *DPR* (document relation prediction). In the former, a part of the input sentence is masked and the model is tasked to predict this masked token. In the latter, given two paragraphs, the model is tasked to predict whether they come from documents that are linked, whether they are subsequent in the same document of whether they are not related at all. During training, the model considers the graph of links between Wikipedia documents.

We train the models with a reproducible seed of *42* and a learning rate of $5e^{-5}$ for *10* epochs with *32* documents in a single batch.

4.3 Higher-level representations

Early-fusion

In order to explore the expressiveness of the joint representations, we construct two different approaches for fusion of representations:

Naive concatenation - We concatenate all the generated representations previously described in the *standalone representations* subsection.

Construction of latent spaces - We first concatenate all the generated representations, then we perform singular-value-decomposition **(SVD)** to obtain a new *joint latent space*. We reduce the proposed space to $d \in \{256, 512, 1024\}$ dimensions.

Late-fusion

Finally, we build ensembles on top of the standalone models (i.e. **late-fusion**). For the final ensemble we use the fine-tuned *FinSim*, *LinkBERT* and the *jointSVD* predictions. The final prediction is based on the majority vote (i.e. the class selected by at least two out of three methods).

5 Results

In this section we report the results of our internal evaluation (with our own splits) together with the final evaluation using the test set of the shared task.

5.1 Internal evaluation

We perform a thorough internal evaluation on our custom data split described in Section 3. We train all our models on the *train* split and optimize the hyper-parameters using the *development* split. For all models we report the evaluation with respect to the F1-score. We use the *test* split for the selection

[3]Implementation and checkpoints from the *huggingface* library.

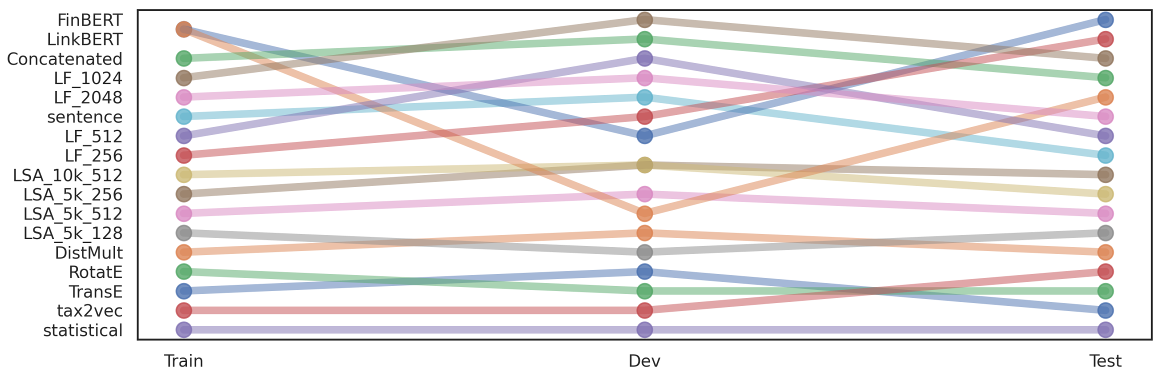

Figure 1: Ranking of the various document representations per split in the dataset (in the internal evaluation phase).

of the final models for submission. Among the knowledge-based methods, the **DistMult** method performs best, achieving a score of 70.91% on the test set, outscoring the *RotatE* by 3.09% and the *TransE* method by 6.83%. The *DistMult* method also outscores the *tax2vec* method by 0.62%.

Method	Dims	Train	Dev	Test
Knowledge based				
TransE	512	71.96	76.10	64.08
DistMult	512	*81.16*	*85.6*	*70.91*
RotatE	512	74.12	72.72	67.82
tax2vec	10	70.18	70.11	70.29
Text based				
statistical	10	59.23	60.38	55.05
LSA	512	89.20	92.43	*88.61*
LSA_1	256	85.53	90.16	85.95
LSA_2	128	83.42	85.59	85.36
LSA_3	512	89.78	*92.41*	88.10
sentence transformers	768	*95.32*	*93.98*	*91.20*
Fine-tuned BERTs				
LinkBERT	512	***100.0***	*92.85*	***95.59***
FinBERT	512	***100.0***	88.50	92.51
Higher level				
Concatenated	3737	*97.52*	96.0	93.49
Latent-fusion	256	93.54	93.0	*94.89*
Latent-fusion	512	94.89	94.69	92.11
Latent-fusion	1024	97.50	***96.32***	93.50
Latent-fusion	2048	95.57	94.40	92.24

Table 2: Internal evaluation of our models in terms of F1-score (%) on our internal data split. The *italic* scores represent the best-performing for each representation paradigm while the **bold** entries represent the best scores all-around.

As the *stylometric* features score the lowest, the next in line are the *LSA*-based features. They improve the scores by nearly 30% compared to the basic statistical features, achieving a score of 88.10%. The best performing methods for this category of methods that did not require fine tuning use a sentence-transformers trained on top of *distilBERT* [Sanh *et al.*, 2019], improving the performance over the *LSA*-based representation by 3%.

The *end2end* fine-tuned BERT models outperform the score of the sentence-transformers by 1.31% for the FinBERT variant and achieve the best score with LinkBERT, improving over FinBERT by 3.08% - reaching a score of 95.59% on our test set.

Finally, the higher level representations improve the performance over our standalone representations by 2.29% for the simple concatenated representations, while the latent representation improves over the naive concatenation by 0.14% - reaching a score of 94.89%.

The ranking of different representations is given in Figure 1, while Figure 2 represents the critical distance diagram between models. We also include the distribution of concepts found in the Knowledge Graph per label in the training set in Figure 3. We see that the distribution of concepts are extremely similar between sustainable and unsustainable sentence, despite unsustainable sentence supposedly not including any reference to ESG-related concepts. However, this analysis is not representative of the distribution of concepts in a full company financial reports, as the sentences in the train set might to have been sampled from reports in an uniform way; some bias might exist due to the sentence selection process during the annotation.

5.2 Final evaluation

We submitted two different approaches for the final evaluation. We opted for the deep latent representation from our standalone representations for the first submission. For the second submission we chose the ensemble of models *LinkBERT, FinBERT* and *latent fusion*, with arbitrary wheights of $2/4$ for LinkBERT and $1/4$ for the other two.

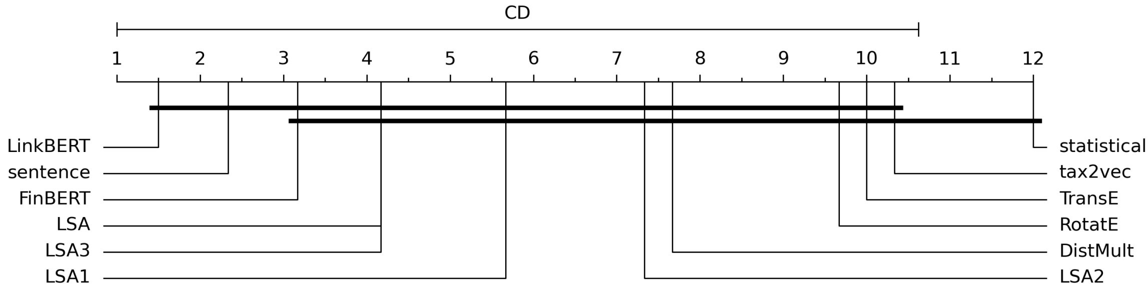

Figure 2: Critical distance plot representing the results of the Nemenyi test. Two classifiers are statistically significantly different in terms of F1-score if a difference between their ranks (shown in brackets next to the classifier name) is larger than the critical distance (CD).

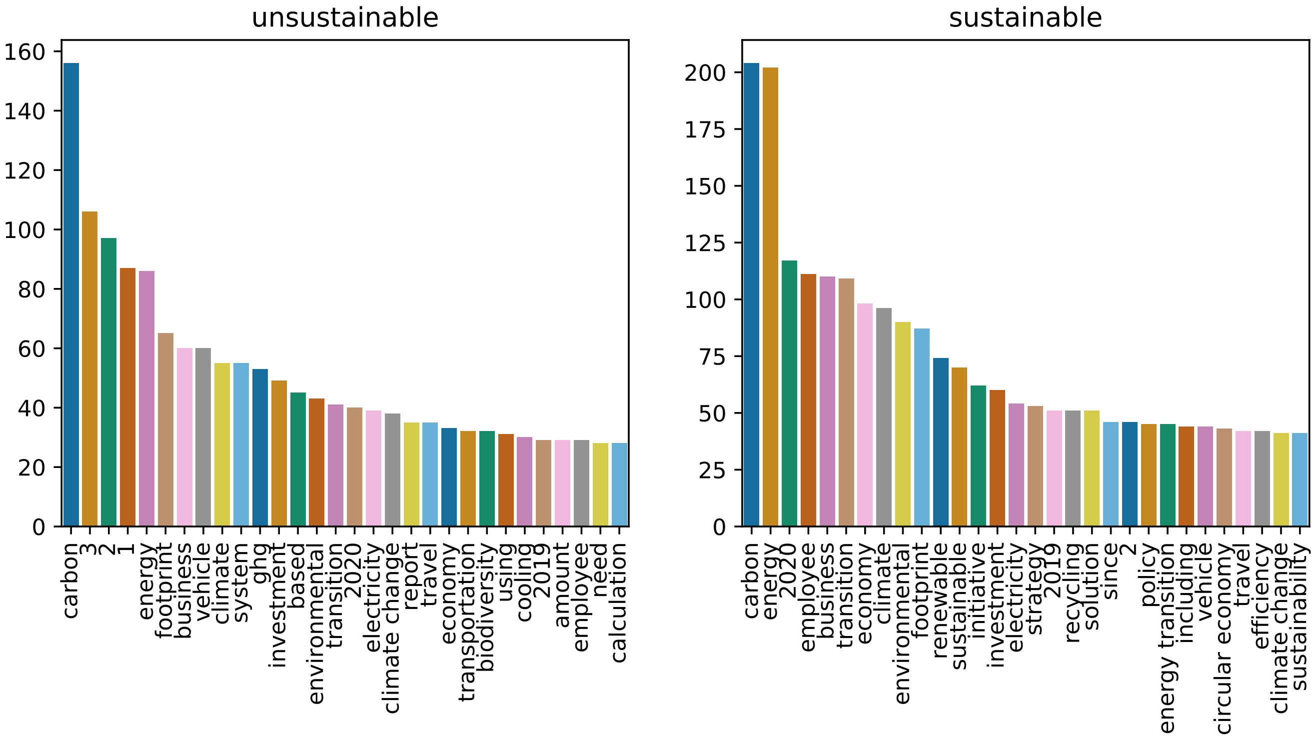

Figure 3: Distribution of extracted concepts from the WikiData5m knowledge graph in the knowledge-enrichment representations, by the respective label in the training set.

	precision	recall	f1-score
Latent representations (early-fusion)			
sustainable	0.86	0.92	0.89
unsustainable	0.91	0.84	0.88
weighted avg	0.89	0.88	0.88
Ensemble of models (late-fusion)			
sustainable	0.83	0.97	0.90
unsustainable	0.96	0.80	0.88
weighted avg	**0.90**	**0.89**	**0.89**

Table 3: Classification report of the final submissions. The **bold** entries represent the best scores between the two fusion approaches with respect to the average scores.

The *ensemble*-based approach achieves an accuracy of ***88.29%*** while the joint latent representation scores ***88.78%***. More granular report of the classification on the final test set is given in Table 3.

6 Conclusions and further work

In this work we developed a system for classification of *ESG* sentences. We used two representation paradigms: text-based and knowledge-based. In the text-based approaches we fine-tuned two BERT variants: *LinkBERT* and *FinBERT*. On top of the standalone representations we built ensembles on two different verticals: at the representation level, where we concatenated the representations and transformed them into a new latent space via SVD, and at the model level, where we stacked various models together for prediction of final labels. Our models scored competitively good, achieving nearly 89% in terms of accuracy. For further work, we consider training deep neural networks on top of the sentence representations to obtain more expressive deep representations that would improve classification performance. We also consider performing feature importance analysis on the representation-level ensembles, to see how representations in the heterogeneous stacks affect the classification on instance level. We also want to include domain-specific knowledge graphs or ontologies and explore their impact on the performance of the models. We also consider using background knowledge as a source for data augmentation, since for various use cases it contributes to better performance [Tang *et al.*, 2022; Shorten *et al.*, 2021; Cashman *et al.*, 2020]. Finally, we want to perform recursive dimensionality reduction to produce better fused document representations.

Availability

The code is available at https://gitlab.com/boshko.koloski/formicca-finsem-esg.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding (No. P2-0103), as well as for funding of the research project *Quantitative and qualitative analysis of the unregulated corporate financial reporting* (No. J5-2554).

References

[Amel-Zadeh and Serafeim, 2018] Amir Amel-Zadeh and George Serafeim. Why and how investors use esg information: Evidence from a global survey. *Financial Analysts Journal*, 74(3):87–103, 2018.

[Armbrust *et al.*, 2020] Felix Armbrust, Henry Schäfer, and Roman Klinger. A computational analysis of financial and environmental narratives within financial reports and its value for investors. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 181–194, Barcelona, Spain (Online), December 2020. COLING.

[Aula, 2010] Pekka Aula. Social media, reputation risk and ambient publicity management. *Strategy & leadership*, 2010.

[Balakrishnan *et al.*, 2010] Ramji Balakrishnan, Xin Ying Qiu, and Padmini Srinivasan. On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3):789–801, 2010.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[Cashman *et al.*, 2020] Dylan Cashman, Shenyu Xu, Subhajit Das, Florian Heimerl, Cong Liu, Shah Rukh Humayoun, Michael Gleicher, Alex Endert, and Remco Chang. Cava: A visual analytics system for exploratory columnar data augmentation using knowledge graphs. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1731–1741, 2020.

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

[Dumais *et al.*, 1988] Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285, 1988.

[Fellbaum, 1998] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[Guo *et al.*, 2020] Tian Guo, Nicolas Jamet, Valentin Betrix, Louis-Alexandre Piquet, and Emmanuel Hauptmann. Esg2risk: A deep learning framework from esg news to stock volatility prediction, 2020.

[Kogan *et al.*, 2009] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics, 2009.

[Koloski *et al.*, 2021] Boshko Koloski, Timen Stepišnik-Perdih, Senja Pollak, and Blaž Škrlj. Identification of covid-19 related fake news via neural stacking. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 177–188. Springer, 2021.

[Koloski *et al.*, 2022] Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, 496:208–226, 2022.

[Loughran and McDonald, 2011] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[Mansar *et al.*, 2021] Youness Mansar, Juyeon Kang, and Ismail El Maarouf. The finsim-2 2021 shared task: Learning semantic similarities for the financial domain. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 288–292, New York, NY, USA, 2021. Association for Computing Machinery.

[Mehra *et al.*, 2022] Srishti Mehra, Robert Louka, and Yixun Zhang. ESGBERT: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications*. Academy and Industry Research Collaboration Center (AIRCC), mar 2022.

[Nagy *et al.*, 2016] Zoltán Imre Nagy, Altaf Kassam, and Linda-Eling Lee. Can esg add alpha? an analysis of esg tilt and momentum strategies. *The Journal of Investing*, 25:113 – 124, 2016.

[Purver *et al.*, 2022 accepted] Matthew Purver, Matej Martinc, Riste Ichev, Igor Lončarski, Katarina Sitar, Valentinčič Aljoša, and Senja Pollak. Tracking changes in ESG representation: Initial investigations in uk annual reports. In *Proceedings of the First Computing Social Responsibility Workshop -NLP Approaches to Corporate Social Responsibilities (CSR-NLP I) 2022, co-located with LREC 2022*, 2022, accepted.

[Qiu *et al.*, 2006] Xin Ying Qiu, Padmini Srinivasan, and Nick Street. Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15, 2006.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[Sammut and Webb, 2010] Claude Sammut and Geoffrey I. Webb, editors. *TF–IDF*, pages 986–987. Springer US, Boston, MA, 2010.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Serafeim and Yoon, 2022] George Serafeim and Aaron Yoon. Stock price reactions to esg news: The role of esg ratings and disagreement. *Review of accounting studies*, pages 1–31, 2022.

[Shorten *et al.*, 2021] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34, 2021.

[Škrlj *et al.*, 2021] Blaž Škrlj, Matej Martinc, Nada Lavrač, and Senja Pollak. autobot: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, Apr 2021.

[Sokolov *et al.*, 2021] Alik Sokolov, Jonathan Mostovoy, Jack Ding, and Luis Seco. Building machine learning systems for automated ESG scoring. *The Journal of Impact and ESG Investing*, 1(3):39–50, January 2021.

[Sun *et al.*, 2019] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019.

[Tang *et al.*, 2022] Zhenwei Tang, Shichao Pei, Zhao Zhang, Yongchun Zhu, Fuzhen Zhuang, Robert Hoehndorf, and Xiangliang Zhang. Positive-unlabeled learning with adversarial data augmentation for knowledge graph completion. *arXiv preprint arXiv:2205.00904*, 2022.

[Tavchioski *et al.*, 2022] Ilija Tavchioski, Boshko Koloski, Blaž Škrlj, and Senja Pollak. E8-IJS@LT-EDI-ACL2022 - BERT, AutoML and knowledge-graph backed detection of depression. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 251–257, Dublin, Ireland, May 2022. Association for Computational Linguistics.

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

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[Yang *et al.*, 2014] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[Yang *et al.*, 2020] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097, 2020.

[Yasunaga *et al.*, 2022] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links, 2022.

[Škrlj *et al.*, 2021] Blaž Škrlj, Matej Martinc, Jan Kralj, Nada Lavrač, and Senja Pollak. tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language*, 65:101104, 2021.