Building Machine Learning Systems for Automated ESG Scoring

November 27, 2020

Alik Sokolov, Jonathan Mostovoy, Jack Ding, Luis Seco

Alik Sokolov is the Managing Director of Machine Learning at RiskLab: a global laboratory headquartered in Toronto that conducts research in financial risk management.
Email: asokolov@math.toronto.edu
Jonathan Mostovoy is the Managing Director of Research and Partnerships at RiskLab Toronto.
Email: mostovoy@math.toronto.edu
Jack Ding is a Researcher at RiskLab Toronto and a PhD Candidate at the University of Toronto with a research focus in symplectic geometry.
mail: xiao.ding@mail.utoronto.ca
Luis Seco is the Head of RiskLab Toronto, Director of the Mathematical Finance Program at the University of Toronto, Director of Fields-CQAM, Director of the GGSJ Centre, and CEO of Sigma Analysis & Management Ltd.
Email: seco@math.toronto.edu

Building Machine Learning Systems for Automated ESG Scoring

November 27, 2020

Key Takeaways:

1. The authors demonstrate the feasibility and advantages to applying state-of-the art Natural Language Processing (NLP) to identify ESG risks using social media data.
2. The authors discuss how advances in modern NLP can be leveraged to continuously build up algorithmic capabilities for processing ESG-relevant documents, by leveraging the capabilities of deep learning models for learning general representations of text data, which can then be applied across many tasks in the ESG domain.
3. The authors discuss results of NLP models can be used for creating aggregated ESG scores, as well design considerations for creating fully or semi-autonomous ESG scoring systems.

Abstract

Although investing in Environment-Social-Governance (ESG) driven portfolios is already a large and growing portion of global assets under management, applications of quantitative techniques to improve and standardize ESG scoring and the construction of ESG portfolios are underutilized. In this paper, we propose an approach to automatically convert unstructured text data into ESG scores by using the latest advances in deep learning for Natural Language Processing (NLP). We also show how a state-of-the-art NLP technique, BERT, can be incorporated to improve the accuracy of assessing relevance and content of documents in as ESG context using social media data as an example, and discuss the relevance of this approach to automating ESG scoring and constructing ESG portfolios.

---

1. Introduction

Environmental, social and governance factors, commonly abbreviated as ESG, have seen an explosion in interest from asset managers in recent years and are rapidly becoming key considerations for asset managers globally. Both retail and institutional investors are becoming increasingly focused on these factors in making investment decisions. It has been shown that corporations' financial results have a positive correlation with the levels of sustainability in their business models, and that ESG investment methodology can help reduce portfolio risk, while generating returns often competitive with those of traditional portfolios(Friede, Busch, and Bassen 2015). Increased demand combined with competitive performance has led to momentum effects and tremendous growth in ESG assets under management, which has in turn led to an increased demand for ESG due diligence and reliable data. However, one barrier for ESG scoring is the lack of relatively complete and centralized information source. Currently, ESG analysts generally leverage financial reports, investor calls, and regulatory disclosures to collect the the necessary data for proper evaluation, relying on companies to disclose pertinent information. However, the majority of corporate directors do not believe disclosures on sustainability matters are important in helping investors make informed decisions(Pinney, Lawrence, and Lau 2019), illustrating the importance of incorporating external data sources for additional validation.

At the same time, investors are increasingly demanding more stringent due diligence of the social responsibility levels of their portfolios. In Canada, the total amount of assets under management in Responsible Investing funds has grown 20% year over year 2015-2017 based on the latest growth numbers reported by the Canadian Responsible Investment Association (Association, n.d.). And according to research from Optimas LLC(Pierron 2019), ESG assets under management now total $30 trillion globally, up from $23 trillion in 2016, and projected to grow to $35 trillion by the end of 2020. According to research from BCG[1]1, Global assets under management in 2018 were at $74 trillion, meaning close to 50% of Global asset managers are incorporating ESG techniques into their investment methodology.(Henriksson et al. 2019)

As demand for ESG grows, investors are set to start demanding more accurate and timely responses to ESG issues, and the incorporation of data sources beyond companies' own reports, surveys, and regulatory filings. In the dynamic investment environment, incorporating relevant data from news and social media is primed to become a key component of ESG investment strategies. Machine learning approaches are in turn extremely promising, as they can limit the human cost of continuously monitoring vast volumes of diverse information related to ESG, while providing access to this information in a consistent, real-time reporting stream. We create a prototype for such a model, detail how it was created and how it can be improved; to the best of our knowledge, our study is the first detailed overview of the benefits of modern machine learning, including representation learning and transfer learning, in its potential applications to ESG. We also discuss the uses of such a model to automatically score the level of a company's exposure to ESG, and to build ESG-focused portfolios. To the best of our knowledge,

2. Literature Review

ESG investing and ESG portfolio construction are studied primarily from the point of view of financial returns and portfolio construction, as well as applying technology to the creation of ESG portfolios. Both of these areas of research mainly focus on the performance of ESG portfolios defined using human judgement, and have so far lagged behind in using the latest advances in machine learning, and specifically Natural Language Processing (NLP). In this paper, we show how start of the NLP can be applied to ESG scoring and the construction of ESG portfolios, as well as elucidate how modern NLP techniques can be applied to make such approaches practical.

The ESG industry has been enjoying significant growth, as seen through sharp increases in AUM detailed in the previous section. In addition, numerous other positive trends have emerged. For example, in recent years, portfolios created using ESG criteria have often demonstrated performance competitive with that of broad market portfolios. Climent and Soriano(Climent and Soriano 2011) show that between 2001-2009, green portfolios (ESG portfolios focused on the Environmental category) have had adjusted returns in line with conventional mutual funds. Some studies even demonstrate superior performance across specific time periods over the last several years, perhaps

due to momentum effects linked to AUM growth: a study by Amundi Asset Management (Nematzadeh et al. 2019) shows that, in the years 2014-2017, incorporating ESG criteria into portfolio construction was a source of excess returns, across all 3 ESG pillars, and in both North America and the Eurozone. And although there is not a clear consensus on the relationship between ESG performance and returns, a large panel study of over 2,200 studies (G., T., and A. 2015) found that 90% of research papers studied found at least a non-negative relationship between ESG ratings on corporate financial performance.

These positive trends have led to significant momentum within the ESG industry, and as a result a rapid spike in interest for ESG data and reliable ESG ratings. Such ESG criteria have been incorporated as part of portfolio construction methodologies; Henriksson et al.(Henriksson et al. 2019) provide a survey of ESG usage for portfolio construction, and show that incorporating industry-specific data for ESG scoring has the potential of creating better ESG portfolios. They also show that ESG criteria have a significant impact on company valuations (but not excess returns), emphasizing the importance of deep and reliable ESG data.

Henriksson et al. (2019) also discuss limitations on current ESG rating methodologies, where input data is typically voluntary, is very sparse, and scoring methodologies are variable across ESG data vendors. This is a significant bottleneck to assessing the ESG performance of companies, especially when coupled with the manual effort and required to continuously source and evaluate this data. The manual effort bottleneck is even more significant when dealing with large volumes of text data, such social media or news data. The approach of ignoring media altogether is unsatisfactory, as many pertinent issues may not be reflected in company financial and regulatory filings, and those filings may significantly lag emerging issues. This creates a strong incentive to use NLP techniques to incorporate this data at scale: several approaches to incorporating such sources via NLP techniques have been studied, summarized below.

Chen et al. (Chen et al. 2019) incorporate news data into a system that predicts stock market movement returns, fusing representations of news data regarding corporate events (including ESG events) with stock time series representations. Nematzadeh et al. (Nematzadeh et al. 2019) address the need to capture discrete controversies by proposing a clustering approach based on manually engineered features. However, research linking recent advances in NLP to the creation of ESG ratings has been very limited. We propose taking a more focused approach to learning text representation (Bengio, Courville, and Vincent 2013) relevant to ESG classification, by creating a human-curated ESG labels at the document level, and utilizing a state-of-the-art NLP representation (Devlin et al., n.d.) as a starting point using transfer learning.

3. Definitions and Notation

Definition 1
We define the *ESG Category* of each *document* d_i as $\{c_j(d_i) : i, j \in , 1 \leq i \leq N_{documents}, 1 \leq j \leq N_{esg\,categories}, and\ c_j(d_i) \in \{0, 1\}\}$, as an indicator function that specifies whether a document belongs to a particular ESG category. We use these as the labels for training the ESG classifier, and evaluate model performance based on comparing $p_j(d_i)$, the predicted probabilities for each document d_i belonging to category j.

Definition 2
We also define a set of documents $\mathscr{D}_t$ as the set of documents generated at time t. We consider daily scores for the purposes of our scoring discussion, so $\mathscr{D}_t$ will refer to all documents generated on a given day.

Definition 3
We define an *ESG score* as an aggregation function $F(t, \mathscr{D}_t)$, which combines the predictions p_i on each day in order to evaluate the aggregate performance of a company in terms of each *ESG Category*, based on a set of input documents for that day. These scores are similar in nature to ESG ratings already prevalent within the industry, but we focus on automating their evaluation and detail several approaches to creating company-level scores using a document-level NLP model. We describe some alternatives for defining $F(t, \mathscr{D}_t)$ in Section 6.

4. Approach

An essential component to our approach is a methodology for processing unstructured text data associated with many documents (Tweets in the context of our study). Historically, a common approach for such a task was creating text ontologies, or lists of words, which could then be used to separate documents into topical categories

(Lee, Tsao, and Chu 2009). Recent advancements in the areas of deep learning have pushed the state-of-the-art in similar NLP tasks towards Transformer-based deep learning model architectures such as BERT (Devlin et al., n.d.). As is typical with modern NLP benchmarks such as GLUE (Wang et al. 2019), we take the approach of curating a "labeled" data set classification a sample of ESG tweets into topic areas, detailed in Section 4.1.

A key advantage of Transformer-based (Vaswani et al. 2017) models is the ability to jointly perform tasks at the document level (i.e. classifying documents into topic areas), and token- or word-level tasks (which may involve more complex rules to be parsed out from the document, such a mention of a company in a specific context). In our study, we focus on the classification tasks, given the very short average length of tweets, as well as our desire create initial ESG representations that can then be applied to more complex subsequent tasks, as explored in section 4.4.

We take a standard approach for fine-tuning a BERT language model for document-level classification to build our ESG NLP model, broadly as labelled data generation (4.1), model training (4.2) and evaluation (4.3), and operationalization (5) . Our aim is to show how modern state-of-the art NLP approaches can be applied in determining the ESG content of news and social media (using Twitter as a prototype data source), as well as show how such a model can be operationalized to create a near real-time ESG scores. We also aim to illustrate the advantages of such an approach over what was previously discussed on ESG scoring and portfolio construction in literature.

An important benefit of a deep-learning approach is that each subsequent task requires less training data, as the learned representations can then be used in a feedback loop in place of a manual feature engineering for many related tasks. As Transformer-based architectures have the ability to tackle both document- and token-level tasks jointly, the models we use can be applied to even more complex tasks, creating even more robust representations and helping solve some of the key challenges we detail in Sections 5 and 6.

4.1. Data & Labelling

We opt for a document-level classification approach in designing the learning task for our initial NLP model. The primary application for our NLP model we considered was to measure the average number of documents linked to ESG over a prolonged period of time. Although the NLP model we chose supports both document and token (word) level tasks, we opt to keep our initial NLP modeling at the document level. This is primarily due the low average token lengths of our documents (tweets), as well as our initial goal being an aggregation of documents to determine relative levels of ESG risk.

- We align our ESG categories to MSCI ratings (Climent and Soriano 2011), and select a number of topic based on likelihood to be observed in the social media sphere.

- Our data was obtained using the Twitter API; in order to make the amount of manual labeling tractable, we pre-filter each ESG category with a set of keywords, which were then further refined using a level of human review.

- Our final training data comprises of labeled pre-filtered tweets (both relevant and irrelevant to ESG), as well supplementary unfiltered tweets labeled negative (assumed to be ground-truth negatives due to the sub-1% incidence of ESG topics occurring in randomly sampled tweets, which we expect will have minimal impact on the model). Our final training dataset contains 6K unique tweets, with 1,468 tweets labeled as belonging to one of the ESG topics, and the remainder deemed unrelated to ESG issues after a level of human review.

Our data set covers 10 ESG topics:

- Governance: Business Ethics, Anti-Competitive Practices, Corruption & Instability.

- Social: Discrimination, Health & Demographic Risk, Supply Chain Labour Standards or Labour Management, Privacy & Data Security, Product Liability.

  - The Discrimination category was added by the RiskLab team due to the prevalence of such issues being expressed on social media.

- Environmental: Climate Change and Carbon Emissions, Toxic Emissions & Waste.

- We also model whether a particular document represents an ESG risk, to account for cases where the document captured represents a positive action towards a particular ESG category (e.g. a company announcing a significant donation to fight climate change).

4.2. Model Training

We utilize a BERT classifier for this problem (Devlin et al., n.d.). BERT is a pre-trained language model, which carries the benefit of being initialized with the ability to map documents to good general representations. These representations can be used to process documents in a variety of ways, including for classification, which is the primary capability we leverage. In addition, we are also able to specialize or "fine-tune" the model to our specific task of ESG analysis using the labeled dataset we created, which generally performs better than training a standalone model (Zhuang et al. 2020).

We utilize a standard pre-training approach with several slight modifications. BERT has recently demonstrated state-of-the-art result on a variety of NLP tasks, and has shown to be robust across many different domains, and ultimately yielded as the best performance across other Transformer-based (Vaswani et al. 2017) architectures tested, such as ALBERT (Lan et al. 2020). We utilize the pre-trained classifier version as provided by HuggingFace (Wolf et al. 2019) due to the well known benefits of transfer learning (Zhuang et al. 2020), which allows us to train a better specialized model with less training data by utilizing a general language model as a starting point. We also add an additional fully-connected layer between the BERT embedding and output layers, in order to simplify fine-tuning. This is achieved through having the initial classifier (trained with the BERT encoder layers frozen) reach a better minimum prior to fine-tuning the BERT layers, as opposed to a single linear output layer. This therefore limits the changes required to the encoder layers of our pre-trained BERT when fine tuning.

Below is the full list of hyperparameters we used to train our classifier:

- We use an output size of 768 and a max sequence length of 128 for the BERT encoder layers, a batch size of 32, and a dropout rate of 0.1.
- We add one hidden layer, of dimension 256, right after the initial (CLS token) BERT embedding, with dropout of 0.5.
- We use the binary cross entropy loss for each category, as the ESG issues may not be mutually exclusive.
- We utilize a two-stage pre-training approach, with a patience of 5 epochs. In the first phase, we train the hidden and output layer; in the second, the full network (with all 10 BERT encoder layers) is fine-tuned.
- We utilize the RADAM (Liu et al. 2020) optimizer with learning rates of $1e^{-4}$ during the first stage of training, and $2e^{-5}$ during the second stage of training (where BERT encoder layers are fine-tuned); the RADAM optimizer simplifies the learning rate hyperparameter search, and removes the need running "warm-up" epochs. We used no warm-up for training of our model.

4.3. Model Evaluation

We optimize our model based on cross-entropy loss, with a test set fixed as a randomly 30% sample of our full dataset.

We evaluate our final models based on the area under the Precision-Recall curve (PRAUC), as well as ROC AUCs for each class to assess model performance. We find that this model, although a prototype, showcases the potential of applying such models at scale. The ROC and PR AUCs for each class, as computed on our holdout set, are presented in Table 1.

ESG Category	**ROC AUC**	**PR AUC**
Anti-Competitive Practices	0.97	0.57
Business Ethics	0.88	0.36
Climate Change	0.94	0.31
Corruption & Instability	0.98	0.55
Discrimination	0.99	0.71
Health & Demographic Risk	0.91	0.19
Privacy & Data Security	0.97	0.66
Product Liability	0.96	0.66
Supply Chain Labour Policies	0.97	0.76
Toxic Emissions & Waste	0.96	0.41
ESG Risk	0.90	0.22
Not an ESG Risk	0.91	0.51

Table 1: NLP Model Evaluation

These results, although they can undoubtedly further improved through collecting additional ground-truth observations, already showcase the potential for applying these models in production. Consider the Precision-Recall curve in Figure 1. Our model for this class has a precision of 60% at an 60% recall, meaning the majority of documents pertaining to ESG issues, with 60% of the retrieved documents being true examples of Privacy & Data Security issues. These performance measures are promising, but given the massive volumes of social media and news data generated daily, would need to be improved further to create a production-ready ESG scoring solution.

It is likely that significant improvements can be made with the addition of more labeled data, as the performance of deep learning models has historically been vastly improved by scaling up the amount of data available for training. The ability to squeeze out additional performance by scaling deep learning models up as additional data becomes available has been consistent across domains(Russakovsky et al. 2015; Devlin et al., n.d.).

As such, the main challenge in creating robust classifiers across different document types and ESG issues lies with curating a trusted dataset with a sufficient number of diverse examples. We also emphasize that these results are based on training and evaluation the model using Twitter data, and the model would likely need to be fine-tuned in order to apply it to other data sources, such as news or SEC filings. A key benefit to our work, however, is to create initial general ESG representations, that can then be fine-tuned for more complex tasks by collecting additional data.

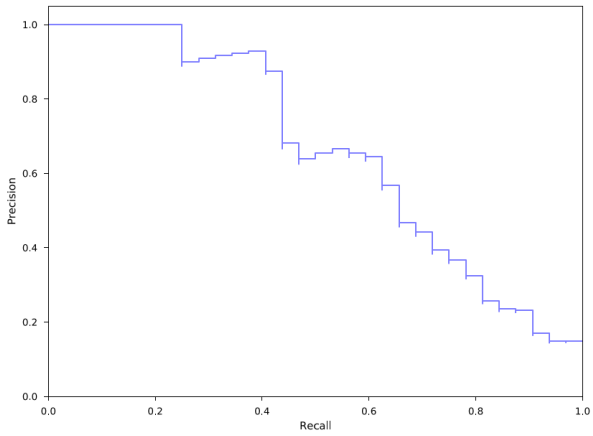

Figure 1: Precision-Recall: Privacy & Data Security

Collecting such data would allow us to create even more robust representations of text data suitable towards evaluation ESG issues. These representations can then be used for creating automated, stable and standardized ESG scores that reflect the overarching objectives of ESG investing. We discuss how a document-level model can be used to create company-level scores, as well as the additional tasks our proposed system would need to incorporate in production, below.

4.4. Exploring ESG Representations

Representation learning enables much greater transferability of models across tasks, which is perhaps the primary benefit of modern deep-learning driven NLP. Some of these benefits can be illustrated visually by examining the inner workings of our network, which we showcase below. As our model learns to distinguish between various ESG issues (an non-issues), intermediate "representations" of text data are created inside the network. We can extract these representations to help us better understand the totality of ESG issues we encounter, and later on use these same representations to help us tackle more complex tasks from a better starting point.

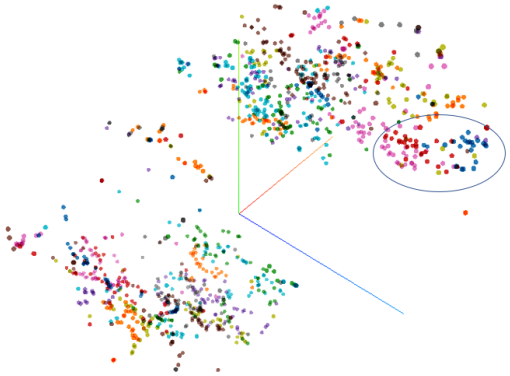

Figure 2: Clusters corresponding to ESG issues

Figure 2 shows a simplified view of ESG issues in our evaluation data set, plotted using the extracted representations from our model. If two tweets are close to one another in this chart, their numerical representations are similar. The 3 larger "clusters" loosely correspond to Environmental, Social, and Governance issues. The colors represent various issues or themes, determined based on how similar the learned representations are across documents.

Even with a moderate amount of labeled data,

our model is already able to distinguish various sub-issues and emerging themes. For example, a common issue with privacy and data security is data breaches, and indeed the blue cluster in Figure 3 contains several examples of data breaches being discussed in the public domain.

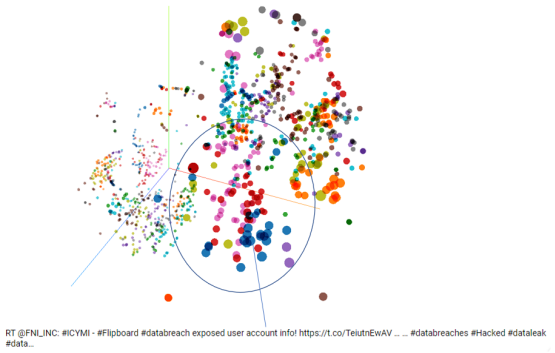

Figure 3: The data breach cluster

On the other hand, similar discussions that appear very similar can represent little ESG risk, as seen in Figure 4. The nearby pink cluster contains tweets related to cybersecurity education, and discussions on how to limit cyber risk. Both of these issues belong to the Privacy & Data Security category of ESG risk, but are being treated differently by our model because of their relation to other classes, with the pink cluster naturally separating and on average being viewed as less likely to constitute an ESG risk by our model.

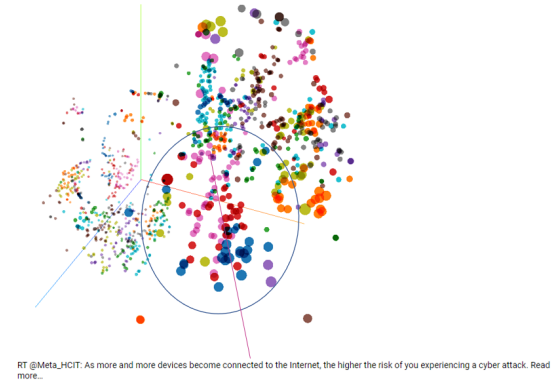

Figure 4: Risk-free cluster

It is encouraging that our model is able to distinguish these subtle differences, and naturally separates risky and non-risky issues. This is done in part by priming our model with labels in addition to ESG categories, such as whether a tweet corresponds to an ESG risk or a positive action related to an ESG issue. By adding these kinds of additional information during the labeling processing and representing these as additional training tasks, we expect the learned representations to become more useful towards understanding ESG risk. Examining representation is a valuable technique, but also carries a very practical and tangible benefit for ESG risk detection. Measuring "drift" in the representation stage can be indicative of emerging risks previously unseen in the data, and these types of analyses can also inform the labeling strategy and in turn lead to even better representations, creating a powerful feedback loop.

5. Using NLP Model Results for ESG Scoring

Our paper demonstrates how state-of-the art machine learning techniques can be applied to improve the parsing of text data, and demonstrates how such models can be creating by detailing the creation of our own prototype model. In order to make the outputs of the model we describe actionable, one has to deal with the additional complexity of converting document-level model outputs, which roughly correspond to the probability of a document belonging to a particular ESG category, into company-level scores. These scores, which update continuously, can then be utilized to construct portfolios that continuously pivot away from ESG exposure, which may be in closer alignment to investor preferences. There are many different ways to accomplish this, for example by incorporating the scores into the views vector in a Black-Litterman model (Black and Litterman 1991)). The scores can also be used as an alternative, or in addition, to existing ESG ratings.

We propose several alternative approaches through which an ESG score can be constructed from unstructured text data sources. All of our proposed approaches involve aggregating the output of the classifier after it is applied to a large number of documents over a period of time. There are several practical considerations that need to be considered to adopt an automated document classifier into an ESG scoring pipeline. One approach is to simply construct a score based on the average predicted probabilities for each ESG Category, by dividing the sum predicted probabilities $p_j(d_i)$ for each category j by the total number of documents for the day N_t.

$$R_j(t) = \sum_{d_i \in \mathscr{D}_t} \frac{1}{N_t} p_j(d_i)$$

Although intuitive, this approach relies on the learned probability estimates for each class corresponding to their real-life probability distributions; when dealing with text documents, this approach is often unreliable as the distribution of the training data may vary in proportion to the real-life data, and the distribution may shift significantly over time. This leads to the estimated probabilities oftentimes not being meaningful as direct estimates of real-life probabilities.

A more robust approach may be to define the score by using our predicted class labels:

$$R_i(t) = \sum_{d_j \in \mathscr{D}_t} \mathbb{1}_i(p_i(d_j))$$

Where $\mathbb{1}_i$ is the indicator function corresponding to class specific cut-offs ε_i, such that:

$$\mathbb{1}_i(p_i, \varepsilon_i) = \begin{cases} 1, & \text{if } p_i \ \geq \varepsilon_i \\ 0, & \text{if } p_i \ < \varepsilon_i \end{cases}$$

Although this approach can be an improvement, it still presents a handful of complications: probability outputs may be proportionally off depending on ratios of training data supplied to the model, and may require manual tuning, and so thresholds ε_i can be difficult to set for each ESG category. In addition, such aggregation approaches are still susceptible to unrelated spikes in chatter (e.g. viral marketing campaign may improve overall ESG scores).

Another improvement may be to first group documents into adverse events. This may be desired as it more closely reflects the overarching business purpose behind ESG monitoring, helping identify emerging issues and controversies by grouping related documents together. Leveraging representations as shown in Figures 3 and 4 can be extremely useful not only for detecting new emerging events, but also for grouping similar documents into events to link documents pertaining to the same issue. This is also an area where token-level capabilities of Transformer models can be leveraged extensively, by building labeled data for extracting mentions of specific types of events or relationships.

With all of these approaches, there is a number of additional considerations that have to be accounted for in order to aggregate document-level model outputs into useful company-level scores. These considerations may have different optimal ranges depending on the applications for the ESG score, and include:

1. Period of time across which ESG scores are computed; as mentions around ESG topics as well as ESG disclosures may not occur very frequently for most companies, a longer time period may be required to ensure that statistically relevant samples of documents are collected, and meaningful changes are being measured; at the same time, scores should be assessed frequently enough to leverage near real-time capabilities as a core advantage of an automated approach. A reasonable approach to combining longer time periods with more frequent updates is to leverage a time-discounting approach such as exponential smoothing.

2. ESG scores need to be carefully normalized, to account for naturally-occurring differences such as some companies having higher average levels of discourse around them than others, which companies should not be penalized for; in addition, spikes in volume of mentions may not always correspond to spikes in the number of adverse ESG events (Retweets are a basic example); therefore, some level of event de-duplication as described above should be employed, although the volume and sentiment of discussion around a specific event is also an important consideration

3. Not all ESG categories should be equally weighted, but it is difficult to assign relative importance to different ESG categories objectively. As such, our proposed approach to leveraging multiple ESG categories is to provide relative weights as a user input where appropriate for the use case. Alternatively, weights can also be learned by leveraging the created ESG scores in a relevant supervised learning task as a proxy, such as whether a security passed ESG screening in the past, whenever such data is available.

6. Future Work

One of the primary areas of future work will be to improve the training data to create a ESG dataset benchmark for NLP, initially focusing on social media data. Given the varied data sources that the industry relies on, it is likely that several

different types of NLP models will be need to incorporate diverse text data such as news, SEC filings, and ESG disclosures. We believe that having benchmark datasets for all of these will be critical to the evolution of the ESG data industry, and will look to contribute to the creation of such benchmarks. Some of the improvements we will aim to incorporate to create the first NLP benchmark for ESG will be:

1. Implement redundancy, such as majority voting, around each proposed document label to ensure data quality
2. Obtain additional training data to ensure broader coverage across time periods, companies, and events
3. Expand the dataset beyond Twitter and into other data sources (other social media such Reddit, blogs and forums, etc., news, SEC filings, ESG disclosures)
4. Expand the dataset to additional token-level tasks described below

There are also additional NLP tasks that would have to be incorporated into an NLP system that captures ESG attitudes across a sufficiently wide number of sources. For example, token level entity recognition models can be used to jointly detect mentions of ESG issues together with references to a company, directly linking a company to an event. In addition, the action of a particular company in relation to an ESG topic may be positive or negative, as captured in the "ESG Risk" classification component of our model. An NLP ESG scoring system would need to appropriately reflect the correct attitudes and relationships in order to limit false positives, especially in datasets with extremely low signal-to-noise rations such as social media.

We will also aim to expand the number of ESG issues being considered for this problem, and capture additional issues prevalent in public discourse but not captured in traditional ESG frameworks.

We will also expand this work into creating ESG portfolios, and evaluate various ESG scoring and ESG portfolio construction methodologies as discussed above, and their impacts on portfolio construction, including stability, re-weighting and associated transaction costs. Lastly, we will evaluate ESG ratings for their efficacy as an additional criteria for portfolio construction, and their impact on portfolio risk-return profiles.

Bibliography

Association, Responsible Investment. "2018 Canadian Responsible Investment: Trends Report." https://www.riacanada.ca/content/uploads/2018/10/2018-RI-Trends-Report-FINAL-WEB-1.pdf. (accessed: 30.11.2019).

Bengio, Y., A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 8 (2013): 1798–1828. 10.1109/TPAMI.2013.50.

Black, F., and R. Litterman. "Asset Allocation." *The Journal of Fixed Income* 1, no. 2 (September 1991): 7–18. https://doi.org/10.3905/2Fjfi.1991.408013. 10.3905/jfi.1991.408013.

Chen, D., Y. Zou, K. Harimoto, R. Bao, X. Ren, and X. Sun. "Incorporating Fine-grained Events in Stock Movement Prediction." In, 31–40. January 2019. 10.18653/v1/D19-5105.

Climent, F., and P. Soriano. "Green and Good? The Investment Performance of US Environmental Funds." *Journal of Business Ethics*, 2011.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." *North American Association for Computational Linguistics (NAACL)*.

Friede, G., T. Busch, and A. Bassen. "ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies "Volume 5, nos. Issue 4 (2015): p. 210–233, 2015. 10.1080/20430795.2015.1118917.

G., Friede, Busch T., and Bassen A. "ESG and financial performance: aggregated evidence from more than 2000 empirical studies." *Journal of Sustainable Finance & Investment* 5, no. 4 (2015): 210–233. https://doi.org/10.1080/20430795.2015.1118917. 10.1080/20430795.2015.1118917.

Henriksson, R., J. Livnat, Pfeifer P., and Stumpp M. "Integrating ESG in Portfolio Construction." *The Journal of Portfolio Management*, nos. 45 (4) (2019): 67–81.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In *International Conference on Learning Representations*. 2020. https://openreview.net/forum?id=H1eA7AEtvS.

Lee, Y., W. Tsao, and T. Chu. "Use of Ontology to Support Concept-Based Text Categorization." In, 201–213. Vol. 22. January 2009. 10.1007/978-3-642-01256-3_17.

Liu, L., H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. "On the Variance of the Adaptive Learning Rate and Beyond." In *International Conference on Learning Representations*. 2020. https://openreview.net/forum?id=rkgz2aEKDr.

Nematzadeh, A., G. Bang, X. Lui, and Z. Ma. "Empirical Study on Detecting Controversy in Social Media," 2019.

Pierron, A. "ESG Data: Mainstrean Consumption, Bigger Spending." 2019. http://www.opimas.com/research/428/detail/. (accessed: 30.11.2019).

Pinney, C., S. Lawrence, and S. Lau. "Sustainability and Capital Markets—Are We There Yet?" *Journal of Applied Corporate Finance* 31, no. 2 (2019): 86–91. https://onlinelibrary.wiley.com/doi/abs/10.1111/jacf.12350.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, et al. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3 2015): 211–252. 10.1007/s11263-015-0816-y.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. "Attention is All You Need." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. NIPS'17. Long Beach, California, USA: Curran Associates Inc. 2017.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman R. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In *International Conference on Learning Representations*. 2019. https://openreview.net/forum?id=rJ4km2R5t7.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, et al. "Huggingface's transformers: State-of-the-art natural language processing," 2019.

Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A Comprehensive Survey on Transfer Learning.