Building Machine Learning Systems to Automate ESG Index Construction

May 19, 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: alik.sokolov@gmail.com
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, and CEO of Sigma Analysis & Management Ltd.
Email: seco@math.toronto.edu

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 modern NLP algorithms can improve the ability of investors to anticipate unforeseen ESG risks
3. The authors discuss applications of such systems to ESG investing and index construction, as well as algorithm design for creating a fully or semi-autonomous ESG rating system

---

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 ESG index construction are underutilized. In this paper, we propose an approach to constructing ESG indexes using news and social media data, combined with deep learning techniques 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, and discuss the relevance of this approach to automating the construction of ESG indexes.

---

1. Introduction

The financial performance of a corporation is correlated with its social responsibility, e.g. the environmental impact of their products and supply chain, employee safety records, or social issues such as utilization of child labour[3]. Both retail and institutional investors are becoming increasingly concerned about these factors in making investment decisions. Environmental, social and governance factors, commonly abbreviated as ESG, are rapidly becoming a key considerations for asset managers globally. It has been shown that corporations financial results have a positive correlation with their sustainability business model and the ESG investment methodology can help reduce portfolio risk, while generating competitive returns.[4]. However, one barrier for ESG evaluation 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[10] do not believe disclosures on sustainability matters are important in helping investors make informed decisions, 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 [8]. And according to research from Optimas LLC[9], ESG assets under management now total $30 trillion globally, up from $23 trillion in 2016, and projected to grow to $35 trillion by 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.[5]

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 information reporting on various ESG issues, while providing access to this information in consistent, real-time

reporting stream.

2. Literature Review

ESG investing and index creation 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. The former research primarily focuses on the performance of ESG portfolios defined using human judgement, whereas the latter category has so far lagged behind 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 the creation of practical indices, as well as elucidate how modern NLP techniques can be applied to make such approaches practical (see the discussion of additional NLP tasks in the *Index Construction* section.

Portfolios that are created using ESG criteria have been shown to generally be inline with the overall market returns in recent years. Climent and Soriano[1] show that between 2001-2009, green portfolios (ESG portfolios focused on the Environmental category) have had adjusted returns in line with conventional mutual funds. An Amundi Asset Management study [7] 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. These results may be a reflection of growing demand for ESG portfolios in those years, as shown above.

ESG criteria have also been incorporated as part of portfolio construction methodologies; Henriksson et al.[5] 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 find that previous studies were able to conclusively show that companies that score higher on ESG criteria have better valuations, but not necessarily excess returns, indicating that at least to some extent ESG criteria are being priced in.

Henriksson et al. [5] also discuss limitations on current ESG scoring 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 ESG portfolio construction, 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 social media and traditional media. 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. [**incorporating˙news**] 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. We propose taking a more focused approach to learning text representation relevant to ESG classification, by creating a human-curated ESG labels at the document level, and utilizing a state-of-the-art NLP representation [2] as a starting point using transfer learning.

Nematzadeh et al. [7] address the need to capture discrete controversies by proposing a clustering approach based on manually engineered features.

Our approach leverages advancements in modern NLP, by which we can create language representations through solving ESG classification and subsequent (segment- and token-level tasks jointly). These joint representations can then be used in a feedback loop in place of a manual feature engineering for many adjacent tasks, such as adverse event detection.

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 ground truth 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 indices for the purposes of our indexing discussion, so $\mathscr{D}_t$ will refer to all documents generated on a given day.

Definition 3
We define an *index* as an aggregation function $F(t, \mathscr{D}_t)$, which aggregates 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. We describe some alternatives for defining $F(t, \mathscr{D}_t)$ in *Section 4.4*.

4. Approach

We took the standard machine learning project approach to creating our ESG classifier, broadly as (3.1) ground truth generation, and (3.2-3.4) model optimization and operationalization. Our aim was 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 index. We also aim to illustrate the advantages of such an approach over what was previously discussed on ESG index construction in literature.

4.1. Data & Labelling

Our approach to ground truth generation was as follows:

- We aligned our ESG categories to MSCI ratings [1], and selected a number of topic based on likelihood to be observed in the social media sphere
- Our data was obtained using the Twitter API, with tweets ranging from June 16 to July 22, 2019. In order to make the amount of manual labeling tractable, we pre-filtered each ESG category with a set of keywords, which were then further refined using a level of human review
- Our final training data comprised of labeled pre-filtered tweets (positive and negative), as well supplementary unfiltered tweets labeled negative (assumed to be ground-truth negatives due to the rarity of ESG topics occurring in randomly sampled tweets). Our final training dataset contained 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
    - The Discrimination category was added by the RiskLab team due to the prevalence of such issues being expressed on social media
- Environmental: Climate Change or Carbon Emissions, Product Quality & Safety, Toxic Emissions & Waste

4.2. Model Training

We utilized a standard BERT classifier for this problem [2], with a couple of slight modifications. We utilize the pre-trained classifier version as provided by HuggingFace [12]. We also add an additional full-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 used 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 our classifier, as the ESG issues may not be mutually exclusive
- We utilized 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 utilized the RADAM [6] 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

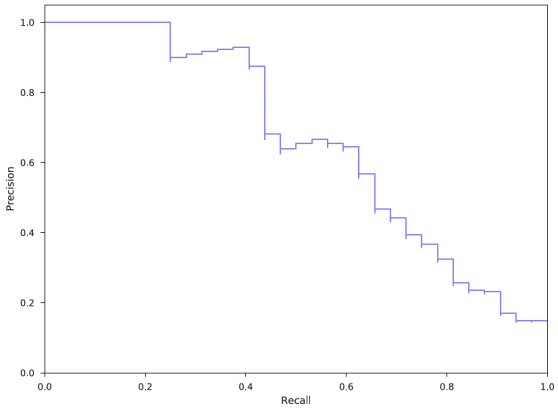
We optimized our model based on cross-entropy loss, with a test set fixed as a randomly 30% sample of our full dataset.

We evaluated 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 below:

ROC-AUC & PR-AUC		
ESG Category	ROC	PR
Corp. Behaviour: Business Ethics	0.88	0.36
Corp. Behaviour: Anti-Comp. Practices	0.97	0.57
Corp. Behaviour: Corruption & Instability	0.98	0.55
Privacy & Data Security	0.97	0.66
Human Cap.: Discrimination (by RiskLab)	0.99	0.71
Human Cap.: Health & Demographic Risk	0.91	0.19
Human Cap.: Supply Chain Labour Policies	0.97	0.76
Climate Change: Carbon Emissions	0.94	0.31
Pollution & Waste: Toxic Emissions & Waste	0.96	0.41
Product Liability: Quality & Safety	0.96	0.66
ESG Risk	0.90	0.22
Not an ESG Risk	0.91	0.51

These results, although they can undoubtedly further improved through collecting ground-truth observations, and through additional refinements of the NLP model, already showcase the potential for applying these models in production. Consider the PR curve presented below.



Precision-Recall: Privacy Data Security

Our model for this class has a precision of $\tilde{6}0\%$ 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 indexing solution. It is likely we would be able to generate such improvements with the addition of more labeled data, as we have historically been able to vastly improve the performance of deep learning models 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[11, 2].

These results show promise for the creation NLP-enabled ESG indexes. 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.

Collecting such data would allow us to create robust representations of text data suitable towards evaluation ESG issues. These representations can then be used for creating practical, stable indices that reflect the overarching objectives of ESG investing. We discuss the construction of such indices, as well as the additional tasks our proposed system would need to incorporate in production, below.

4.4. Index Construction

We propose several alternative approaches through which an ESG index can be constructed from unstructured text data sources by aggregating the output of our classifier applied to large collections of

dated documents, as well as propose some future improvements to make such indices more robust.

There are several practical considerations that need to be considered to adopt an automated document classifier into an indexing pipeline. One approach is to simply construct a rating 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 sigmoid functions for each class corresponding to real-life probabilities; 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 index 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
- 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.

This is another area where modern NLP carries significant benefits, as we are able to better reason about unstructured text data by examining the inner workings of our network. 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:

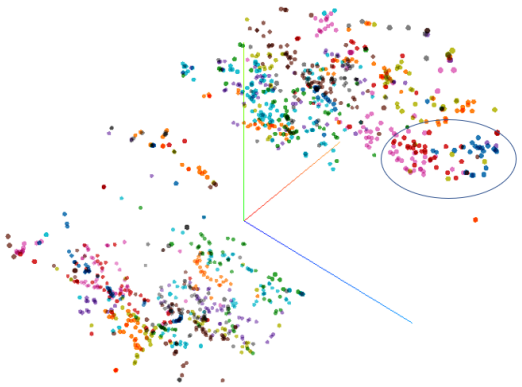


Figure 1: Clusters corresponding to ESG issues.

The chart above 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 were 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 below contains several examples of data breaches being discussed in the public domain:

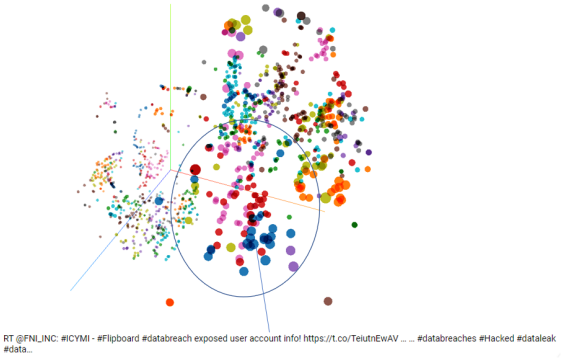


Figure 2: The data breach cluster.

On the other hand, similar discussions that appear very similar can represent little ESG risk. 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:

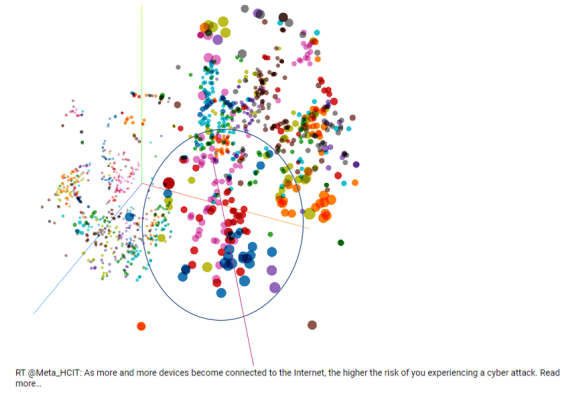


Figure 3: Risk-free cluster.

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 level of risk or relevance to a specific company. By adding these kinds of additional information during the labeling processing, we expect the learned representations to become more useful towards understanding ESG risk. Examining representation is a valuable technique, and can also be used as a feedback loop for the ESG model: measuring "drift" in the representation stage can be indicative of emerging risks, and these types of analyses can also inform the labeling strategy and in turn lead to even better representations.

5. Future Work

The primary area of future work will be to improve the training data to create a ESG dataset benchmark:

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 (Reddit, other social media, traditional news)

4. Expand the dataset to additional 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, a news article may mention an ESG issue together with a name of a company, without directly relating the company to the issue at hand. In addition, the action of a particular company in relation to an ESG topic may be positive or negative. An ESG indexing system incorporating NLP would have to take such cases into account, and evaluate documents to appropriately reflect the correct attitudes and relationships in such situations.

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 indexes, and evaluate various index construction methodologies as discussed above, and their impacts on portfolio construction, including stability, re-weighting and associated transaction costs.

Lastly, we will evaluate ESG indexes for their efficacy as an additional criteria for portfolio construction, and the impact on portfolio returns and risk profiles.

References

[1] Climent, F., Soriano, P. "Green and Good? The Investment Performance of US Environmental Funds". In: (2011).

[2] Devlin, J., Chang, M., Lee, K., and Toutanova, K. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *North American Association for Computational Linguistics (NAACL)* (). arXiv: 1810.04805 [cs.CL].

[3] Albert Einstein. "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]". In: *Annalen der Physik* 322.10 (1905), pp. 891–921. DOI: http://dx.doi.org/10.1002/andp.19053221004.

[4] Friede, G., Busch, T., Bassen, A. "ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies (October 22, 2015)". In: *Journal of Sustainable Finance Investment* Volume 5.Issue 4 (), p. 210–233, 2015. DOI: 10.1080/20430795.2015.1118917.

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

[6] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., J. Gao, J., and Han J. "On the variance of the adaptive learning rate and beyond". In: (2019). arXiv: 1908.03265 [cs.LG].

[7] Nematzadeh, A., Bang, G., Lui, X., Ma., Z. "Empirical Study on Detecting Controversy in Social Media". In: (2019). arXiv: 1909.01093 [cs.SI].

[8] Pierron, A. *ESG Data: Mainstream Consumption, Bigger Spending*. URL: http://www.opimas.com/research/428/detail/. (accessed: 30.11.2019).

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

[10] Pinney, C., Lawrence, S., Lau, S. "Sustainability and Capital Markets—Are We There Yet?" In: *Journal of Applied Corporate Finance* 31.2 (2019), pp. 86–91. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jacf.12350. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jacf.12350.

[11] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.,Berg, A.,Fei-Fei, L. *ImageNet Large Scale Visual Recognition Challenge*. 2014. arXiv: 1409.0575 [cs.CV].

[12] Wolf, T. , Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. "Huggingface's transformers: State-of-the-art natural language processing". In: (2019). arXiv: 1910.03771 [cs.CL].