Weak Supervision and Black-Litterman for Automated ESG Portfolio Construction

March 24, 2021

Alik Sokolov[1], *Kyle Caverly*[1], *Jonathan Mostovoy, Talal Fahoum, 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
Kyle Caverly is a Machine Learning Researcher at RiskLab Toronto.
Email: kylebcaverly@gmail.com
Jonathan Mostovoy is the Managing Director of Research & Partnerships at RiskLab.
Email: mostovoy@math.toronto.edu
Talal Fahoum is a research analyst at RiskLab.
Email: talal.fahoum@mail.utoronto.ca
Luis Seco is the Head of RiskLab Toronto, Director of the Mathematical Finance Program at the University of Toronto, CEO of GGSJ Centre, and CEO of Sigma Analysis & Management Ltd.
Email: seco@math.toronto.edu

[1] A.S.(Alik Sokolov) contributed equally to this work with K.C. (Kyle Caverly).

Key Takeaways:

1. The authors describe a theoretical framework for how an automated NLP system can be used to incorporate ESG criteria into portfolio optimization decisions.
2. The authors demonstrate the technical implementation details for incorporating ESG signals into augmented portfolio weights.
3. The authors demonstrate the competitive performance of such a portfolio through a long-term historical back-test with the S&P 500 Index.

Abstract

We propose an approach that combines modern machine learning techniques in Natural Language Processing with portfolio optimization to incorporate views of companies' ESG performance. This is automatically done through curating and subsequently converting large scale news-data into portfolio management decisions. We train a machine learning news data classifier to automatically identify several key ESG issues in news data over time. We then aggregate these issues over time to generate a "views vector" under the Black-Litterman portfolio framework, and finally compare the performance of an ESG-tilted portoflio against a standard Black-Litterman portfolio. Additionally, we show how this can be achieved at scale, in a fully automated manner, and with consistency over large periods of times. Our methodology thus demonstrates a reasonable and agile method for asset managers to incorporate ESG considerations into their portfolios free of any exclusionary frameworks and without sacrificing performance.

---

1. Introduction

The last several years have seen an explosion of interest in ESG, and a continued trend of growth in assets under management for ESG portfolios even throughout the COVID-19 pandemic. Recent years saw some of the world's largest and more prominent investors, including the world's largest private fund manager, BlackRock, and prominent sovereign and pension funds like Japan's GPIF, signal a willingness to commit to ESG investment principles. A recent research report by Morningstar [1] cites continued growth, with the number of sustainable passive funds tripling in the trailing five years leading to June 2019. Sustainable funds have been shown to benefit from significantly higher net inflow of funds in recent years [2], which is consistent with overall changes in assets under management.

The COVID-19 pandemic has served to accelerate these trends. A recent survey by J.P. Morgan demonstrated that the majority of investors see the crisis as a catalyst for ESG investing [3]. This appears to have been consistent with empirical performance: Canada has seen an especially significant uptrend in ESG investing according to TD Securities, with inflows into passive sustainable funds already more than tripling both 2018 and 2019 numbers as of June 2020 [4]. Additionally, recent studies such as S&P Global's paper [5], which focused on exchange-traded ESG funds and ESG mutual funds, as well as a Morningstar's [6], which focused on large-cap US ESG equity indices, show ESG funds outperforming their peers during the crisis. Although these results only represent a short period of time and are not based on robust statistical practices, at the very least, the press generated has the potential to push the demand for ESG funds even further.

In spite of the positive indicators detailed above, ESG data remains a significant challenge for the industry. Self-reported ESG data is extremely challenging to manage, with inconsistency and varying imputation methodologies bringing significant challenges [7]; a recent study has even shown that increased

corporate disclosures tends to produce higher variability of ESG ratings [8]. The lack of standardization in the rating industry, combined with views on a specific company impacting a rater's views of ESG performance across specific categories [9], mean that the availability and quality of ESG data has not kept up with the growth of the industry. As such, we propose a more automated approach be taken to incorporating unstructured ESG information into ESG scores and portfolios, which we believe has the advantages of being more principled and likely to be more trusted by stakeholders.

2. Framework

The purpose of this paper is to present a framework to develop an agile ESG approach to portfolio construction, tilted towards increasing allocations to companies with few mentions of ESG issues in news articles over a long time horizon. Our framework consists of three steps: first, we use news data sources (specifically, New York Times data, obtained via the NYT Developer API [10]), which provides historical author-curated data tagging that we use as proxies for several key ESG categories. Second, to address the fact that these tags may not be consistently provided, and can vary in their meaning, we introduce a weak supervision approach to create a useful ESG classifier. Specifically, we build a machine learning model using these tags in order to ensure that the ESG scoring is consistent when applied over several decades. This is achieved by isolating the time periods during which tagging was most consistent for tags related to ESG for use as weak labels. Finally, we aggregate the results of our model applied at the article level to generate ESG signals and apply a modified Black-Litterman model [11] to construct ESG portfolios. These portfolios uniquely take into account both ESG and investment risk criteria to create optimized portfolios. We end the paper with an empirical study of the performance of ESG portfolios constructed through a thorough back-testing setting.

2.1. Data Collection & Labelling

Our data set covers the following ESG topics:

- Governance: Business Ethics, Anti-Competitive Practices, Corruption & Instability
- Social: Health & Demographic Risk, Supply Chain Labour Standards or Labour Management, Privacy & Data Security
- Environmental: Climate Change or Carbon Emissions, Product Quality & Safety, Toxic Emissions & Waste

We utilize 2 key sources for ESG data in this study: New York Times news articles from 1998 to mid 2019 (overlapping with the availability of Compustat data described below), as well as equity price data for companies in the S&P 500 index, aligned to the same time period.

In order to generate an appropriate corpus of News Articles relevant to ESG topics, we leverage New York Times Developer API [10] to collect News Articles from the business section of the New York times, published from 1998 to mid 2019. This corpus contains hundreds of thousands of articles related to a variety of news topics, some of which are relevant for ESG. We utilize author-provided tags as weak supervised labels, using relevant tags as proxies for ESG categories listed above. To ensure class imbalance is not a concern during training, we randomly sample the corpus keeping all articles with tags relevant to an ESG category, combined with a sample of articles without relevant tags. In addition, we use company names (also provided as tags via the New York Times API) and, thus, did not use a separate named entity recognition model to identify companies.

Our equity price data was obtained from Compustat, for which we use data starting in 1998 and ending with mid-2019 for our portfolio analysis since S&P index data is no longer available past this date. We also apply several data quality improvements to our dataset to ensure consistency over time, such as imputing any missing trading dates and accounting for companies that trade under multiple

Table 1: BERT Model Hyperparameter Choices

Parameter	Definition	Value
Stage 1 Learning Rate	Learning rate during stage 1 of training (classification layer)	10^{-3}
Stage 2 Learning Rate	Learning rate during stage 2 of training (classification and encoder layer)	10^{-4}
Early Stopping Eval Limit	Consecutive evals with no improvement before stopping	5
Max Sequence Length	Max sequence length for BERT	256
Batch Size	Batch Size used in training	16
Evaluate Every n Samples	Run evaluation every n samples	400

tickers. Outside of these changes we take the price data as is, and use it solely for portfolio backtesting described in Section 2.3.

Finally, to align news article data against Compustat data we use a collection of text similarity algorithms to generate a dictionary between Company names available in the NYT and the company names associated with Compustat data, and then evaluate and refine this dictionary, curating a final map between NYT company names and S&P 500 tickers.

2.2. Model Training

We train our model as a standard BERT [12] classifier, using a weak supervision approach by leveraging author-provided tags as proxies for ESG categories of interest. We use the pre-trained model from HuggingFace [13], and apply two-stage fine-tuning to train our model, whereby we first freeze the encoder layers and only train the classification layers of our model, and then train the full model briefly after our early-stopping criteria are triggered. We train a single model to classify all of our ESG categories simultaneously in order to combine the labeled data from individual categories and create good general ESG text embeddings, and select our hyperparameters based on the weighted average precision scores across our classes. Our final hyperparemeter choices are presented in Table 1.

We train one network to model for all of the ESG categories described in Section 3.1. We do this in order to utilize all of our labeled data to fine tune a single representation, thus boosting the total number of labels available for fine tuning; another benefit of this approach is to create better general representations of news data for ESG, which can then be applied towards some potential improvements to the model described in Section 4.

2.3. Portfolio Construction

A key contribution we make is to demonstrate a working example of how the results of a machine-learning NLP model may be used to continuously incorporate views on ESG performance into portfolio construction decisions, even in the presence of noisy labels for ESG categories. We believe our proposed approach addresses a number of key issues inherent in incorporating unstructured data into portfolio construction decisions, demonstrates the technical feasibility of such a combination, and showcases interesting results in the process.

There is no single clear approach through which the results of an NLP model, even when aggregated into an ESG score at a company level, can be used to make investment decisions. We accomplish this by leveraging a widely used Black-Litterman Model [11] optimization approach to incorporate ESG views, which is a very natural approach for this task. The Black-Litterman model allows portfolio managers to incorporate a Bayesian vector of views, in combination with an imputed vector of views expressed by the market, as priors. These priors, used as inputs into the classical Markowitz portfolio optimization algorithm[14], have the effect of pivoting away from investments the portfolio manager views more negatively relative to the market. Converting scored ESG data into a views vector is therefore

particularly pivotal to our approach. There are several key decision points to be made in order to convert news data ESG scores into signals. We elaborate on each of these below:

- Initially, our machine learning classifier outputs a number between 0 and 1 for each article, which roughly corresponds to the probability of the article belonging to a particular ESG category. As seen in the precision-recall estimates we present in Section 3.1, we see significant differences across categories in terms of the number of false positives one has to tolerate in order to meet a particular recall cut-off. To add to this complexity, our model performance is often under-estimated by human error and labeling inconsistency, where articles labeled by our model but not by NYT are in fact legitimate members of the category. These considerations lead us to pick a specific probability cut-off for each category for treating an article as indication of ESG risk.
- After these signals are generated, we need to account for natural variance in reporting, both over time as the news cycle moves towards big newsworthy events and away from reporting on corporate issues, as well as companies which simply get more press on average. We accomplish this by first aggregating signals over the last quarter, clipping the highest signals at the 95^{th} percentile (making the overall distribution less skewed), and scaling the signals so that the 95^{th}-percentile signal receives the value of 1.
- Lastly, the normalized weights get turned into a views vector Q. As per the standard Black-Litterman approach [11], we add our views vector to the expected returns vector in order to tilt the portfolio away from companies linked to a significant amount of ESG press.

After producing our ESG views vector as described above, we follow the standard Black-Litterman portfolio optimization approach. As in the seminal work [11]. We compute the posterior returns estimate as over n assets as:

$$E(R) = [(\tau\Sigma)^{-1} + \Omega^{-1}]^{-1}[(\tau\Sigma)^{-1}\Pi + \Sigma^{-1}Q]$$

Where $E(R)$ represents the posterior returns expectation, Σ represents the prior covariance matrix and τ is used as a scalar tuning constant for adding additional uncertainty, Ω as the $n \times n$ uncertainty matrix of views, Π gives the prior vector of expected returns, and Q gives the prior vector of asset manager views. As we discussed above, estimating Q is pivotal to our approach. We further discuss the impact Q has on the make-up of our ESG portfolio in Section 3.2.

In addition, we compute the posterior covariance matrix as:

$$\hat{\Sigma} = \Sigma + [(\tau\Sigma)^{-1} + \Omega^{-1}]^{-1},$$

Where $\hat{\Sigma}$ represents the posterior covariance matrix, Σ represents the prior covariance matrix, τ is used as as a scaling constant adding additional uncertainty, and Ω represents the portfolio manager's estimate on the uncertainty of their views. We did not compute an ESG-specific estimate Ω, instead using market covariance as the uncertainty matrix estimate. Some ways in which Ω could potentially be estimated when using inputs of a predictive model would be by using model uncertainties to aggregate the total uncertainty of NLP models predictions by using estimates of false positive and false negative rates, as well as the number of predictions made. One could also build on this approach by estimating the statistical distributions and estimating the covariance matrix of model errors.

We use this strategy with the constraint of taking long-only positions to reflect the needs of the long-only ESG-focused investor, and with the optimization objective of maximizing the Sharpe Ratio. Some periods have no feasible solution to the optimization problem with the long-only constraints, during which times we use market weights for both the ESG and the standard Black-Litterman portfolios; the results are summarized in Section 3.2.

Table 2: NLP Model Performance

E/S/G	ESG Sub-Category	ROC AUC	PR AUC
Social	Product Liability	0.92	0.65
Social	Human Capital	0.90	0.62
Social	Privacy	0.91	0.59
Social	Data Security	0.93	0.63
Governance	Anti-Competitive Practices	0.93	0.75
Environmental	Pollution & Waste	0.95	0.70
Environmental	Climate Change	0.96	0.73

Table 3: Model Thresholds

E/S/G	ESG Sub-Category	Threshold Precision	Threshold Recall	Sampled Precision at Cut-Off
Social	Product Liability	19%	69%	50%
Social	Human Capital	39%	46%	58%
Social	Privacy	40%	46%	57%
Social	Data Security	49%	31%	56%
Governance	Anti-Competitive Practices	29%	80%	54%
Environmental	Pollution & Waste	17%	62%	58%
Environmental	Climate Change	12%	50%	40%

3. Results

3.1. ESG News Data Classifier and Weak Supervision

The choices of hyperparameters detailed in Table 1 result in model performance metrics computed by using author-curated tags directly, summarized in Table 2. These results are very promising, and based on our analysis of false-positive and false negatives are under-stated (meaning our models are able to generalize well despite the prevalence of missing tags in the historical NYT data), as detailed in Table 3. Given that the fully labeled hold-out set is based on author-provided tasks, we find that in general the true performance of the model for classifying ESG issues is understated by this direct computation. We discuss our estimates of effective performance of the NLP model below, as well as the implications towards using the model within a portfolio optimization framework. We note that good generalization of NLP models despite limitations and inconsistencies in labeled data is expected, and is consistent with recent research on weak supervision in NLP, for example [15].

Based on these results, we are able to choose cut-offs (detailed in Table 3) that provide precision-recall trade-offs that are sufficiently high for us to make a determination. We note that the performance of our models at our chosen thresholds is empirically better than what we estimate using NYT labeled data only, due to limitations and a level of inconsistency in historical labeling, as Table 3 shows. Our choice of cut-off was made such that the approximate precision of each of our NLP models was $\sim 50\%$, and aimed to maximize recall under this constraint. For example, based on the precision-recall curve we estimate for the Product Liability ESG category as can be seen in Figure 1, our chosen cut-off allows us to capture $\sim 80\%$ of the total number of mentions of Product and Liability as an ESG issue in relation to S&P500 companies, and we estimate the effective precision as this threshold to be $\sim 50\%$. Other classes similarly yield results that are sufficiently precise to provide accurate estimates of ESG risk exposure when averaging over a large number of news articles. The intuitive results and favorable performance of the Black-Litterman-ESG portfolio, as described in Section 3.2, yield credence to satisfactory performance of our NLP models.

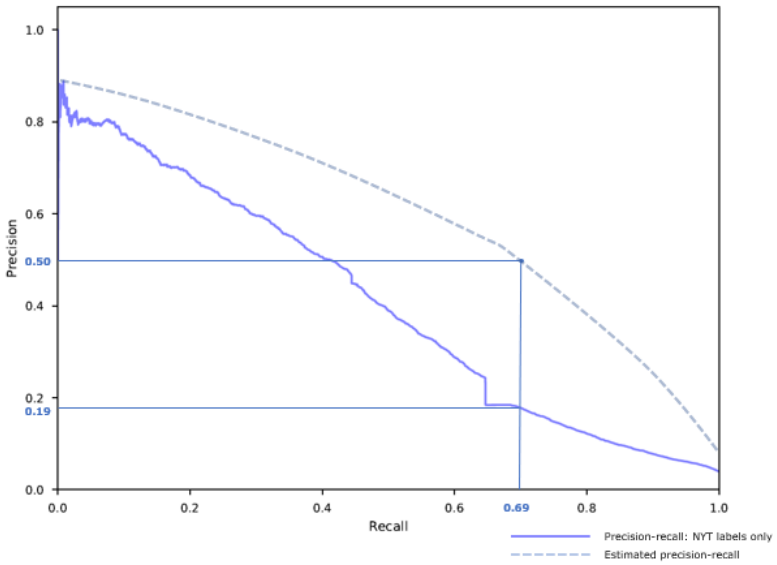

Figure 1: Threshold Choice for the Product Liability ESG Category

3.2. Portfolio Comparison

Although ESG criteria are generally not performance driven, we compare the historical performance of three ESG portfolios. The first is the market-weighted portfolio provided for reference, the second a Black-Litterman Benchmark portfolio with no additional priors (i.e. a zero-vector used for the portfolio manager's views), and the third is the Black-Litterman ESG portfolio utilizes the article scores generated by our NLP model to generate a views vector, where a high prevalence of ESG-related news is taken as a negative signal, as described in Section 2.3 above. Figure 2 shows a comparison of the three portfolios.

Note that the Black-Litterman ESG portfolio performs in-line with the benchmark Black-Litterman portfolio, with both having higher returns than the market portfolio. Table 4 includes some additional metrics, which also shows that the performance is of the ESG portfolio is in line with the benchmark Black-Litterman portfolio across a range of metrics commonly used in portfolio management. The information ratio is computed using the market portfolio as a benchmark, while the Sharpe Ratio and standard deviation of excess returns are computed using US Treasury spot interest rates. We purposefully avoid running a high number of experiments in order to avoid the pitfalls of P-Hacking in quantitative finance [16], and therefore we do not make the claim that an investment strategy based on ESG investing would produce reliable excess returns in the future. One might hypothesize that this level of performance is a signal of an ESG momentum factor, but a deeper analysis would be required to conclusively determine that. Nevertheless, we believe it is interesting that such a portfolios could have performed in line with an unconstrained portfolio in the past, perhaps indicating possible advantages of implementing ESG portfolio "tilt" strategy, rather than treating ESG considerations as exclusionary constraints.

Figure 3 are the industries that were under- or over-indexed in the BL-ESG portfolio throughout our testing period. Note much of this is intuitive: the Finance and Insurance industries are under-represented during market downturns, especially during the Great Financial Crisis, due to the increases in negative publicity and therefore in mentions of ESG issues. Similarly, the Information industry becomes more under-represented in recent years, as privacy issues have become much more prevalent and drawn additional public scrutiny. Also note that during periods where the market becomes highly correlated, there is no under- or -over indexing as all strategies track the market portfolio during these periods. We also note that the two Black-Litterman portfolios are significantly different, with an average overlap of only 19.85% in their holdings over the periods where Black-Litterman optimization is performed.

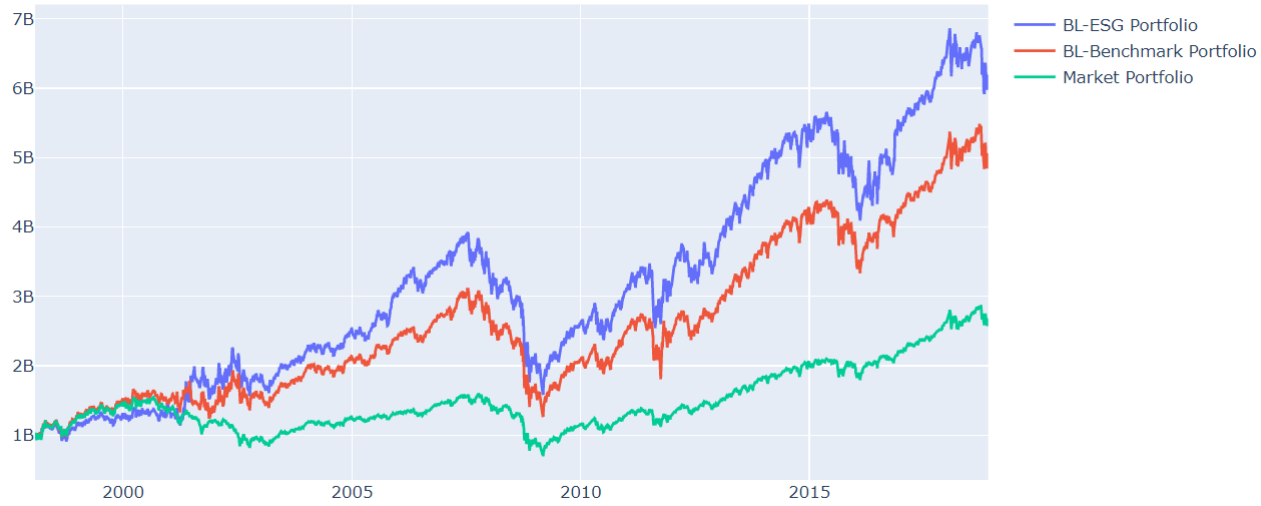

Figure 2: Comparison of Market Portfolio, ESG-BL Portfolio and Benchmark BL Portfolio

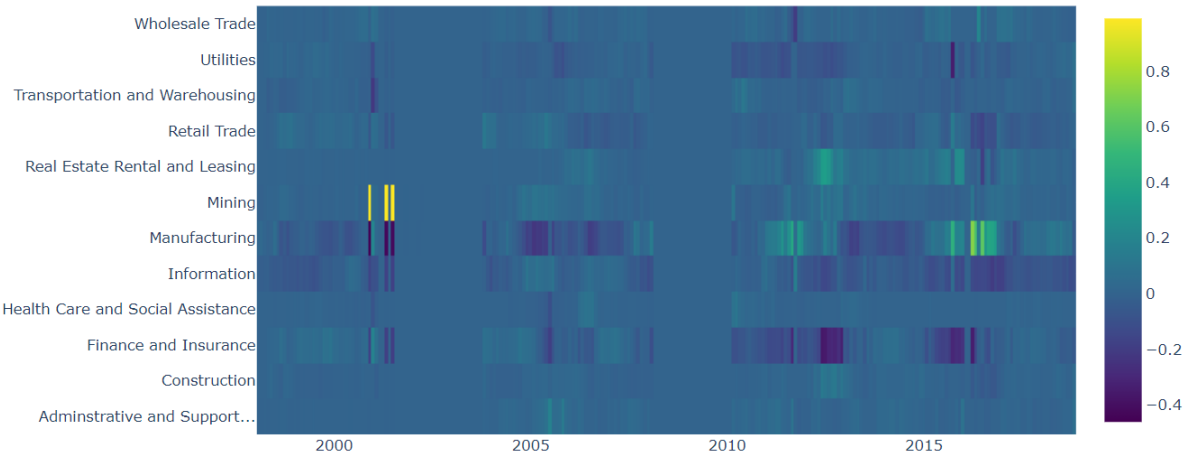

Figure 3: Industry under- and over-Indexing for the ESG Portfolio

Table 4: Portfolio Metrics

Portfolio	**Information Ratio**	**Sharpe Ratio**	**Annualized Returns**	**Volatility of Excess Returns**
BL-ESG Portfolio	**0.0147**	**0.400**	**9.14%**	24.59%
BL-Benchmark Portfolio	0.0142	0.368	8.08%	**23.18%**
Market Portfolio	0.0	0.237	4.77%	21.63%

4. Conclusion and Future Work

In conclusion, we developed and demonstrated an approach that can be used to extract ESG signals from unstructured text data using deep learning models for NLP, then subsequently incorporate these signals into portfolio construction decisions. Such an approach can generate portfolios that are tilted towards companies with fewer signals of ESG exposure, but is otherwise competitive with a benchmark portfolio across a variety of metrics.

There are a number of opportunities to build on our work described above, and we detail some of the approaches that we deem most promising for improving on our work.

As the state of the art in natural language processing is evolving rapidly, one area of opportunity is incorporating newer NLP models. There are two primary areas of research pushing the state-of-the-art over the large Transformer-based models (BERT [12], ALBERT [17]) we utilized for our NLP work. The first area of research is in pushing the performance on common NLP benchmark tasks and generalizability of NLP models, primarily (but not exclusively [18]) through scaling up transformer architectures to be even larger, e.g. XLNet [19], [20] or GPT-3 [21]. Another key area of research, which is perhaps even more useful in the domain of ESG, is focused on achieving similar results but dramatically scaling down the size of the model, such as DistilBERT [22], TinyBERT [23]. This is very promising due to the large volumes of data related to ESG, where lowering inference costs and reducing / eliminating the need for multiple models (e.g. TF-IDF based pre-filtering) has the potential to save significant engineering and computational costs. One can also incorporate additional detail into estimating the posterior covariance matrix $\hat{\Sigma}$ by incorporating the statistical distribution of NLP model errors to reflect the inherent uncertainty in the views vector.

Another key area of improvement is in collecting additional data, and potentially in combining data from various ESG data sources. Relevant sources of text data for ESG include social media data, news data, regulatory filings, and company ESG disclosures. Building generalized NLP models that are robust to dealing with multiple data sources also has the potential for significantly reducing engineering costs for building ESG scoring systems, and also for creating better robust general representations of ESG text data by pooling labeled data across sources. These better representations can then be used beyond supervised classification tasks and expanded to additional tasks such as detection of ESG events or emerging ESG risks.

References

[1] Alex Bryan et al. *Passive Sustainable Funds: The Global Landscape 2020*. Tech. rep. Sept. 2020.

[2] Samuel M. Hartzmark and Abigail B. Sussman. "Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows". In: *The Journal of Finance* 74.6 (2019), pp. 2789–2837. DOI: 10.1111/jofi.12841. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12841. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12841.

[3] J.P. Morgan. *Why COVID-19 Could Prove to Be a Major Turning Point for ESG Investing*. July 2020. URL: https://www.jpmorgan.com/insights/research/covid-19-esg-investing (visited on 10/12/2020).

[4] Divya Balji. "Sustainable investing surges in Canada amid pandemic, protests". In: *BNN Bloomberg* (June 2020). URL: https://www.bnnbloomberg.ca/sustainable-investing-surges-in-canada-amid-pandemic-protests-1.1453425 (visited on 10/12/2020).

[5] Esther Whieldon, Robert Clark, and Michael Copley. "ESG funds outperform S&P 500 amid COVID-19, helped by tech stock boom". In: *S&P Global Market Intelligence* (Aug. 2020). URL: https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/esg-funds-outperform-s-p-500-amid-covid-19-helped-by-tech-stock-boom-59850808 (visited on 10/12/2020).

[6] Briegel Leitao. "How ESG ETFs Have Performed in the Sell-Off". In: *Morningstart ETF Research & Insights* (Apr. 2020). URL: https://www.morningstar.co.uk/uk/news/201154/how-esg-etfs-have-performed-in-the-sell-off.aspx (visited on 10/12/2020).

[7] Sakis Kotsantonis and George Serafeim. "Four Things No One Will Tell You About ESG Data". In: *Journal of Applied Corporate Finance* 31.2 (2019), pp. 50–58. DOI: 10.1111/jacf.12346. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jacf.12346. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jacf.12346.

[8] Dane Christensen. "Why is Corporate Virtue in the Eye of The Beholder? The Case of ESG Ratings". In: 2019.

[9] Florian Berg, Julian F. Kolbel, and R. Rigobón. "Aggregate Confusion: The Divergence of ESG Ratings". In: 2020.

[10] New York Times. *Article Search API*. 2020. URL: https://developer.nytimes.com/docs/articlesearch-product/1/overview.

[11] Fischer Black and Robert B Litterman. "Asset Allocation". In: *The Journal of Fixed Income* 1.2 (1991), pp. 7–18. ISSN: 1059-8596. DOI: 10.3905/jfi.1991.408013. eprint: https://jfi.pm-research.com/content/1/2/7.full.pdf. URL: https://jfi.pm-research.com/content/1/2/7.

[12] 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].

[13] 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].

[14] Harry Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91. URL: https://www.jstor.org/stable/2975974.

[15] Alexander J. Ratner, B. Hancock, and C. Ré. "The Role of Massively Multi-Task and Weak Supervision in Software 2.0". In: *CIDR*. 2019.

[16] de Prado, Marcos Lopez. "Clustered Feature Importance (Presentation Slides)". In: (2020). SSRN: https://ssrn.com/abstract=3517595.

[17] Zhenzhong Lan et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[18] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

[19] Zhilin Yang et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 5753–5763. URL: http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf.

[20] Alexis Conneau et al. *Unsupervised Cross-lingual Representation Learning at Scale*. 2020. arXiv: 1911.02116 [cs.CL].

[21] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[22] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *NeurIPS EMC^{2}Workshop*. 2019.

[23] Xiaoqi Jiao et al. *TinyBERT: Distilling BERT for Natural Language Understanding*. 2020. URL: https://openreview.net/forum?id=rJx0Q6EFPB.