Modelling ESG with Twitter Data

Knowsis: From Text to Signal

Armando Marozzi and Filippo Lanza

*Armando Marozzi is a Postdoctoral Researcher at the European Institute, London School of Economics (a.marozzi@lse.ac.uk) and Head of Quantitative Research at Knowsis (armando@knows.is).
Filippo Lanza (CFA) is CEO and CIO of Numen Capital Ltd and Director of Knowsis (filippo.lanza@numencapital.com). He has also earned the CFA Institute Certificate in ESG Investing.

Abstract

The integration of Environmental, Social and Governance (ESG) considerations into investment strategies, business decisions, legislative and regulatory initiatives have grown exponentially over the past few years. As a result, many financial and non-financial organizations, as well as governments and regulators, raced to rate companies on the basis of their ESG performance. However, the current scoring systems suffer from two substantive limitations: on the one hand, scores are usually *backward-looking* and, on the other hand, they are only available at *low-frequency*. We aim to overcome these drawbacks by providing a real-time Twitter-based score for every letter of E-S-G. In detail, firstly, we develop an accurate classification model, relying on state-of-the-art neural networks, that identifies tweets related to E-S-G topics. Secondly, based on a collection of more than a thousand professional documents related to ESG taxonomy, we create a unique ESG dictionary containing polarized unigrams and ngrams for environmental, social and corporate governance matters. Thirdly, we construct a new scoring algorithm that closely resembles the natural language and is able to capture social media nuances. We highlight the following results: first, the model proves to *accurately and timely track ESG events* for more than 4000 stocks. Second, our ESG metrics often *anticipates* market movements. Third, further empirical applications show the *forecasting power* of our ESG scores in predicting stock returns *throughout the pandemic period*.

Keywords: NLP, ESG, Machine Learning, Forecasting, Finance

1. Introduction

Environmental, Social and Governance (ESG) considerations have gained increasing attention over the past few years; so much so that, on August 19 2019, America's Business Roundtable announced the release of a new Statement on the Purpose of a Corporation (SPC) signed by 181 CEOs. The SPC formally stated that companies should no longer advance only the interests of shareholders, but must also invest in their employees, protect the environment and deal ethically with suppliers. While each previous version of the document issued since 1978 had endorsed principles of "shareholder primacy", meaning that corporations exist principally to serve shareholders, the new announcement marked the arrival of a modern standard for corporate responsibility.

This new set of principles - also fostered by the stewardship of global partnerships such as the United Nations Environment Programme Finance Initiative (UNEPFI) and the Principles for Responsible Investment (PRI) - has inevitably led companies to integrate ESG topics into corporate strategies, business models and codes of conduct. In parallel, market participants have progressively embedded ESG factors into stock and bond valuations[1]. Financial and non-financial organizations, as well as governments and regulators, have consequently proposed various taxonomies and principles to facilitate the implementation of an appropriate ESG framework for companies and investors.

The emergence of several scoring and rating solutions, currently dominated by MSCI and Sustainalytics, has in fact allowed investors to analyze and assess the ESG performance and compliance of their investee companies and consistently integrate those analyses into portfolio solutions. However, despite being a powerful guidance for investors, existing score systems suffer from several practical limitations. Firstly, they rely on companies voluntary and unaudited disclosures around ESG themes. Secondly, they are heavily dependent on the qualitative assessment of their analysts and their judgment calls. Thirdly, and unsurprisingly, commonly used ESG ratings are not easily comparable and are not even consistent among themselves, raising several issues for the investment community (see, for example, Chatterji et al., 2016; Berg

[1] Given the complexity of the ESG factors, ranging from renewable energy to business ethics, a growing literature has investigated the impact of ESG practices on markets, company performance, productivity and industry trends.

et al., 2020; Brandon et al., 2021). Furthermore, from an investor point of view, there are two other substantial drawbacks: on the one hand, scores are usually *backward-looking*, that is, they only offer an assessment of how virtuous ESG-companies have been. Whenever forward-looking valuations are included, they are often based on company sustainability reports (CSR) that are produced by the companies themselves. Such a methodology, therefore, doesn't appear too robust as companies are incentivized to present a more positive picture of their ESG performance and future goals. This incentive into green-washing and ESG-washing has grown more relevant as ESG dedicated funds raised more money, making an ESG better rating a boost for company's share price. On the other hand, the most widely-used ESG ratings are only updated *infrequently* (typically, quarterly or annually). This means that investors receive information with a significant time lag, thereby missing the market timeliness of ESG controversies.

We aim to overcome these drawbacks by providing a real-time Twitter-based score calibrated for every letter of E-S-G. Although the informational richness of social media data for financial topics is well established (Sun et al., 2016; Gu and Kurov, 2020), we restate four main advantages of our Twitter dataset: first, we receive tweets in *real-time directly* from Twitter via a fat API pipeline. Therefore, the timeliness of our source of information has no delay. Moreover, this allows us to provide *intraday* historical time series for more than 4000 stocks. Second, tweets are often *forward-looking* in content since they capture the process of expectation formation. Third, unlike rankings computed on discretionary rules and/or companies' own reports, Twitter is somehow "ruthlessly democratic" in so far as it gives everyone who feels entitled the possibility to express his own view on a topic of interest and an instantaneous and transparent voting system around it. More in detail, increasingly consumers, investors, journalists, senior managers and company's founders are becoming vocal and active on social media, thereby making their opinions heard and amplified very quickly. In the case of sustainable topics, for example, ESG perception around a company is extremely relevant since it dictates not only spending and consumption behaviors on the company's products and services but also the investment decision of a retail cohort of investors where consumption and investment decision processes frequently overlap. This creates a self-reinforcing cycle that amplifies news' impact. Fourth, from an investment point of view, it is noteworthy that most ESG controversies and especially severe ESG events, such as corporate scandals and frauds, have originated typically from a whistleblower either inside the company or outside (e.g., an investigative journal or consumer group). Historically, such scandals (e.g., Wirecard, Parmalat, Volkswagen, etc.) were detected or flagged

neither by credit rating agencies nor by sell side analysts. Nowadays, instead, the voicing and flagging of ESG concerns increasingly occur on social media and, predominantly, on Twitter which has become the dominant "virtual speaker's corner" for the globe. Hence, Twitter is well-suited to serve as a real-time anomaly detector.

To model our Twitter dataset, we proceed in three steps. Firstly, we develop an accurate classification model that identifies tweets related to E-S-G topics. Secondly, based on a in-house collection of more than a thousand professional documents related to ESG taxonomy, we create a unique ESG dictionary containing polarized unigrams (single words) and ngrams (sequence of words) for environmental, social and corporate governance matters. Thirdly, we construct a new scoring algorithm that closely resembles the natural language and is able to capture social media nuances.

There are three main results to highlight: first, the model proves to *track accurately and timely ESG events* for more than 4000 stocks. Second, our metrics often moves *ahead of the market*. Third, further empirical applications show the *forecasting power* of our ESG scores in predicting stock returns *throughout the pandemic period*, that is, a rather challenging period for professional forecasters where many long-lived models have failed to remain accurate.

The rest of the paper is organized as follows: Section 2 outlines the ESG model, Section 3 documents the empirical applications, Section 4 discusses the integration of ESG into investment strategies and Section 5 concludes the paper[2].

2. Delving into the Model

In this section, we present our E-S-G model. In particular, Section 2.1 sheds light on the classification model, Section 2.2 presents our letter-specific (E-S-G) dictionary, Section 2.3 introduces our rule-based scoring algorithm and Section 2.4 compares the performance of the model with cutting edge alternatives.

[2] For the sake of brevity, we don't provide in the paper a formal treatment of our models. If interested, please contact us for more information.

2.1. Classifying E-S-G Tweets

Social media, like a jester in the court telling truth to the king, can be very noisy. In fact, we receive, on average, 20 tweets per second, more than one million tweets per day and roughly 30 million tweets per month. We assign every tweet to an asset[1] on the basis of a list of keywords specific to each asset[2]. Then, we need to filter the content properly. In particular, we aim to identify only tweets whose content is related to ESG topics. To be as granular as possible, we classify tweets according to their environmental (E), social (S), governance (G) or other (O) relevance using a multiclass Tranformer model.

Transformers are the state-of-the-art of natural language processing (NLP) models and represent the backbone of Artificial Intelligence (AI) models developed by Google AI, Facebook AI, and Open AI (i.e. GPT-2, GPT-3, XLNet, BERT, RoBERTa). A Transformer is a deep learning model that relies on self-attention mechanisms (Vaswani et al., 2017) and is designed to manage sequential input data, such as natural language. Unlike recurrent neural networks (RNNs), transformers do not need to process the data in sequential order. Rather, the attention mechanism provides context for any position in the input sequence. For instance, if the input data is a natural language sentence, the transformer does not necessarily have to process the beginning of the sentence before the end. Instead, it identifies the context that confers meaning to each word in the sentence. This feature allows for higher accuracy, more parallelization and lower training times than standard RNNs.

To customize the transformer model to our ESG framework, we constructed a training set with around 1 million and 300 thousand tweets, of which 800 thousands were classified as "Other", 200 thousands as "Environmental", 150 thousands as "Social" and 150 thousands as "Governance"[3]. The training set was manually classified by ESG experts and "buy-side" professionals who followed an accurate ESG mapping we developed internally. The multiclass transformer model was then trained accordingly.

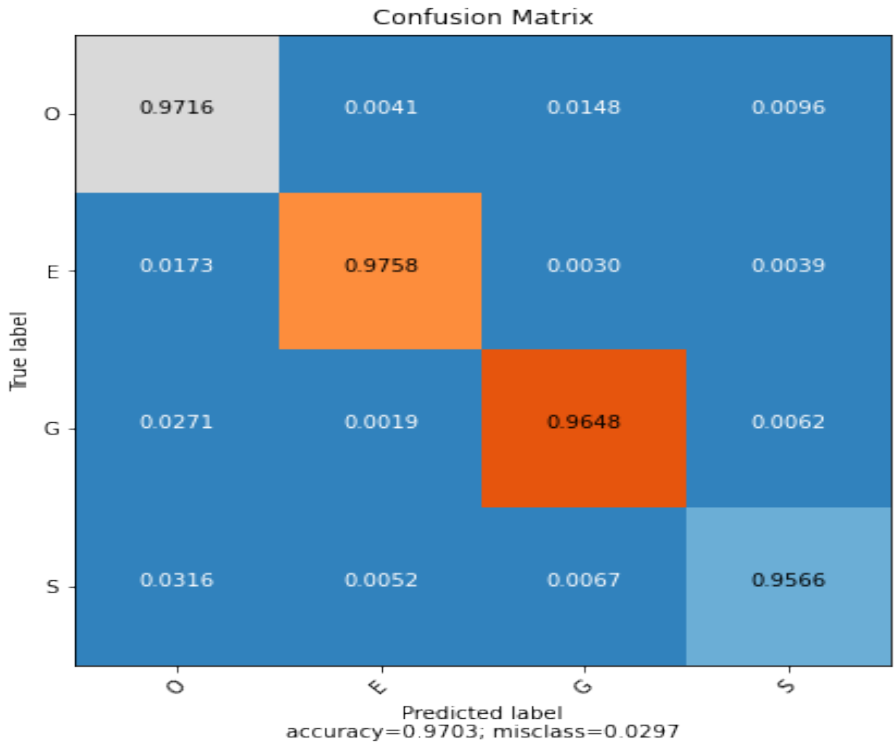
[1] In Knowsis, an "asset" can be any publicly owned company traded on a relevant exchange, any commodity, any stock index, any private company etc. We have the flexibility to add as many assets as we are interested in.

[2] For example, the keywords for Tesla are: "tesla","teslamotors", "elonmusk", "teslainc", "tsla", "teslamotorsinc", "elon musk", "deepak ahuja", "jeffrey straubel", "veronica wu".

[3] In building our training set, we attempted to replicate the structure of tweets where the majority of tweets are not about ESG and, among ESG topics, environmental tweets are more frequent than social and governance ones. The training set is based on tweets sampled across assets (companies) and time (from 2012 onward) to get a robust and representative training set.

We now analyze the performance of the classification model. Figure 2 illustrates the confusion matrix. The transformer model appears to have high-level performance. In fact, focusing on the diagonal of the confusion matrix, we can observe high percentages for every class, that is, the model is able to predict the true label, on average, more than 97% of the times.

Figure 1: Confusion Matrix



Note: The figure documents the outcome of the confusion matrix.

We can derive additional statistics from the confusion matrix to help us shed further light on the performance of the classification model. Table 1 shows that the transformer model proves to be extremely accurate with negligible deviations among classes. We instead observe more prominent variations among classes when we look at precision, namely the proportion of predicted positives that are actually positive. In particular, "Social" and "Governance" tweets tend to be classified less precisely (respectively, 88% and 93%) than "Environmental" and "Other" ones (respectively, 98% and 99%). This may be due to the nature of tweets' content: while environmental and non-ESG relevant tweets have neater boundaries, the perimeter of

social and governance topics is much more blurred[4]. That said, Table 1 overall documents the solidity of our classifier.

Table 1: Classification Transformer: Performance

	Environmental	Social	Governance	Other
Accuracy	0.976	0.956	0.965	0.972
F-score	0.978	0.919	0.950	0.979
Precision	0.981	0.885	0.935	0.986
Recall	0.977	0.941	0.951	0.985

Note: The table illustrates the performance of the classification model showing precision, accuracy and F-score for each letter.

2.2. A Letter-Specific Dictionary

There is a growing literature that applies NLP models to ESG topics. For example, Shahi et al. (2014) apply a variety of classifiers including a feed-forward neural network to company sustainability reports (CSR) in order to predict the sustainability score. Similarly, Raman et al. (2020) analyze the transcripts of corporate earning calls and find that 15% of the discussions during earning calls pertain to ESG, implying that ESG factors are integral to business strategy.

Data sources, such as news and social media have also been used in the ESG domain. Hisano et al. (2020) use adverse media coverage to predict companies that are likely to be blacklisted in sustainable investment practices. Nematzadeh et al. (2019) combine event study techniques with sentiment classification in order to identify Twitter controversies around a given company and associate the controversies with the stock performance of the company. Moreover, Ardia et al. (2020) construct a Media Climate Change Concerns index using news about climate change published by major U.S to test whether green firms outperform brown firms when concerns about climate change increase unexpectedly. On the same vein, Schmidt (2019) investigates the effect of ESG related news sentiment on the stock market performance of the Dow Jones Industrial Average (DJIA) constituents.

[4] For instance, if we focus on "Social", we find that out of the false positives for "S" (i.e tweets that have been wrongly classified as "Social") approximately 80% were "Other". The cost of falsely predicting "Other" as "Social" is however mitigated by two factors: tweets are asset-related and the dictionary is ESG-specific.

While these papers provide interesting insights on the ESG universe, the main limitation consists in the reliance on a generic vocabulary to quantify ESG-specific textual information. ESG language is in fact challenging due to its multi-faceted nature and cannot be accurately approximated using existing polarized dictionaries. For instance, if we take the Valence Aware Dictionary and sEntiment Reasoner (VADER, Hutto and Gilbert, 2014) - by far the most popular and widely used dictionary among researchers - to score an ESG sentence such as "I would like to reduce carbon emissions", we would be calculating a misleading score. In fact, VADER would only identify "reduce" as a negative word since it only contains generic unigrams.

To overcome these issues, we create - to the best of our knowledge - the first ESG-sensitive dictionary currently on the market. The dictionary combines generic unigrams and ngrams, E-S-G specific words and sequence of words as well as emoticons, totalling around 30,000 features - roughly 10,000 features per letter[5]. In particular, the dictionary contains 15,610 E-S-G specific ngrams that were extracted from Twitter and cross-validated by "buy-side" professionals and ESG experts[6]. Table 2 shows a sample of ngrams and unigrams for each letter of E-S-G[7].

[5] From the gold-standard valence aware sentiment with 7,502 uni-grams of casual sentiment expressions (Hutto and Gilbert, 2014), a total of 1,898 expressions have been modified (1,177) and removed (721) to fit in the ESG context.

[6] Out of 15,610 features, 8,829 are unique ESG-specific ngrams and are composed of (i) 5,376 unigrams; (ii) 2,520 bigrams, (iii) 790 tri-grams; and (iv) 143 n-grams (4+). The selection was based on our ESG Mapping.

[7] Alongside the E-S-G dictionary, we also collected a list of so-called valence shifters (namely, adversative conjunctions, negations, amplifiers and deamplifiers) whose use will be made clearer in Section 2.3.

Table 2: Sample of N-grams and Unigrams for Each Letter of E-S-G

Environmental	**Social**	**Governance**
Positive	Positive	Positive
– address climate challenge*	– adapt to employee need*	– corporate social responsibility
– reduce carbon emission*	– above minimum wage	
– address climate change	– address social issue	– environmental policy
– absence of pollution*	– affordable housing	– equal opportunity
– absence of emission*	– employee benefit*	– democratization
– zero emission*	– alleviate poverty	– business ethics
– carbon offset	– drinkable water	– board diversity
– capture co2	– unionization	– gender equity
– carbon free	– inclusion	– transparency
– recycling	– diversity	– donation*
		– integrity
Negative	Negative	Negative
– environmental devastation	– crime against humanity	– breach of confidentiality
– climate breakdown	– displace local econom*	– antivoting right
– chemical weapon*	– underrepresentation	– civil allegation*
– carbon footprint*	– domestic violence	– tax avoidance
– carbon emission*	– dodgy reputation	– tax evasion
– carbon intensity	– forced labour	– corruption
– glacial melting	– enslavement	– antitrust
– greenwashing	– dictatorship	– zero tax
– ecocide	– unfair wage	– briber*
– ghg	– abuser*	– flirt

Note: The table shows a sample of n-grams and unigrams for each letter of E-S-G. * indicates that the dictionary contains both the singular and the plural of a specific unigram/n-gram.

Therefore, unlike existing attempts that throw away information due to a generic vocabulary, our dictionary enables us to fully capture all the information on Twitter since: (i) the dictionary is calibrated to span every dimension of environmental, social and governance matters, (ii) it is integrated with generic unigrams and ngrams and (iii) it includes Twitter-specific expressions such as emoticons, peculiar punctuation and acronyms.

2.3. Rule-Based Scoring Algorithm

Once we are able to distinguish ESG relevant tweets from non-ESG ones, the last step is to apply the dictionary we constructed in Section 2.2 in a meaningful way. Building on Hutto and Gilbert (2014), Rinker (2019) and Marozzi (2021), we develop a rule-based algorithm that approximates

the natural language as closely as possible and is thereby able to accurately measure the tone of a tweet[8]. Intuitively, the algorithm embeds the dictionary into generalizable heuristics. If a tweet contains a unigram/ngram in the dictionary, the score can be modified whether before of after the unigram/ngram: i) there is an amplifier/deamplifier (e.g. "significantly"/"marginally") that increases/decreases the tone of a tweet; ii) there is an adversative conjunction (i.e. "but") that causes a shift in the sentiment score; iii) there is a negation (e.g. "not") that flips the score; iv) there is either peculiar punctuation (e.g., a sequence of exclamation or question marks) or capital letters (e.g., "ALL-CAPS") that emphasize the sentiment. A more formal explanation of the rule-based score is available upon request.

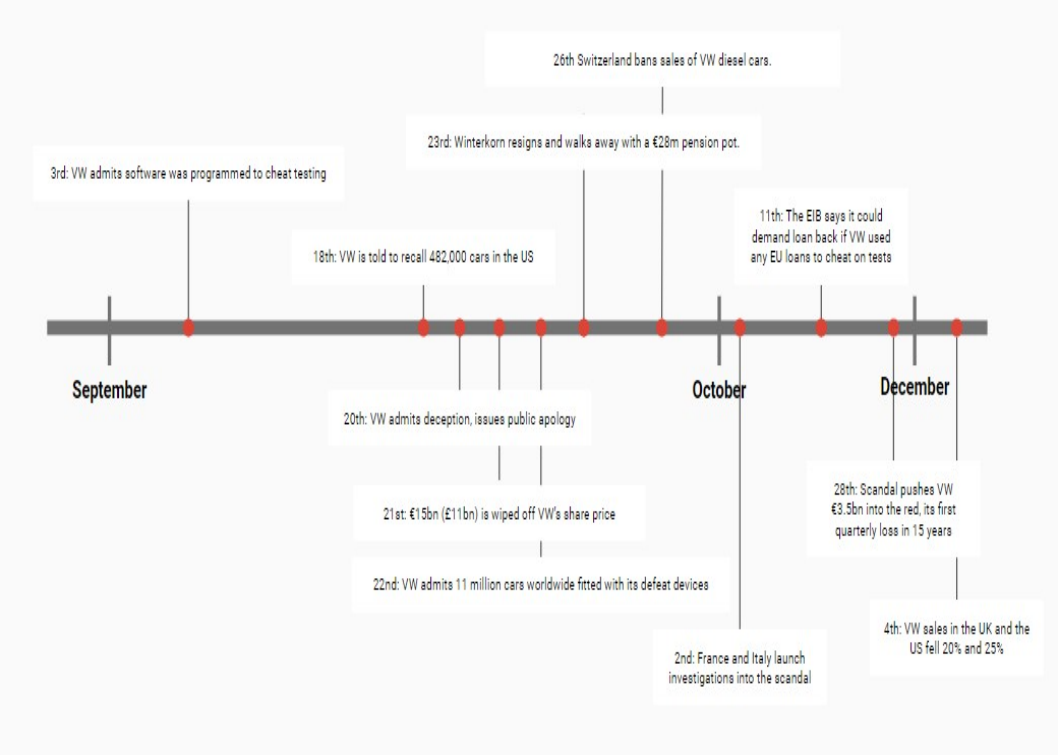
2.4. A Glance over Model Performance

In this section we offer some preliminary insights on the performance of the model[9]. To do so, we focus on one of the most popular ESG events over the past few years: the Volkswagen emission scandal in 2015, often known as "Dieselgate". The scandal broke out in September 2015 when the United States Environmental Protection Agency (EPA) issued a notice of violation of the Clean Air Act to Volkswagen. The EPA found that Volkswagen intentionally programmed turbocharged direct injection (TDI) diesel engines to activate their emission controls only during laboratory emissions testing, while they emitted up to 40 times more nitrogen oxides in real world driving. Figure 2 provides a timeline of the scandal.

[8] The alternative could have been to rely on a machine learning (ML) model to score the tweets. However, ML models, in particular state-of-the-art deep learning (DL) ones, are difficult to interpret due to their lack of transparency. For this reason, we decided for a rule-based algorithm that is transparent, reliable and accountable, without sacrificing performance for transparency.

[9] The model outputs E-S-G intraday scores that range between -1 and 1. Since the nature of raw textual scores is to be jagged, we always smooth the time series with a moving average.

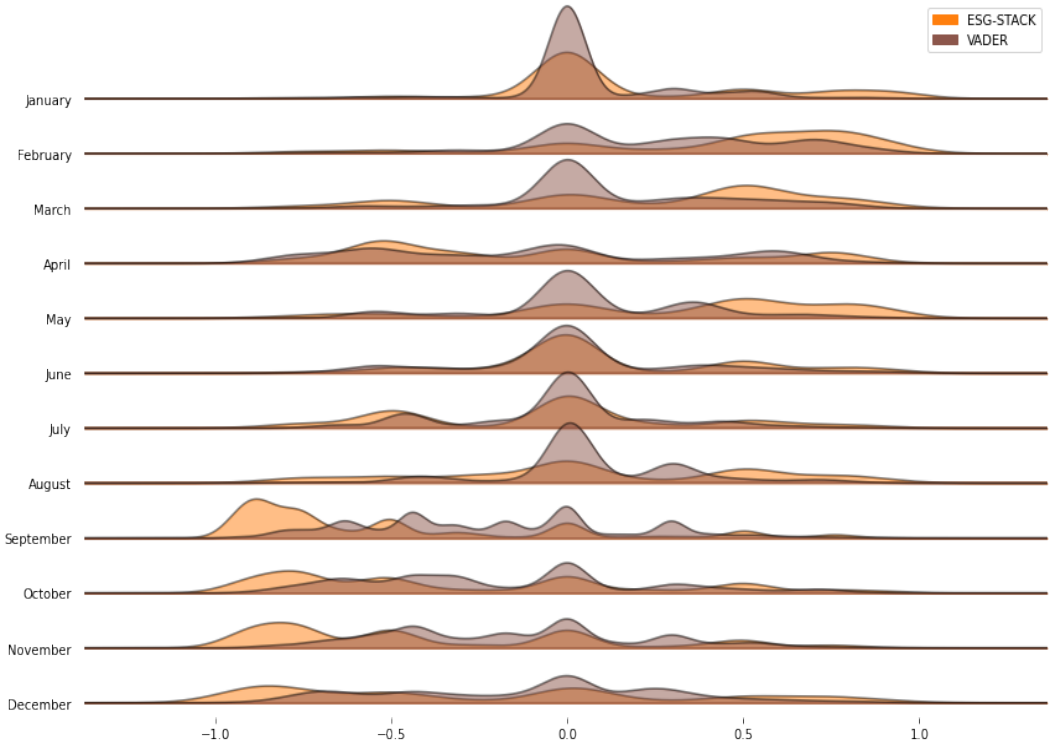
Figure 2: Timeline of Volkswagen Emission Scandal



Note: The figure summarizes the timeline of Volkswagen emission scandal.

We begin by testing the validity of our ESG dictionary and scoring algorithm against a baseline model widely used among researchers, namely, the VADER. We compare the distribution of the sentiment scores obtained with our model and those obtained with the VADER for the period January-December 2015. As Figure 3 illustrates, our model has fatter tails than the VADER. In particular, as the scandal broke out and heightened (September-December), our model shows a much more prominent positive skewness than the baseline one. This is to say that the mass of the distribution is concentrated on the left side, i.e. around negative values. Therefore, if we had used a generic dictionary, we would have not accurately grasped all the information contained in the tweets.

Figure 3: ESG Model vs VADER

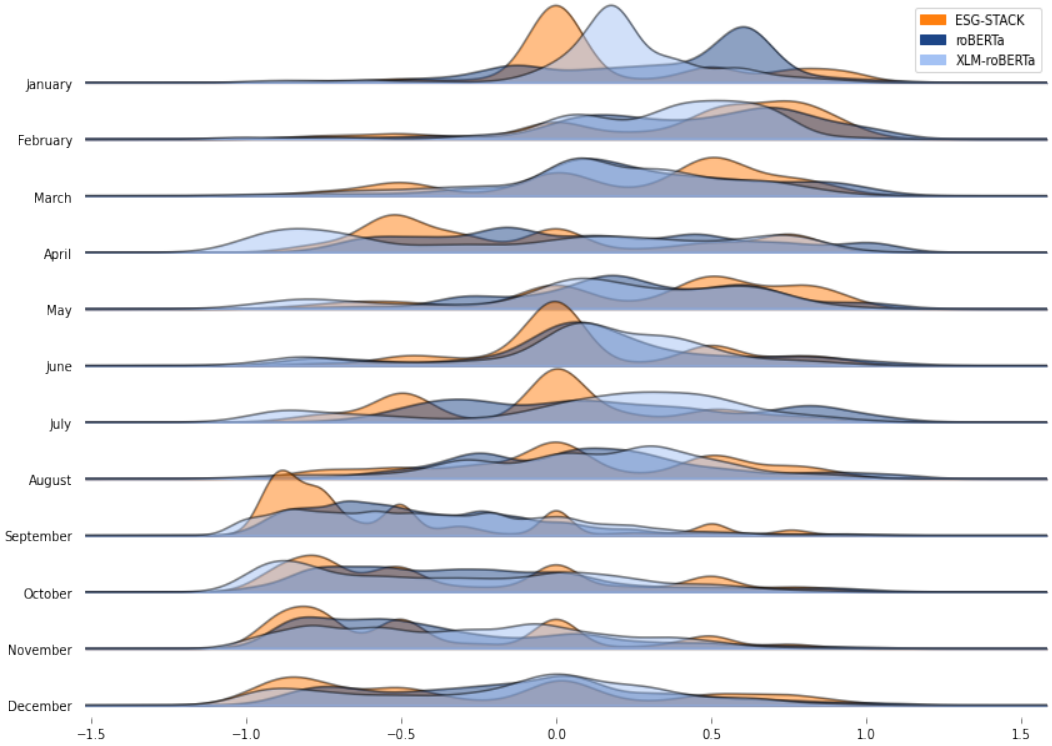


Note: The figure shows the distribution of sentiment scores for our ESG model vs VADER. The results are based on Volkswagen over the period January-December 2015.

Next, we test the performance of our classification model *vis-à-vis* state-of-the-art classificators such as BERT (Devlin et al., 2019), roBERTa (Liu et al., 2019) and XLM-roBERTa (Conneau et al., 2020). In this exercise, the dictionary and scoring algorithm are held constant, while the classification model is changed. Figure 4 shows the results[10]. If we leave aside a few outliers, there doesn't seem to be a significant difference in the distribution of the scores among different classification models. However, it can be noticed that, from September onward, our model more prominently catches negative scores compared to its competitors.

[10] The results for BERT are not included in Figure 4 since they overlapped with roBERTa's ones and made the figure harder to read. BERT's results can be seen in Table 3.

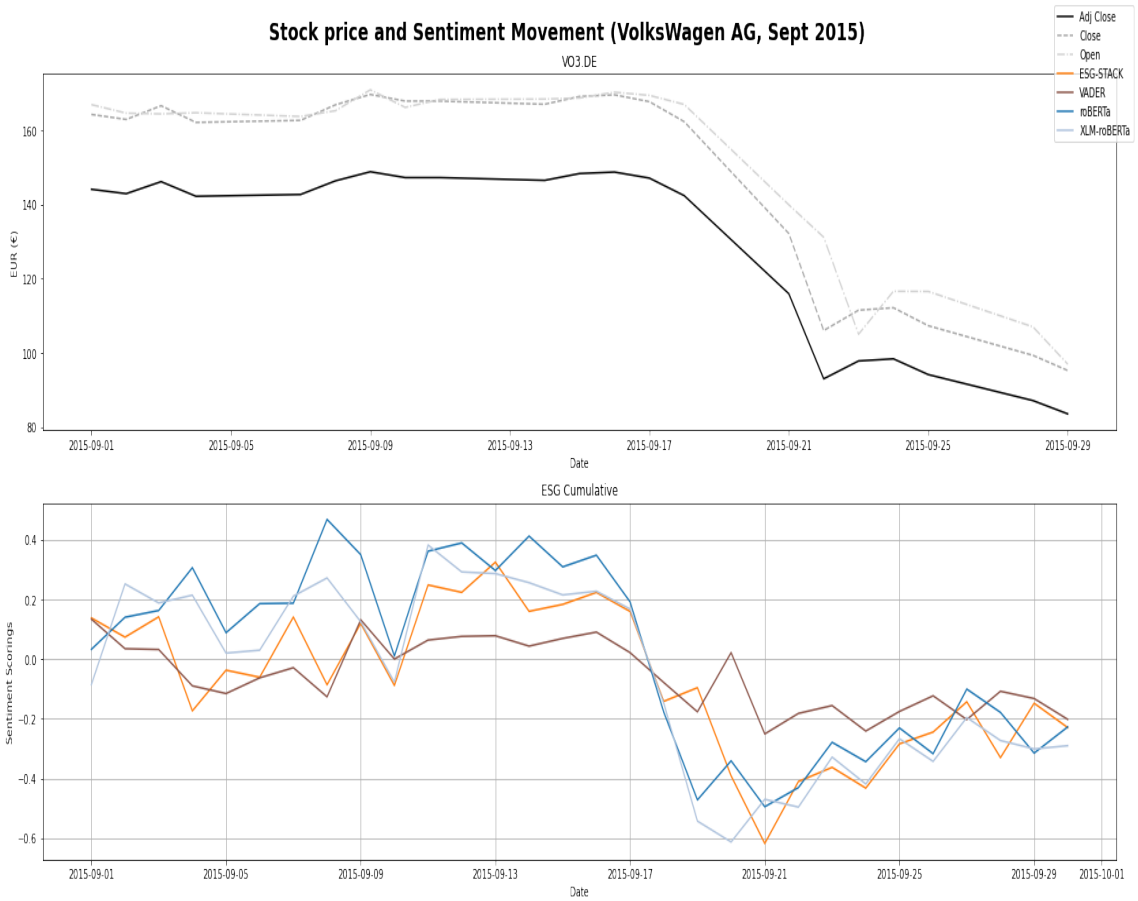
Figure 4: ESG Transformer Model vs State-of-the-Art Classificators



Note: The figure shows the distribution of the sentiment scores computed using our transformer model vs the scores obtained using state-of-the-art classificators such as roBERTa adn XLM-roBERTa. The results are based on Volkswagen over the period January-December 2015.

The results in Figure 3 and Figure 4 can be presented differently. Figure 5 shows the daily evolution of the stock price (first row) alongside that of the sentiment scores (second row) computed with different methodologies throughout September 2015, when the scandal materialized. The first striking result to highlight is that our model *anticipates the market sell-off by at least two days*. Second, a generic dictionary like VADER is unable to seize the severity of the event. It in fact remains rather flat over the time horizon and even peaks in the proximity of the scandal. Last, our model strongly correlates with other state-of-the-art models such as roBERTa and XLM-roBERTa.

Figure 5: Volkswagen Stock Price vs Sentiment Scores



Note: The figure shows the stock price (first row) and the sentiment scores of Volkswagen (second row) throughout September 2015.

Table 3 builds on the aforementioned points. Notably, TextBlob (Loria, 2013) - another generic dictionary and scoring system often used in NLP application - proves to be completely inadequate in tracking the evolution of the stock price at a sensitive juncture. Moreover, our model appears to be the most correlated one with the stock price of Volkswagen in September 2015, outperforming state-of-the-art models such as BERT, roBERTa and XLM-roBERTa.

Table 3: Pearson Correlation between Stock Price and Sentiment Scores

	Correlation	p-value
ESG Model	**0.60**	**0.000**
VADER	0.51	0.000
TextBlob	0.15	0.014
BERT	0.58	0.000
roBERTa	0.59	0.000
XLM-roBERTa	0.51	0.000

Note: The table illustrates the correlation between the stock price of Volkswagen from January to December 2015 and various sentiments computed with different methodologies.

In summary, the results presented in this section show that our model outperforms not only generic dictionaries and scoring algorithms that are widely applied to ESG topics but also state-of-the-art classification models that represent the backbone of models developed by companies such as Google AI and Facebook AI.

3. Empirical Applications

In this section, we provide a few empirical applications to further validate the robustness and usefulness of our model. In particular, Section 3.1 investigates the relation between our ESG metrics and the stock price at key ESG historical events. Section 3.2 proposes a Long Short-Term Memory (LSTM) model, trained using our E-S-G score, to forecast the stock returns. For our forecasting exercise, we chose a rather challenging time frame: the pandemic period, from January 2020 onward. In such a period, in fact, many long-standing and well-established models have failed to provide accurate results both in macroeconomic an financial applications.

3.1. Zooming in on ESG Events

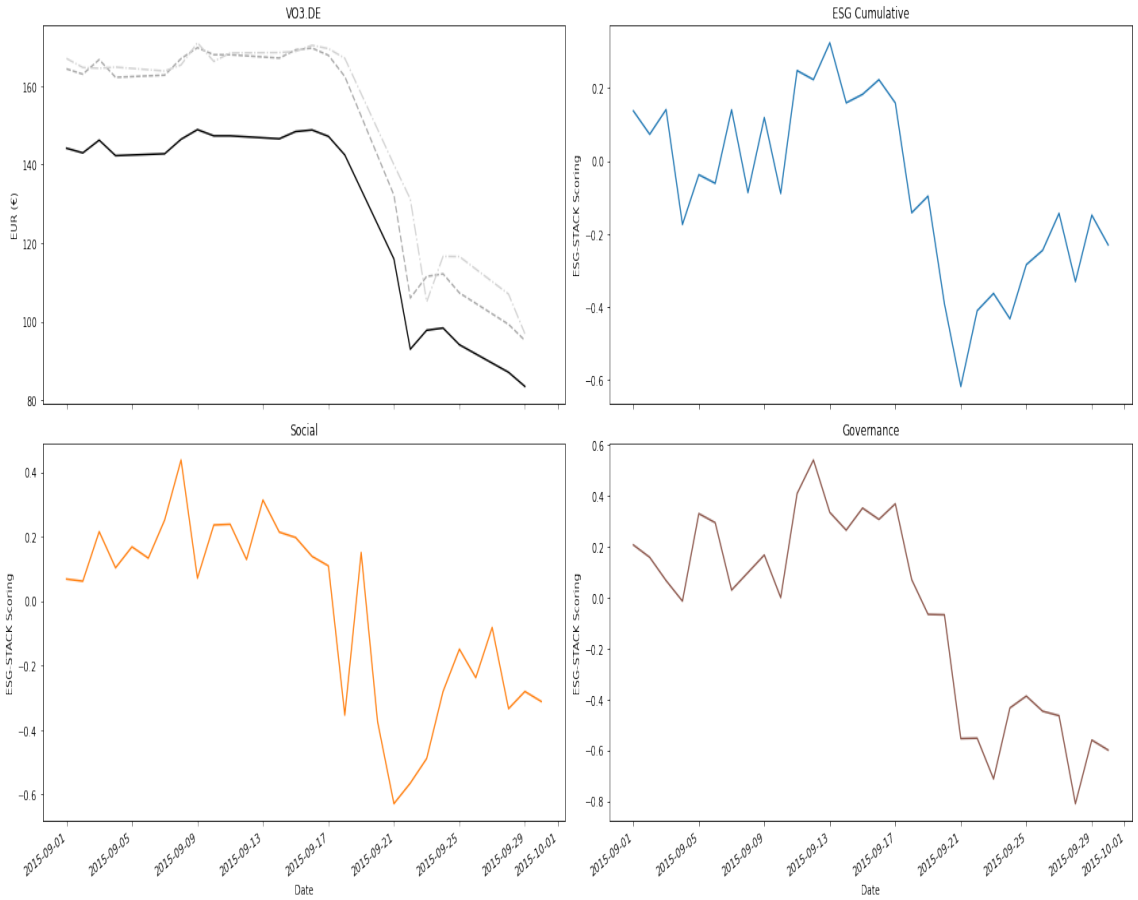
We document four case studies: Volkswagen, Facebook, Amazon and Bayer. For each company we study the relation of our ESG score with the stock price at key ESG historical events.

3.1.1 Volkswagen

The case of Volkswagen has already been studied in Section 2.4. We now complement that analysis with additional insights. For instance, Figure 6 provides more granular results for Social, Governance and ESG scores over the period January-December 2015[1]. Since we classified such a scandal as a governance topic, we begin focusing on the governance score (fourth quadrant). There is a significant correlation (71%) between the stock price and the "G" score. More importantly, the "G" score severely drops (from 0.25 to around -0.8) *2 days ahead of the stock price*. Similarly, the Social score (third quadrant) also dipped due to spillover effects before quickly recovering and stabilizing around -0.2. It is in fact hard to imagine that a corporate scandal doesn't affect the social dimension of a company. Averaging the "S" and "G" scores, we obtain the ESG score (second quadrant) that collapses *ahead of the market sell-off* but then recovers more quickly than the stock price, signalling *early* the stabilization of the market perception of Volkswagen's failure.

[1] The environmental score isn't displayed since the scandal is treated as a governance scandal. Therefore, "E" is not informative in this case

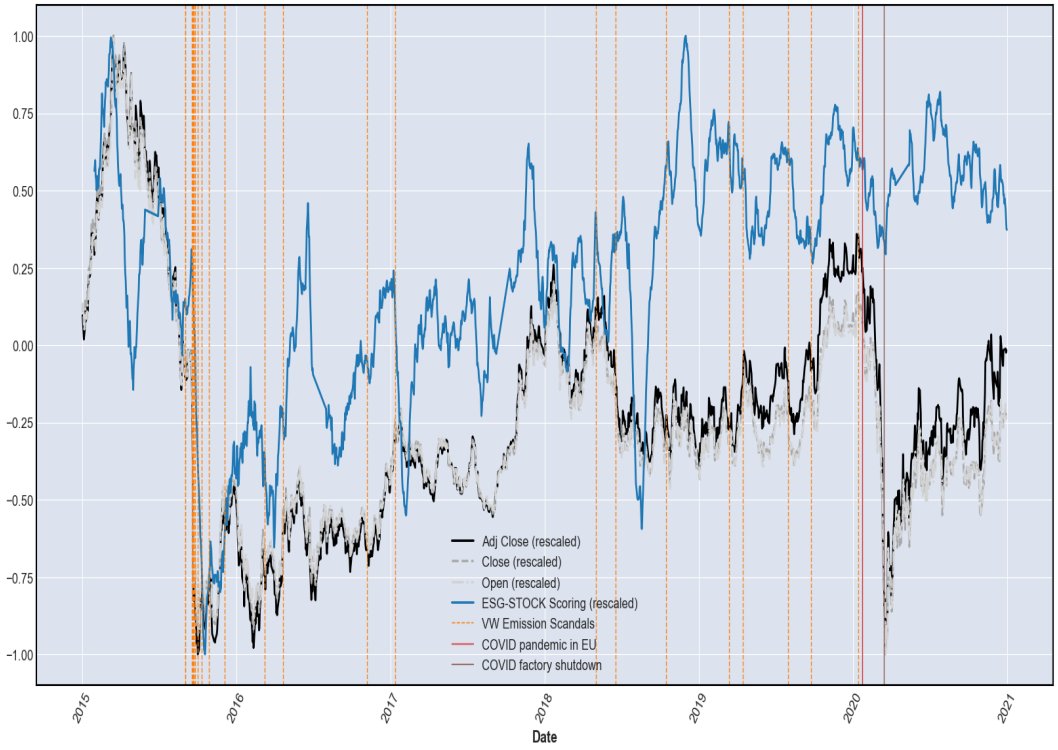
Figure 6: The Diesel Scandal



Note: The figure shows the stock price (first quadrant), ESG score (second quadrant), Social score (third quadrant) and Governance score (fourth quadrant) for Volkswagen over the time period January-December 2015.

Figure 7 displays the evolution of the stock price and of the sentiment score over the time frame 2015-2021. Two important results stand out: first, the sentiment score drops in correspondence of orange vertical lines denoting emission scandals. This provides evidence of the fact that our score is accurate and sensitive enough in tracking ESG events. Second, the stock price and the sentiment score appear to be decently correlated until the break-out of the pandemic. More precisely, while the stock price collapses at the start of the Covid pandemic (following also factory shutdowns), the ESG score does not seem to fully capture this event. This is evidence that our model is ESG-specific and isn't calibrated to respond to any generic, despite meaningful, event for the stock price.

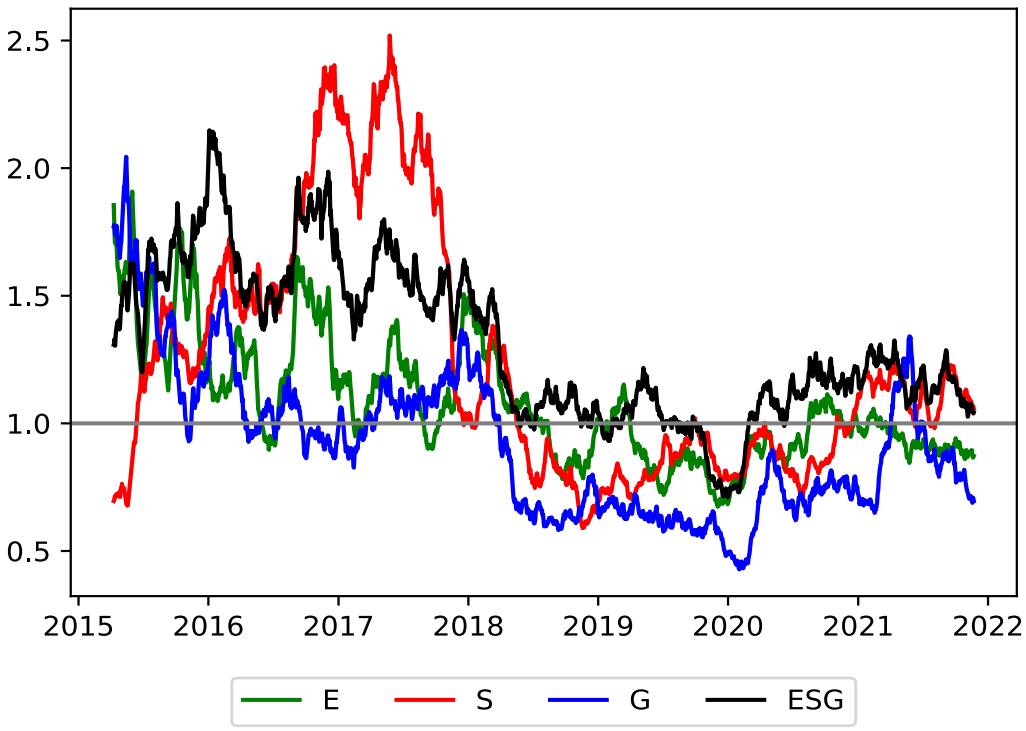
Figure 7: Stock Price vs ESG Sentiment Score



Note: The figure shows the stock price vs the ESG-sentiment score of Volkswagen from 2015 to 2021.

Drifting somewhat away from studying events, the last exercise we propose is to compare Volkswagen's scores with Tesla's ones, an important competitor in the auto-maker industry. To do so, we compute the ratio between Tesla's score for E, S, G and ESG and their corresponding scores for Volkswagen. Figure 8 illustrates the results.

Figure 8: Ratio between Tesla's Scores and Volkswagen's Ones



Note: The figure shows the ratio between Tesla's scores and Volkswagen's ones from 2015 to 2021.

According to our metrics, Tesla enjoyed a significant advantage over Volkswagen, especially on environmental and social factors, from 2015 to 2018. However, from 2018 onward, Volkswagen appear to have closed the gap with Tesla on any ESG dimension. Such analyses can be of great interest to investors since they can reveal ESG trend among competitors, thereby suggesting meaningful investment decisions.

3.1.2 Facebook

The second case study we propose is Facebook (FB)[2]. Since its foundation in 2004, FB has experienced various incidents highlighting inadequate and insufficient measures to protect data privacy. The most important one is certainly the "Cambridge Analytica scandal". On March 17, 2018, an exposé was published by *The Guardian* and *The New York Times*, initially reporting

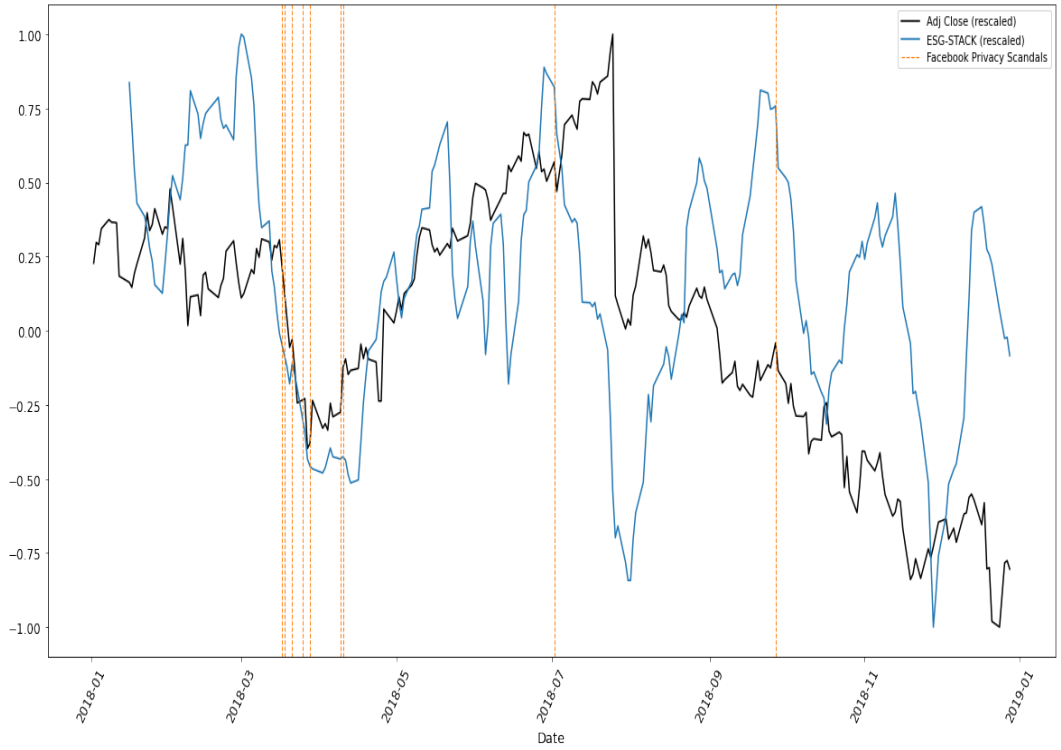
[2] The American social media multinational, founded in 2004, has now a market capitalization of around $954 billion and is currently owned by Meta.

that 50 million Facebook profiles were harvested by Cambridge Analytica[3]. The exposé relies on information provided by Christopher Wylie, a former employee of SCL Elections and Global Science Research, the creator of the "thisisyourdigitiallife app". Wylie claimed that the data from that app was sold to Cambridge Analytica, which used the data to develop "psychographic" profiles of users, and target users with pro-Trump advertising[4]. As a result of the news, Facebook wiped off more than $100 billion of market capitalization in the following days. Politicians in the US and UK also demanded answers from Facebook CEO Mark Zuckerberg. The negative public response to the media coverage eventually led to him agreeing to testify in front of the United States Congress.

Since the breach of data privacy is an ESG matter[5], we decided to focus on the time period January-December 2018 to test the responsiveness of our ESG sentiment score to different ESG-relevant events. Figure 9 illustrates Facebook's stock price *vis-à-vis* our ESG sentiment metrics over the period of interest.

[3] The figure was later revised to "up to 87 million" profiles.
[4] This claim was denied by Cambridge Analytica.
[5] Following our ESG Mapping classification, we classified incidents like *breach of data privacy* as a "Social" matter, under "Product Responsibility".

Figure 9: Stock Price vs ESG Sentiment Score



Note: The figure shows the stock price vs the ESG-sentiment score of Facebook from January to December 2018.

Figure 9 shows that our ESG metrics is incredibly sensitive to ESG news. In fact, the solid blue line plunges in correspondence of news related to data privacy scandals denoted by orange vertical lines. More importantly, our ESG score *anticipates the market sell-off* in two circumstances. First, while our ESG metrics starts collapsing on March 1, the market begins selling Facebook's shares on March 17 2018, when the exposé is published by *The Guardian* and *The New York Times*, *two weeks later*. Second, our metrics significantly dips once again on July 2 2018, when *The Washington Post* reports that the US Securities and Exchange Commission (SEC), Federal Trade Commission (FTC), and Federal Bureau of Investigation (FBI) joined the Department of Justice inquiry into the Facebook/Cambridge Analytica data scandal. Since then, our ESG score embarks in a free fall that hits -0.8 around 15 July. Instead, the market waits mid-July before selling FB's shares, *two week later* than the momentum signalled by the ESG score[6]. From this moment onward, FB's stock price begins a consistent declining path.

[6] Around mid-July, two pieces of news were published: first, on July 11, *The Wall Street Journal* reported that the SEC was separately investigating if Facebook adequately hard warned investors in a timely manner about

The ESG metrics, on the other hand, suffers from two other severe downturns: the first one on September 28, when FB discloses details of a security breach which affected 50 million users and, the second one, on November 14 when *The New York Times* publishes another exposé on the Facebook data privacy scandal, citing interviews of more than 50 people, including current and former Facebook executives and employees.

3.1.3 Amazon

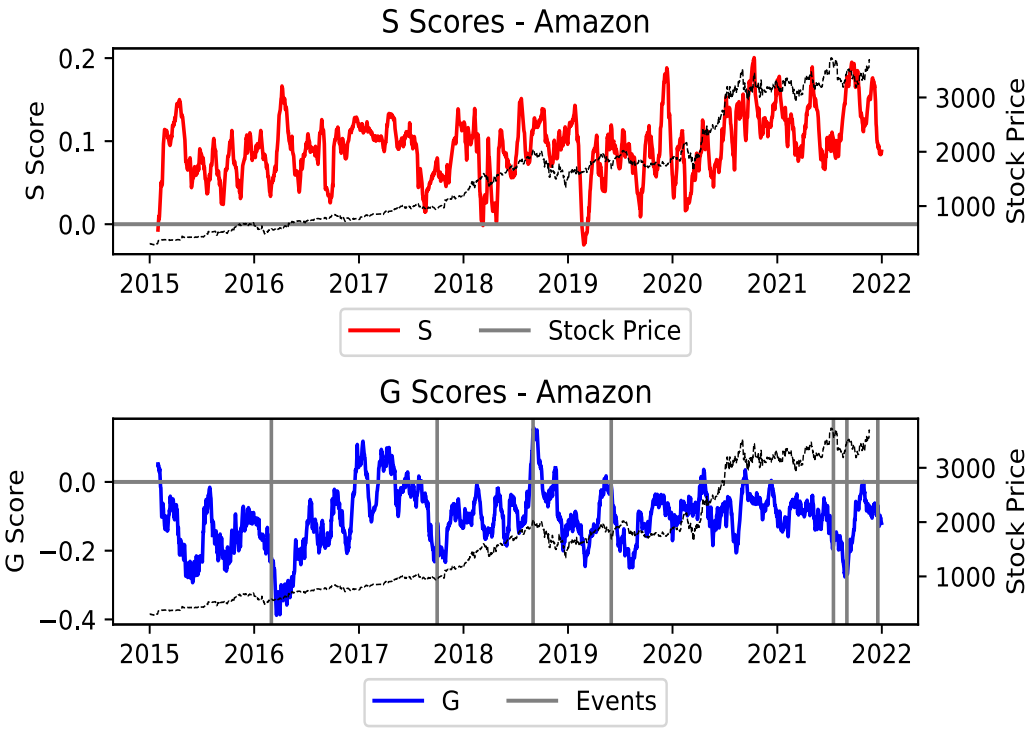
The third case study is Amzon, the e-commerce giant with current market capitalization of roughly 1.7 trillions. During his lifetime, Amazon has experienced various challenges both from a "social" and "governance" standpoint. Let's then investigate whether our metrics can shed some light on a few events that occurred to the e-commerce company.

Starting with our "Social" measure, we observe from the first row of Figure 10 that the red solid line is consistently in a positive territory throughout the timespan of interest. However, Amazon has frequently been accused of neglecting workers' needs in favor of productivity. Most of the times, news around Amazon's working conditions originated directly from Amazon workers who complained about unsafe working environments, excessive productivity goals, too few bathroom breaks and intrusive surveillance technologies.

How can we then match these accusations with a trending positive social score on Twitter? The reason lies in Amazon's "social policy". For instance, when Amazon has surplus inventory, the e-commerce giant donates the eligible items to charity organizations. Similarly, *AmazonSmile*, since its inception in 2013, donates every year considerable sums to charitable organizations that focus on animals, education, health and disasters. Besides, Jeff Bezos, Amazon's founder, is a philanthropist who very often tops the rankings of the 50 Americans who every year give the most to charity. Upon closer inspection, in fact, several Twitter users praise, rather than criticizing Amazon, for its social commitment. That's why we can observe the positive score for the online retail company.

the possible misuse and improper collection of user data. The same day, the UK assessed a £500,000 fine to Facebook, the maximum permitted by law, over its role in the data scandal. Second, on July 12, a CNBC report indicated that a privacy loophole was discovered and closed. A Chrome plug-in intended for marketing research called "Grouply.io" allowed users to access the list of members for private Facebook groups.

Figure 10: Social and Governance Scores: an Overview



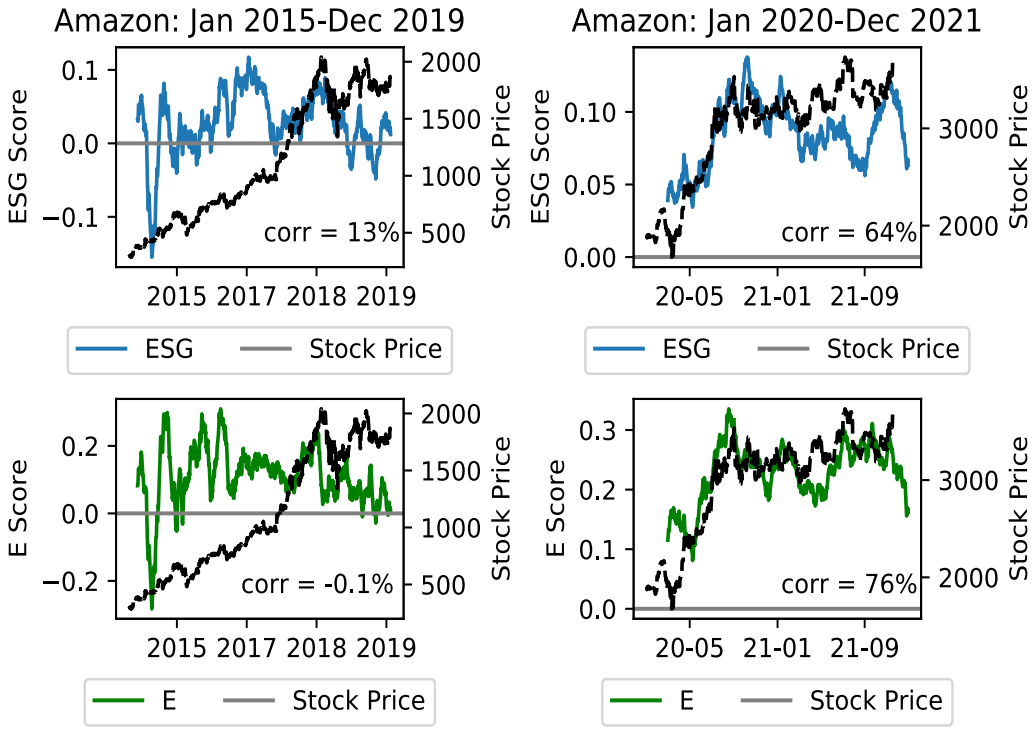
Note: The figure shows the social score (first row) and the governance score (second row) against the stock price of Amazon from January 2015 to December 2021.

We now focus on the "G" metrics. Unlike "S", "G" shows a persistent downward bias. This downtrend is primarily driven by contentious corporate governance events concerning tax-related issues and antitrust cases. For example, the governance score plunges at the beginning of February 2016 as a result of an article by *The Guardian*, revealing new details of an elaborate tax avoidance scheme called "Project Goldcrest" that European and US authorities began scrutinizing. The solid blue line slumps again in early June 2017 when *The Washington Post* publishes the news that antitrust regulators had put Amazon under the watch of the Federal Trade Commission (FTC). Similarly, on 19 Sept 2018, the "G" metrics starts declining again as European Competition Commissioner Margrethe Vestager begins questioning merchants on Amazon's use of their data.

Next, we take a closer look at Amazon's environmental performance. Figure 11 highlights some interesting findings. First, while the ESG and E scores are little correlated with the stock price from 2015 to 2019, the correlation rockets in the sub-period 2020-2021 where our scores accurately track the evolution of the stock price. We find a similar pattern in many other stocks

that, for the sake of brevity, we don't display. In our opinion, this is preliminary evidence that investors are increasingly embedding ESG-considerations into stock valuations and investment decisions. Second, Amazon's environmental score appears to be volatile from 2015 to 2016, it then remains rather stable in a positive territory until 2019 before marginally declining in 2019 and strongly rebounding from 2020 onward.

Figure 11: ESG and E Scores in Sub-periods of Time



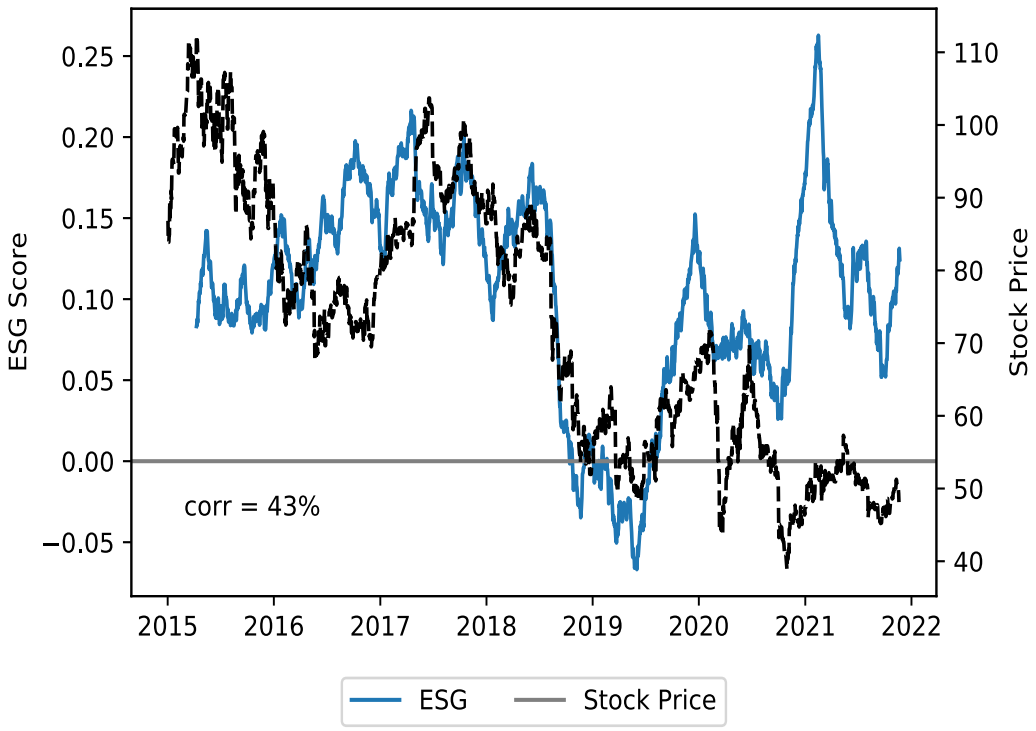
Note: The figure shows the stock price vs the ESG score (first row) and the stock price versus the environmental score (second row).

If we zoom in on the timeframe 2020-2021, we can claim that our environmental metrics is consistent with Amazon's renewed commitment to fighting climate change with initiatives such as *The Climate Pledge Fund*, *Sustainable Operations* and the launch of 100,000 electric delivery vans.

3.1.4 Bayer

We conclude our case studies with Bayer AG, one of the largest pharmaceutical companies in the world. Let's start by highlighting, from Figure 12, a 43% correlation between our ESG score and the evolution of Bayer's stock price.

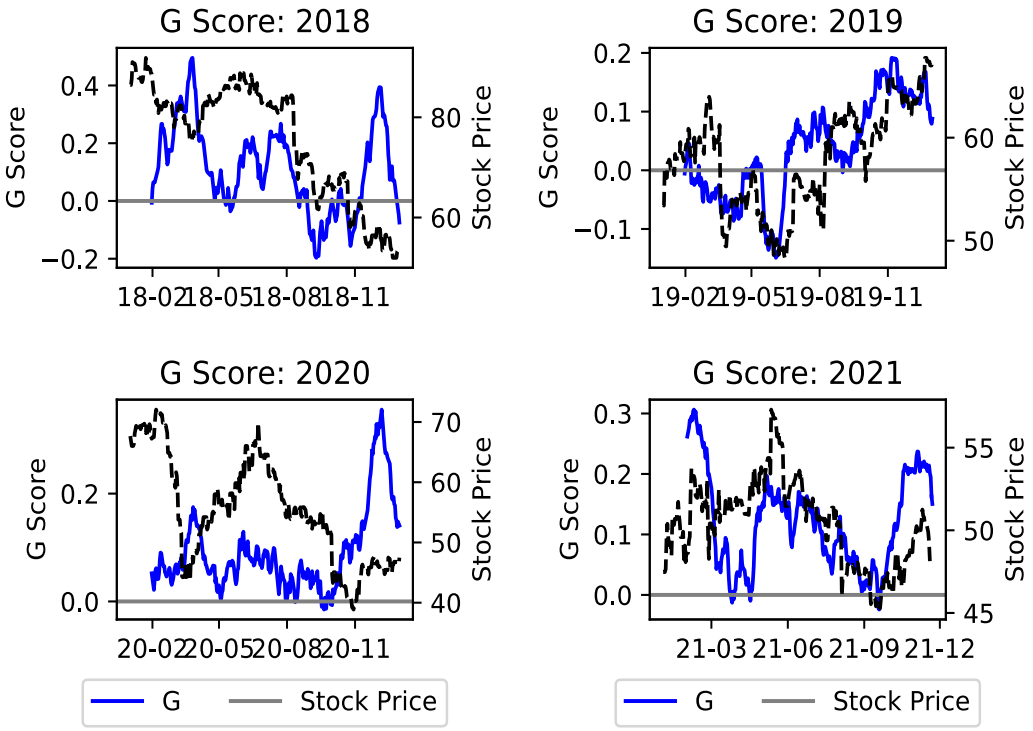
Figure 12: Stock Price vs ESG Sentiment Score



Note: The figure shows the stock price vs the ESG-sentiment score of Bayer from January to December 2018.

Figure 13 also shows a strong correlation, for each year from 2018 to 2021, between our "G" score and the stock price at different points in time.

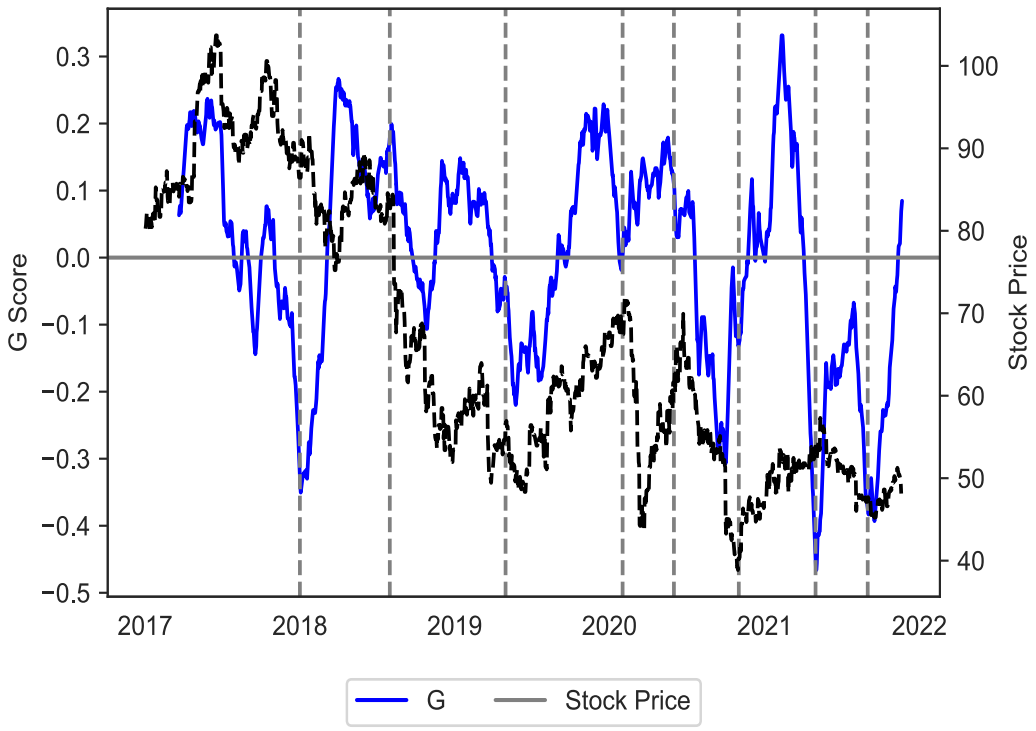
Figure 13: Stock Price vs G Score at Different Points in Time



Note: The figure shows the stock price vs the ESG-sentiment score of Bayer from January to December 2018.

However, rather than closely matching the stock price, the model is primarily built in order to timely detect ESG anomalies. Therefore, since Bayer has been affected by various ESG-relevant episodes, we now study whether our model is suited to capture those events. Figure 14 displays the results for the period 2017-2021.

Figure 14: Stock Price vs G Score



Note: The figure shows the stock price vs the G-sentiment score of Bayer from 2017 to December 2021. Gray shaded lines are G-relevant events.

For instance, in August 2018, two months after Bayer acquired Monsanto, a US jury ordered Monsanto to pay $289 million to a school groundskeeper who claimed his Non-Hodgkin's lymphoma was caused by regularly using Roundup, a glyphosate-based herbicide produced by Monsanto. In correspondence of the verdict, the G-score severely declines. Similarly, on 13 May 2019, a United States Superior Court Judge ordered Bayer to pay more than $2.5 billion in damages to a couple in California, both of whom contracted non-Hodgkin's Lymphoma[7]. As a result of this news, the G-score plunges once again into a deep negative territory. The last event we want to pay attention occurs on April 26 2021 when a Milan judge indicted the Italian units of Bayer on charges of operating a scheme to cheat the regional public health service in Lombardy. As a result of the news, the G-metrics plummets to hit an all-time low[8].

[7] The fine was later cut to $87 million on appeal.

[8] A few months later, on August 10, our score drops again following the news around Bayer having lost a third appeal against U.S. court verdicts that awarded damages to customers blaming their cancers on use of its glyphosate-based weedkillers. A California appeals court late on Monday upheld an $86 million verdict that found Bayer responsible for a couple's cancer after using Bayer's glyphosate-based Roundup against weeds.

3.1.5 Key Takeaways

We want to highlight four main takeaways from this section: 1) our metrics is proven to be extremely sensitive to environmental, social and governance events, responding in real-time and often *ahead of the market*; 2) correlation coefficients are not stable over the period of interest for the majority of the stocks. This is however an expected finding since our model isn't calibrated to respond to any generic, despite meaningful, event for the stock price. ESG factors, in fact, not always drive price dynamics. Rather than tracking the stock price closely, we believe that our metrics work at its finest as a real-time anomaly detector, capturing ESG relevant *momenta*; 3) although correlation coefficients might often be low across 2015-2021, they are significantly higher when the sample size is narrowed to be *within* a year of time frame (see for example, Amazon's example). This suggests that, with a shorter timespan, our score can also shed lights on share prices' dynamics; 4) for many stocks we analyzed, correlation coefficients tend to increase over the last couple of years. We believe this is not a fluke but it is rather likely due to the increasing integration of ESG factors into stock valuations and portfolio strategies.

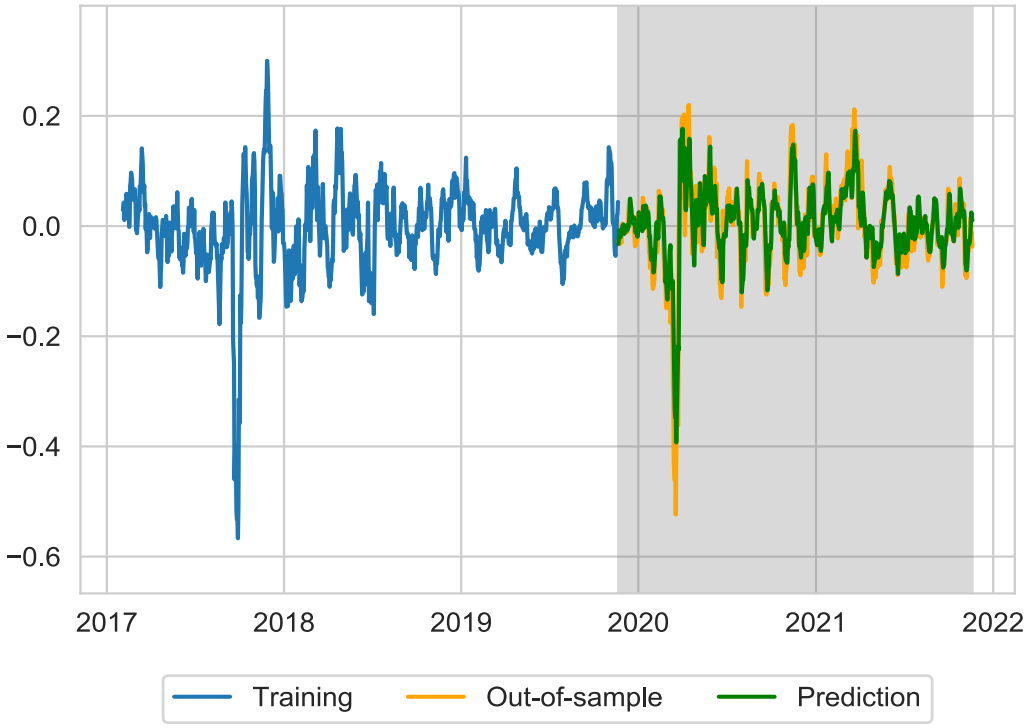
3.2. Forecasting Stock Returns

We then move to our forecasting exercises where we focus on Volkswagen and Facebook. For any company, our forecast horizon is January 2020-December 2021. We restate, once again, how this time frame is rather challenging from a forecasting standpoint. In fact, many well-established and long-standing time series models have failed to keep up with the evolution of the pandemic.

3.2.1 Volkswagen

To forecast stock returns, we employ an LSTM model (Hochreiter and Schmidhuber, 1997) that we train, from January 2015 to December 2019, with our ESG score. We then aim to predict out-of-sample Volkswagen's stock returns from January 2020 onward. Figure 8 shows the results. Despite our ESG score not being constructed to navigate the pandemic period, the forecasting exercise is quite successful. In fact, the predicted returns (green line) approximates the actual ones (orange line) throughout the forecast horizon. Strikingly, the model is also able to capture the severe drop due to the break-out of the pandemic.

Figure 15: Forecasting Volkswagen Stock Returns with ESG Sentiment Scores



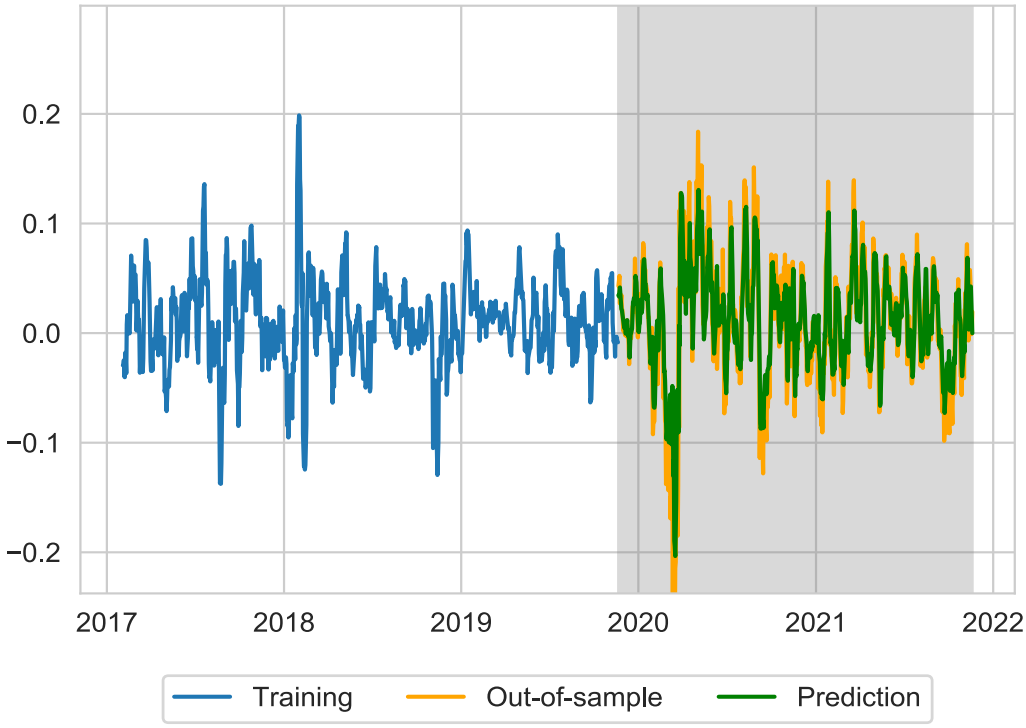
Note: The figure shows the result of our forecasting exercise where we used the ESG sentiment score to predict Volkswagen's stock returns. We used a LSTM model for this forecasting exercise.

The accurate performance of the model is also confirmed by standard statistics such as the mean squared forecast error (MSFE) and the mean absolute error (MAE), respectively being rather low at 0.0396 and 0.1252.

3.2.2 Facebook

Next, we use our ESG sentiment score to forecast Facebook's stock returns from January 2020 onward. In particular, we employ an LSTM model, trained on ESG scores, to predict out-of-sample Facebook's returns. Figure 10 documents the results.

Figure 16: Forecasting Facebook Stock Returns with ESG Sentiment Scores



Note: The figure shows the result of our forecasting exercise where we used the ESG sentiment score to predict Facebook's stock returns. We used a LSTM model for this forecasting exercise.

The result of the forecasting exercise is remarkable. The predicted returns (green line) accurately track the daily evolution of the actual ones (orange line). More strikingly, the model is able to capture the dips caused by the break-out of the Covid pandemic and the following lockdown. In this exercise, MSFE and MAE are even lower than the previous exercise, 0.0167 and 0.0993 respectively.

3.2.3 Key Takeaways

The key takeaway here is the following: although our ESG metrics can sometimes drift away from the evolution of the stock price, it might still provide a valuable integration to existing forecasting models.

4. ESG Integration in the Investment Process

Although it is beyond the scope of this paper to provide a fully-fledged model that integrates our ESG metrics into a comprehensive framework, we want to briefly touch upon a few potentially relevant applications for the ESG investment community. From a qualitative perspective, our model could help improve on "ESG red flags" as well as on "ESG score cards". As for the former, our model is equipped to flag emerging ESG risks and threats at a very early stage, enabling investors to capture the speed and intensity of their possible development. With regard to the latter, instead, our model would produce different moving averages as an additional score to be used in absolute and relative terms against sectors, geographic regions and index benchmarks. Those additional metrics would help "score cards" to have a more continuous computation and relative signalling power. From a quantitative point of view, we notice that the score momentum is a potentially valuable additional factor for either factor model portfolios or as a trigger for exclusionary-screening adjustments (see, for instance, PRI). Furthermore, from an active ownership standpoint, the model detects in real-time areas of further discussion and engagement with investee's companies and alerts them to respond timely to issues raised on social media.

According to PRI, a responsible investment is a strategy and practice that incorporates environmental, social and governance factors into investment decisions and active ownership. Within this framework, and compared to the existing tools, we attempted to prove that our ESG model offers many compelling properties (real-time, continuous, and forward looking) that could complement traditional approaches. We leave to our readers the task to further explore the integration of our metrics into a portfolio perspective and maximize the usage of our product.

5. Conclusion

Over the past few years, ESG has marked a revolution in the corporate and financial world. Nowadays, companies cannot avoid ESG considerations into their business models and corporate operations. At the same time, the "buy side" has been developing criteria through which it is possible to assess the ESG compliance and performance of publicly and privately owned companies. However, current benchmarks suffer from being *backward-looking* and *lagged*.

We propose a product that can overcome these difficulties. In fact, we provide a real-time

Twitter-based score tailored to each letter of ESG for more than 4000 publicly owned companies. We first develop an accurate classification model, relying on state-of-the-art neural networks, that identifies tweets related to ESG topics. We then create a unique ESG dictionary containing polarized ngrams and unigrams for environmental, social and governance matters. Last, we construct a new scoring algorithm that closely approximates the natural language and is able to capture social media nuances.

There are three main results to highlight: first, the model proves to *track accurately and timely ESG events* for more than 4000 stocks. Second, our metrics often moves *ahead of the market*. Third, further empirical applications show the *forecasting power* of our ESG scores in predicting stock returns *throughout the pandemic period*.

We therefore aim to equip investors with an additional instrument that, unlike existing benchmarks, is: *real-time*, *high-frequency* and *forward-looking*. Such a tool can complement current ESG scoring and rating systems, by enhancing the performance of many existing ESG investment strategies both quantitative and qualitative.

References

Ardia, D., K. Bluteau, K. Boudt, and K. Inghelbrecht (2020). Climate change concerns and the performance of green versus brown stocks. *National Bank of Belgium, Working Paper Research*.

Berg, F., J. Kölbel, and R. Rigobon (2020). Aggregate Confusion: The Divergence of ESG Ratings. *SSRN*.

Brandon, R. G., P. Krueger, and P. S. Schmidt (2021). Esg rating disagreement and stock returns. *Financial Analysts Journal 77*(4), 104–127.

Chatterji, A. K., R. Durand, D. I. Levine, and S. Touboul (2016). Do ratings of firms converge? implications for managers, investors and strategy researchers. *Strategic Management Journal 37*(8), 1597–1614.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov (2020). Unsupervised cross-lingual representation learning at scale.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Gu, C. and A. Kurov (2020). Informational role of social media: Evidence from twitter sentiment. *Journal of Banking & Finance 121*, 105969.

Hisano, R., D. Sornette, and T. Mizuno (2020). Prediction of esg compliance using a heterogeneous information network. *Journal of Big Data 07*(22).

Hochreiter, S. and J. Schmidhuber (1997, nov). Long short-term memory. *Neural Comput. 9*(8), 1735–1780.

Hutto, C. and E. Gilbert (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media 8*(1), 216–225.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach.

Loria, S. (2013). Textblob: Simplified text processing. https://github.com/sloria/textblob.

Marozzi, A. (2021). The ECB's Tracker: Nowcasting the Press Conferences of the ECB. *ECB Working Paper Series* (2609).

Nematzadeh, A., G. Bang, X. Liu, and Z. Ma (2019). Empirical study on detecting controversy in social media.

Raman, N., G. Bang, and A. Nourbakhsh (2020). Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction 2*(4), 453–468.

Rinker, T. W. (2019). *sentimentr: Calculate Text Polarity Sentiment.* Buffalo, New York. version 2.8.0.

Schmidt, A. (2019). Sustainable news - a sentiment analysis of the effect of esg information on stock prices. *SSRN*.

Shahi, A. M., B. Issac, and J. R. Modapothala (2014). Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications 13*(01).

Sun, A., M. Lachanski, and F. Fabozzi (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis 48*(C), 272–281.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need.