Semi-supervised Text Mining for Monitoring the News About the ESG Performance of Companies

Samuel Borms, Kris Boudt, Frederiek Van Holle, and Joeri Willems

Abstract We present a general monitoring methodology to summarize news about predefined entities and topics into tractable time-varying indices. The approach embeds text mining techniques to transform news data into numerical data, which entails the querying and selection of relevant news articles and the construction of frequency- and sentiment-based indicators. Word embeddings are used to achieve maximally informative news selection and scoring. We apply the methodology from the viewpoint of a sustainable asset manager wanting to actively follow news covering environmental, social, and governance (ESG) aspects. In an empirical analysis, using a Dutch-written news corpus, we create news-based ESG signals for a large list of companies and compare these to scores from an external data provider. We find preliminary evidence of abnormal news dynamics leading up to downward score adjustments and of efficient portfolio screening.

1 Introduction

Automated analysis of textual data such as press articles can help investors to better screen the investable universe. News coverage, how often news discusses a certain topic, and textual sentiment analysis, if news is perceived as positive or negative, serve as good proxies to detect important events and their surrounding perception.

S. Borms (✉)
Université de Neuchâtel, Neuchâtel, Switzerland

Vrije Universiteit, Brussels, Belgium
e-mail: samuel.borms@unine.ch

K. Boudt
Universiteit Gent, Ghent, Belgium

Vrije Universiteit, Brussels, Belgium
e-mail: kris.boudt@ugent.be

F. Van Holle · J. Willems
Degroof Petercam Asset Management, Brussels, Belgium
e-mail: f.vanholle@degroofpetercam.com; j.willems@degroofpetercam.com

S. Consoli et al. (eds.), *Data Science for Economics and Finance*,
https://doi.org/10.1007/978-3-030-66891-4_10

Text-based signals have at least the advantage of timeliness and often also that of complementary information value. The challenge is to transform the textual data into useful numerical signals through the application of proper text mining techniques.

Key research in finance employing text mining includes [13, 14, 24, 3]. These studies point out the impact of textual sentiment on stock returns and trading volume. Lately, the focus has shifted to using text corpora for more specific goals. For instance, Engle et al. [11] form portfolios hedged against climate change news based on news indicators.

This chapter takes the use of textual data science in sustainable investment as a running example. Investors with a goal of socially responsible investing (SRI) consider alternative measures to assess investment risk and return opportunities. They evaluate portfolios by how well the underlying assets align with a corporate social responsibility (CSR) policy—for instance, if they commit to environmental-friendly production methods. A corporation's level of CSR is often measured along the environmental, social and corporate governance (ESG) dimensions.

Investors typically obtain an investable universe of ESG-compliant assets by comparing companies to their peers, using a best-in-class approach (e.g., including the top 40% companies) or a worst-in-class approach (e.g., excluding the bottom 40% companies). To do so, investors rely on in-house research and third-party agency reports and ratings. Berg et al. [6], Amel-Zadeh and Serafeim [2], and Escrig-Olmedo et al. [12], among others, find that these ESG ratings are diverse, not transparent, and lack standardization. Moreover, most agencies only provide at best monthly updates. Furthermore, ratings are often reporting-driven and not signal-driven. This implies that a company can be ESG-compliant "by the book" when it is transparent (akin to greenwashing), but that the ratings are not an accurate reflection of the true current underlying sustainability profile.

In the remainder of the chapter, we introduce a methodology to create and validate news-based indicators allowing to follow entities and topics of interest. We then empirically demonstrate the methodology in a sustainable portfolio monitoring context, extracting automatically from news an objective measurement of the ESG dimensions. Moniz [19] is an exception in trying to infer CSR-related signals from media news using text mining in this otherwise largely unexplored territory.

2 Methodology to Create Text-Based Indicators

We propose a methodology to extract meaningful time series indicators from a large collection of texts. The indicators should represent the dimensions and entities one is interested in, and their time variation should connect to real-life events and news stories. The goal is to turn the indicators into a useful decision-making signal. This is a hard problem, as there is no underlying objective function to optimize, text data are not easy to explore, and it is computationally cumbersome to iterate frequently. Our methodology is therefore semi-supervised, altering between rounds of algorithmic estimation and human expert validation.

2.1 From Text to Numerical Data

A key challenge is to transform the stream of qualitative textual data into quantitative indicators. This involves first the selection of the relevant news and the generation of useful metadata, such as the degree to which news discusses an entity or an ESG dimension, or the sentiment of the news message. We tackle this by using domain-specific keywords to query a database of news articles and create the metadata. The queried articles need to undergo a second round of selection, to filter out the irrelevant news. Lastly, the kept corpus is aggregated into one or more time series.

To classify news as relevant to sustainability, we rely on keywords generated from a word embedding space. Moniz [19] uses a latent topic model, which is a probabilistic algorithm that clusters a corpus into a variety of themes. Some of these themes can then be manually annotated as belonging to ESG. We decide to go with word embeddings as it gives more control over the inclusion of keywords and the resulting text selection. Another approach is to train a named entity recognition (NER) model, to extract specific categories of concepts. A NER model tailored to ESG concepts is hard to build from scratch, as it needs fine-grained labeled data.

The methodology laid out below assumes that the corpus is in a single language. However, it can be extended to a multi-language corpus in various ways. The go-to approach, in terms of accuracy, is to consider each language separately by doing the indicators construction independently for every language involved. After that, an additional step is to merge the various language-specific indicators into an indicator that captures the evolution across all languages. One could, for simplicity, generate keywords in one language and then employ translation. Another common way to deal with multiple languages is to translate all incoming texts into a target language and then proceed with the pipeline for that language.

2.1.1 Keywords Generation

Three types of keywords are required. The **query lexicon** is a list of keywords per dimension of interest (*in casu*, the three ESG dimensions). Its use is twofold: first, to identify the articles from a large database with at least one of these keywords, and second, to measure the relevance of the queried articles (i.e., more keywords present in an article means it is more relevant). The **sentiment lexicon** is a list of words with an associated sentiment polarity, used to calculate document-level textual sentiment. The polarity defines the average connotation a word has, for example, -1 for "violence" or 1 for "happy." **Valence shifters** are words that change the meaning of other words in their neighborhood. There are several categories of valence shifters, but we focus on amplifiers and deamplifiers. An amplifier strengthens a neighboring word, for instance, the word "very" amplifies the word "strong" in the case of "very strong." Deamplifiers do the opposite, for example, "hardly" weakens the impact of "good" when "hardly good." The reason to integrate valence shifters in the sentiment

calculation is to better account for context in a text. The unweighted sentiment score of a document i with Q_i words under this approach is $s_i = \sum_{j=1}^{Q_i} v_{j,i} s_{j,i}$. The score $s_{j,i}$ is the polarity value attached in the sentiment lexicon to word j and is zero when the word is not in the lexicon. If word $j-1$ is a valence shifter, its impact is measured by $v_{j,i} = 1.8$ for amplifiers or $v_{j,i} = 0.2$ for deamplifiers. By default, $v_{j,i} = 1$.

To generate the keywords, we rely on expansion through a word embedding space. Word embeddings are vector representations optimized so that words closer to each other in terms of linguistic context have a more similar quantitative representation. Word embeddings are usually a means to an end. In our case, based on an initial set of seed keywords, analogous words can be obtained by analyzing the words closest to them in the embedding space. Many word embeddings computed on large-scale datasets (e.g., on Wikipedia) are freely available in numerous languages.[1] The availability of pretrained word embeddings makes it possible to skip the step of estimating a new word embedding space; however, in this chapter, we describe a straightforward approach to do the estimation oneself.

Word2Vec [18] and GloVe [21] are two of the most well-known techniques to construct a word embedding space. More recent and advanced methods include fastText [7] and the BERT family [9]. Word2Vec is structured as a continuous bag-of-words (CBOW) or as a skip-gram architecture, both relying only on local word information. A CBOW model tries to predict a given word based on its neighboring words. A skip-gram model tries to use a given word to predict the neighboring words. GloVe [21] is a factorization method applied to the corpus word-word co-occurrence matrix. A co-occurrence matrix stores the number of times a column word appears in the context of a row word. As such, GloVe integrates both global (patterns across the entire corpus) and local (patterns specific to a small context window) statistics. The intuition is that words which co-occur frequently are assumed to share a related semantic meaning. This is apparent in the co-occurrence matrix, where these words as a row-column combination will have higher values.

GloVe's optimization outputs two v-dimensional vectors per word (the word vector and a separate context word vector), that is, $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^v$. The final word vector to use is defined as $\boldsymbol{w} \equiv \boldsymbol{w}_1 + \boldsymbol{w}_2$. To measure the similarity between word vectors, say $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, the cosine similarity metric is commonly used. We define $cs_{ij} \equiv \boldsymbol{w}_i \boldsymbol{w}_j / \|\boldsymbol{w}_i\| \left\|\boldsymbol{w}_j\right\|$, where $\|.\|$ is the ℓ_2-norm. The measure $cs_{ij} \in [-1, 1]$, and the higher the more similar words i and j are in the embedding space.

Figure 1 displays the high-level process of expanding an initial set of seed words into the final three types of keywords needed. The seed words are the backbone of the analysis. They are defined manually and should relate strongly to the study domain. Alternatively, they can be taken from an existing lexicon, as done in [25] who start from the uncertainty terms in the Loughran and McDonald lexicon [17]. The seed words include both query seed words and sentiment seed words (often a

[1]For example, pretrained word embeddings by Facebook are available for download at https://fasttext.cc/docs/en/crawl-vectors.html.

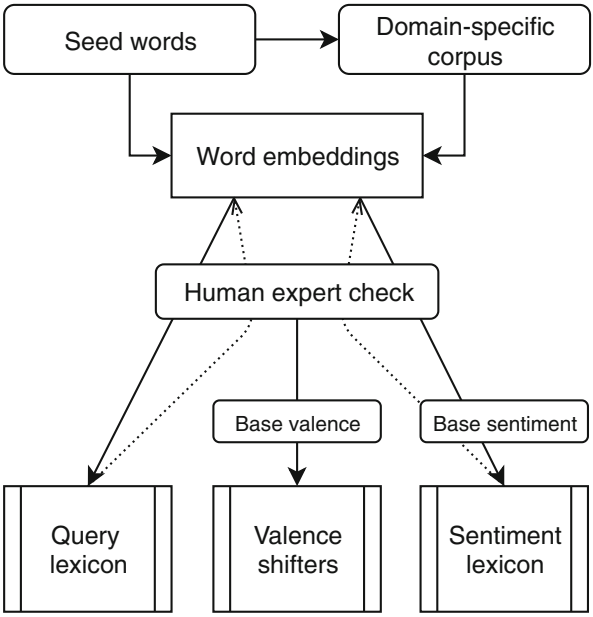

Fig. 1 Representation of the flow from seed words to the keywords of interest

subset of the former). The base valence and base sentiment word lists are existing dictionaries in need for a domain-specific twist to the application of interest.

All seed words are first used to query a more confined corpus from which the word embeddings will be estimated. The seed words are then expanded into the final query keywords by adding words that are similar, based on a ranking using the cs_{ij} metric and a human check. The human expert chooses between keeping the word, discarding the word, and assigning the word as a valence shifter. The same step is done for the sentiment seed words. As sentiment lexicons are typically larger, the words from a base sentiment lexicon not too far from the obtained query lexicon are added as well. The words coming from the word embeddings might be considered more important and thus weighted differently. The valence shifters are a combination of a base valence shifters list with the words assigned as a valence shifter. Section 3.2.1 further explains the implementation for the ESG use case.

This keywords generation framework has as limitation that it only considers unigrams, i.e., single words. Maintaining a valence shifters list adds a contextual layer in the textual sentiment calculation, and the number of keywords present in an article is a good overall indicator of the ESG relevance of news.

2.1.2 Database Querying

The database of texts is the large corpus that contains the subset of news relevant for the analysis. The task is to extract that subset as accurately as possible. The trade-off at play is that a large subset may guarantee full relevance, but it also adds more noise so it requires to think more carefully about the filtering step. In the process described in Fig. 1, a first query is needed to obtain a decent domain-specific corpus to estimate the embeddings.

Once the final query lexicon is composed, the batch of articles including the words in this lexicon as well as the entities to analyze needs to be retrieved and stored. To avoid a very time-consuming query, the querying is best approached as a loop over pairs of a given entity and the query lexicon keywords. A **list of entities** with the exact names to extract needs to be curated, possibly dynamic over time to account for name changes. Only the articles in which at least one entity name and at least one of the keywords is present are returned.

2.1.3 News Filtering

Keywords-based extraction does not guarantee that all articles retrieved are pertinent. It must be expected that a considerable degree of noise still remains. For example, press articles about a thief driving a BMW is not ESG-worthy news about the company BMW. Therefore, we recommend the following negative filters:

- Removal of texts that have no connection with the topic to study, for example, articles dealing with sports or lifestyle.
- Removal of articles that are too long (e.g., lengthy interviews) or too short (being more prone to a biased measurement of relevance and sentiment). Instead of removing the longer-than-usual articles, one could proceed with the leading paragraph(s) or a summary.
- Removal of exact duplicated entries or highly related (near-duplicated) entries.
- Removal of texts that are subject to database-specific issues, such as articles with a wrong language tag.

The level of filtering is a choice of the researcher. For instance, one can argue to leave (near-)duplicates in the corpus if one wants to represent the total news coverage, irrespective of whether the news rehashes an already published story or not. In this sense, it is also an option to reweight an article based on its popularity, proxied by the number of duplicates within a chosen interval of publication or by the number of distinct sources expressing related news.

2.1.4 Indicators Construction

A corpus with N documents between daily time points $t = 1, \ldots, T$ has a $N \times p$ matrix $\boldsymbol{Z}$ associated to it. This matrix maps the filtered corpus for a given entity

to p numerical metadata variables. It stores the values used for optional additional filtering and ultimately for the aggregation into the time series indicators. Every row corresponds to a news article with its time stamp. The number of articles at time t is equal to N_t, such that $N \equiv N_1 + \ldots + N_T$.

The ultimate indices are obtained applying a function $f : \boldsymbol{Z} \mapsto \boldsymbol{I}$, where $\boldsymbol{I}$ is a $U \times P$ time series matrix that represents the "suite" of P final text-based indices, with $U \leq T$. The (linear or nonlinear) aggregation function depends on the use case.

Specific computation of the metadata and the aggregation into indices are elaborated upon in the application described in Sect. 3.

2.2 Validation and Decision Making

Not all ESG information is so-called material. The created indicators only become useful when explicitly mapped into practical and validated decision-making signals.

Qualitative validation involves surveying the news to assess the remaining irrelevance of the articles. It also includes a graphical check in terms of peaks around the appearance of important events. Quantitative validation statistically measures the leading properties in regard to a certain target variable (e.g., existing sustainability scores) and the effectiveness of an investment strategy augmented with text-based information (in terms of out-of-sample risk and return and the stability and interpretation of formed portfolios).

In a real-life setting, when wanting to know which companies face a changing sustainability profile ("positives") and which not ("negatives"), false positives are acceptable but false negatives are typically not; in the same vein doctors do not want to tell sick patients they are healthy. It is more important to bring up all cases subject to a potentially changed underlying ESG profile (capturing all the actual positives at the cost of more false positives), rather than missing out on some (the false negatives) but bringing only the certain cases to the surface (merely a subset of the true positives). In machine learning classification lingo, this would mean aiming for excellent recall performance. An analyst will always proceed to investigation based on the signals received before recommending a portfolio action. Still, only an amount of signals that can reasonably be coped with should get through.

3 Monitoring the News About Company ESG Performance

In this section, we further motivate the integration of news-based ESG indices in sustainable investment practices. Secondly, we implement the described methodology and validate its applicability.

3.1 Motivation and Applications

We believe there is a high added value of news-implied time-varying ESG indicators for asset managers and financial analysts active in both risk management and investment. These two main types of applications in the context of sustainable investment are motivated below.

3.1.1 Text-Based ESG Scoring as a Risk Management Tool

According to [22], social preferences are the driving factor behind why investors are willing to forgo financial performance when investing in SRI-compliant funds. This class of investors might be particularly interested in enhanced ESG risk management. An active sustainable portfolio manager should react appropriately when adverse news comes out, to avoid investors becoming worried, as the danger of reputational damage lurks.

The degree to which a company is sustainable does not change much at a high frequency, but unexpected events such as scandals may immediately cause a corporation to lose its ESG-compliant stamp. An investor relying on low-frequency rating updates may be invested wrongly for an extended time period. Thus, it seems there is the need for a timelier filter, mainly to exclude corporations that suddenly cease to be ESG-compliant. News-based indicators can improve this type of negative screening. In fact, both negative and positive ESG screenings are considered among the most important future investment practices [2]. A universe of stocks can be split into a sustainable and a non-sustainable subuniverse. The question is whether news-based indicators can anticipate a change in the composition of the subuniverses.

Portfolio managers need to be proactive by choosing the right response among the various ESG signals they receive, arriving from different sources and at different times. In essence, this makes them an "ESG signals aggregator." The more signals, the more flexibility in the ESG risk management approach. An important choice in the aggregation of the signals is which value to put on the most timely signal, usually derived from news analysis.

Overall, the integration of textual data can lead to a more timely and a more conservative investment screening process, forcing asset managers as well as companies to continuously do well at the level of ESG transparency and ESG news presence.

3.1.2 Text-Based ESG Scoring as an Investment Tool

Increased investment performance may occur while employing suitable sustainable portfolio strategies or strategies relying on textual information. These phenomena are not new, but doing both at the same time has been less frequently investigated. A global survey by Amel-Zadeh and Serafeim [2] shows that the main reason for

senior investment professionals to follow ESG information is investment performance. Their survey does not discuss the use of news-based ESG data. Investors can achieve improved best-in-class stock selection or do smarter sector rotation. Targeted news-based indices can also be exploited as a means to tilt portfolios toward certain sustainability dimensions, in the spirit of Engle et al. [11]. All of this can generate extra risk-adjusted returns.

3.2 Pipeline Tailored to the Creation of News-Based ESG Indices

To display the methodology, we create text-based indices from press articles written in Dutch, for an assortment of European companies. We obtain the news data from the combined archive of the Belga News Agency and Gopress, covering all press sources in Belgium, as well as the major press outlets from the Netherlands. The data are not freely available.

The pipeline is incremental with respect to the companies and dimensions monitored. One can add an additional company or an extra sustainability (sub)dimension by coming up with new keywords and applying it to the corpus, which will result in a new specified time series output. This is important for investors that keep an eye on a large and changing portfolio, who therefore might benefit from the possibility of building the necessary corpus and indicators incrementally. The keywords and indicators can be built first with a small corpus and then improved based on a growing corpus. Given the historical availability of the news data, it is always easy to generate updated indicators for backtesting purposes. If one is not interested in defining keywords, one can use the keywords used in this work, available upon request.

3.2.1 Word Embeddings and Keywords Definition

We manually define the seed words drawing inspiration from factors deemed of importance by Vigeo Eiris and Sustainalytics, leading global providers of ESG research, ratings, and data. Environmental factors are for instance climate change and biodiversity, social factors are elements such as employee relations and human rights, and governance factors are, for example, anti-bribery and gender diversity. We define a total of 16, 18, and 15 seed words for the environmental, social, and governance dimensions, respectively. Out of those, we take 12 negative sentiment seed words. There are no duplicates across categories. Table 1 shows the seed words.

The time horizon for querying (and thus training the word embeddings) spans from January 1996 to November 2019. The corpus is queried separately for each dimension using each set of seed words. We then combine into a large corpus, consisting of 4,290,370 unique news articles. This initial selection assures a degree

Table 1 Dutch E, S, G, and negative sentiment seed words

E	S	G	Sentiment[a]
milieu (*environment*), energie (*energy*), mobiliteit (*mobility*), nucleair (*nuclear*), klimaat (*climate*), biodiversiteit (*biodiversity*), koolstof (*carbon*), vervuiling (*pollution*), water, verspilling (*waste*), ecologie (*ecology*), duurzaamheid (*sustainability*), uitstoot (*emissions*), hernieuwbaar (*renewable*), olie (*oil*), olielek (*oil leak*)	samenleving (*society*), gezondheid (*health*), mensenrechten (*human rights*), sociaal (*social*), discriminatie (*discrimination*), inclusie (*inclusion*), donatie (*donation*), staking (*strike*), slavernij (*slavery*), stakeholder, werknemer (*employee*), werkgever (*employer*), massaontslag (*mass fire*), arbeid (*labor*), community, vakbond (*trade union*), depressie (*depression*), diversiteit (*diversity*)	gerecht (*court*), budget, justitie (*justice*), bestuur (*governance*), directie (*management*), omkoping (*bribery*), corruptie (*corruption*), ethiek (*ethics*), audit, patentbreuk (*patent infringement*), genderneutraal (*gender neutral*), witwaspraktijken (*money laundering*), dierproeven (*animal testing*), lobbyen (*lobbyism*), toploon (*top wage*)	vervuiling, verspilling, olielek, discriminatie, staking, slavernij, massaontslag, depressie, omkoping, corruptie, patentbreuk, witwaspraktijken

[a] These are a subset of the words in E, S, and G

of domain specificity in the obtained word vectors, as taking the entire archive would result in a too general embedding.

We tokenize the corpus into unigrams and take as vocabulary the 100,000 most frequent tokens. A preceding cleaning step drops Dutch stop words, all words with less than 4 characters, and words that do not appear in at least 10 articles or in more than 10% of the corpus. We top the vocabulary with the 49 ESG seed words.

To estimate the GloVe word embeddings, we rely on the R package **text2vec** [23]. We choose a symmetric context window of 7 words and set the vector size to 200. Word analogy experiments in [21] show that a larger window or a larger vector size does not result in significantly better accuracy. Hence, this hyperparameters choice offers a good balance between expected accuracy and estimation time. In general, small context windows pick up substitutable words (e.g., due to enumerations), while large windows tend to better pick up topical connections. Creating the word embeddings is the most time-consuming part of the analysis, which might take from start to finish around half a day on a regular laptop. Figure 2 shows the fitted embedding space, shrunk down to two dimensions, focused on the seed words "duurzaamheid" and "corruptie."

To expand the seed words, for every seed word in each dimension, we start off with the 25 closest words based on cs_{ij}, i.e., those with the highest cosine similarity. By hand, we discard irrelevant words or tag words as an amplifying or as a deamplifying valence shifter. An example in the first valence shifter category is "chronische" (*chronic*), and an example in the second category is "afgewend" (*averted*). We reposition duplicates to the most representative category. This leads to

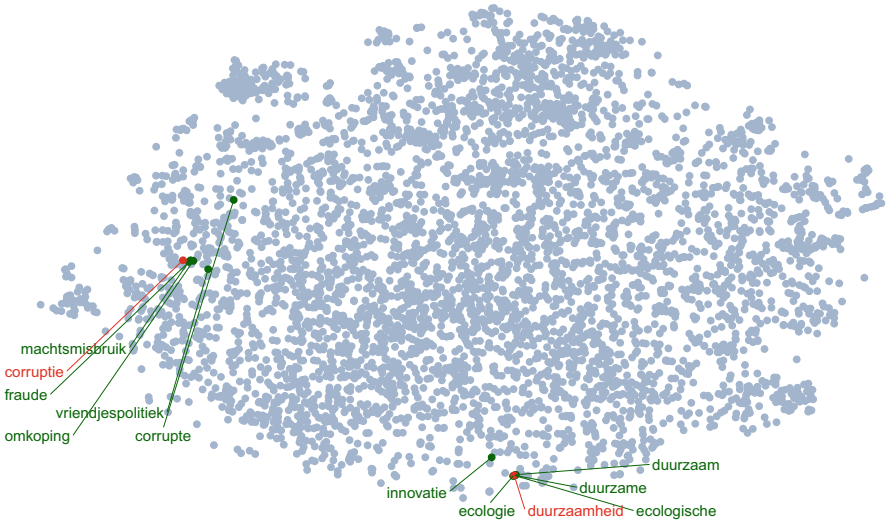

Fig. 2 Visualization of the embedding for a 5% fraction of the 100,049 vocabulary words. The t-distributed stochastic neighbor embedding (t-SNE) algorithm implemented in the R package **Rtsne** [15] is used with the default settings to reduce the 200-dimensional space to a two-dimensional space. In red, focal seed words "duurzaamheid" and "corruptie," and in green the respective five closest words according to the cosine similarity metric given the original high-dimensional word embeddings

197, 226, and 166 words, respectively, for the environmental, social, and governance dimensions.

To expand the sentiment words, we take the same approach. The obtained words (151 in total) receive a polarity score of -2 in the lexicon. From the base lexicon entries that also appear in the vocabulary, we discard the words for which none of its closest 200 words is an ESG query keyword. If at least one of these top 200 words is a sentiment seed word, the polarity is set to -1 if not already. In total, the sentiment lexicon amounts to 6163 words, and we consider 84 valence shifters.

3.2.2 Company Selection and Corpus Creation

To query the news related to companies, we use a reasonable trade-off between their commonplace name and their legal name.[2] Counting the total entity occurrences

[2]Suffixes (e.g., N.V. or Ltd.) and too generic name parts (e.g., International) are excluded. We also omit companies with names that could be a noun or a place (for instance, Man, METRO, Partners, Restaurant, or Vesuvius). Our querying system is case-insensitive, but case sensitivity would solve the majority of this problem. We only consider fully merged companies, such as Unibail-Rodamco-Westfield and not Unibail-Rodamco.

(measured by $n_{i,t}$; see Sect. 3.2.3) happens less strict by also accounting for company subnames. Our assumption is that often the full company name is mentioned once, and further references are made in an abbreviated form. As an example, to query news about the company Intercontinental Hotels, we require the presence of "Intercontinental" and "Hotels," as querying "Intercontinental" alone would result in a lot of unrelated news. To count the total matches, we consider both "Intercontinental" and "Intercontinental Hotels."

We look at the 403 European companies that are included in both the Sustainalytics ESG dataset (ranging from August 2009 to July 2019) and (historically) in the S&P Europe 350 stock index between January 1999 and September 2018. The matching is done based on the tickers.

We run through all filters enumerated in Sect. 2.1.3. Articles without minimum 450 or with more than 12,000 characters are deleted. To detect near-duplicated news, we use the locality-sensitive hashing approximate nearest neighbor algorithm [16] as implemented in the R package **textreuse** [20].

In total, 1,453,349 company-specific and sustainability-linked news articles are queried, of which 1,022,898 are kept after the aforementioned filtering. On average 33.4% of the articles are removed. Most come from the removal of irrelevant articles (20.5 p.p.); only a minor part is the result of filtering out too short and too long articles (6.4 p.p.). Pre-filtering, 42.2%, 71%, and 64.3% are marked belonging to the E, S, or G dimension, respectively. Post-filtering, the distribution is similar (38.1%, 70.2%, and 65.9%). Additionally, we drop the articles which have only one entity mention. The total corpus size falls to 365319. The strictness of this choice is to avoid the inclusion of news in which companies are only mentioned in passing [19]. Furthermore, companies without at least 10 articles are dropped. We end up with 291 of the companies after the main filtering procedure and move forward to the index construction with for each company a corpus.

3.2.3 Aggregation into Indices

As discussed in Sect. 2.1.4, we define a matrix $\boldsymbol{Z}_e$ for every entity e (i.e., a company) as follows:

$$\boldsymbol{Z}_e = \begin{bmatrix} n_{1,1} & n^E_{1,1} & n^S_{1,1} & n^G_{1,1} & a^E_{1,1} & a^S_{1,1} & a^G_{1,1} & s_{1,1} \\ n_{2,1} & n^E_{2,1} & n^S_{2,1} & n^G_{2,1} & a^E_{2,1} & a^S_{2,1} & a^G_{2,1} & s_{2,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{i,t} & n^E_{i,t} & n^S_{i,t} & n^G_{i,t} & a^E_{i,t} & a^S_{i,t} & a^G_{i,t} & s_{i,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{N^e-1,T} & n^E_{N^e-1,T} & n^S_{N^e-1,T} & n^G_{N^e-1,T} & a^E_{N^e-1,T} & a^S_{N^e-1,T} & a^G_{N^e-1,T} & s_{N^e-1,T} \\ n_{N^e,T} & n^E_{N^e,T} & n^S_{N^e,T} & n^G_{N^e,T} & a^E_{N^e,T} & a^S_{N^e,T} & a^G_{N^e,T} & s_{N^e,T} \end{bmatrix}. \tag{1}$$

The computed metadata for each news article are the number of times the company is mentioned (column 1); the total number of detected keywords for the E, S, and G dimensions (columns 2 to 4); the proportions of the E, S, and G keywords w.r.t. one another (columns 5 to 7); and the textual sentiment score (column 8). More specifically, n counts the number of entity mentions; n^E, n^S, and n^G count the number of dimension-specific keywords; and s is the textual sentiment score. The proportion $a^d_{i,t}$ is equal to $n^d_{i,t}/(n^E_{i,t} + n^S_{i,t} + n^G_{i,t})$, for d one of the sustainability dimensions. It measures something distinct from keywords occurrence—for example, two documents can have the same number of keywords of a certain dimension yet one can be about one dimension only and the other about all three.

The sentiment score is calculated as $s_{i,t} = \sum_{j=1}^{Q_{i,t}} \omega_{j,i,t} v_{j,i,t} s_{j,i,t}$, where $Q_{i,t}$ is the number of words in article i at time t, $s_{j,i,t}$ is the polarity score for word j, $v_{j,i,t}$ is the valence shifting value applied to word j, and $\omega_{j,i,t}$ is a weight that evolves as a U-shape across the document.[3] To do the sentiment computation, we use the R package **sentometrics** [4].[4]

The metadata variables can also be used for further filtering, requiring, for instance, a majority proportion of one dimension in an article to include it. We divide $\boldsymbol{Z}_e$ into $\boldsymbol{Z}_{e,E}$, $\boldsymbol{Z}_{e,S}$, and $\boldsymbol{Z}_{e,G}$. In those subsets, we decide to keep only the news entries for which $n^d_{i,t} \geq 3$ and $a^d_{i,t} > 0.5$, such that each sustainability dimension d is represented by articles maximally related to it. This trims down the total corpus size to 166020 articles.[5]

For a given dimension d, the time series matrix that represents the suite of final text-based indices is a combination of 11 frequency-based and 8 sentiment-adjusted indicators. We do the full-time series aggregation in two steps. This allows separating out the first simple from the subsequent (possibly time) weighted daily aggregation. We are also not interested in relative weighting within a single day; rather we will utilize absolute weights that are equally informative across the entire time series period.

We first create daily $T \times 1$ frequency vectors $\boldsymbol{f}$, $\boldsymbol{p}$, $\boldsymbol{d}$, and $\boldsymbol{n}$ and a $T \times 1$ vector $\boldsymbol{s}$ of a daily sentiment indicator. For instance, $\boldsymbol{f} = (f_1, \ldots, f_t, \ldots, f_T)'$ and $\boldsymbol{f}_{[k,u]} = (f_k, \ldots, f_t, \ldots, f_u)'$. The elements of these vectors are computed starting from the

[3]Notably, $\omega_{j,i,t} = c\left(j - (Q_{i,t} + 1)/2\right)^2$ with c a normalization constant. Words earlier and later in the document receive a higher weight than words in the middle of the document.

[4]See the accompanying package website at https://sentometricsresearch.github.io/sentometrics for code examples, and the survey paper by Algaba et al. [1] about the broader sentometrics research field concerned with the construction of sentiment indicators from alternative data such as texts.

[5]For some companies the previous lower bound of 10 news articles is breached, but we keep them aboard. The average number of documents per company over the embedding time horizon is 571.

submatrix $\boldsymbol{Z}_{e,d}$, with at any time $N_t^{e,d}$ articles, as follows:

$$f_t = N_t^{e,d}, \quad p_t = 1/N_t^{e,d} \sum_{i=1}^{N_t^{e,d}} a_{i,t}^d, \quad d_t = \sum_{i=1}^{N_t^{e,d}} n_{i,t}^d, \quad n_t = \sum_{i=1}^{N_t^{e,d}} n_{i,t}. \tag{2}$$

For sentiment, $s_t = 1/N_t^{e,d} \sum_{i=1}^{N_t^{e,d}} s_{i,t}$. Missing days in $t = 1, \ldots, T$ are added with a zero value. Hence, we have that $\boldsymbol{f}$ is the time series of the number of selected articles, $\boldsymbol{p}$ is the time series of the average proportion of dimension-specific keyword mentions, $\boldsymbol{d}$ is the time series of the number of dimension-specific keyword mentions, and $\boldsymbol{n}$ is the time series of the number of entity mentions. Again, these are all specific to the dimension d.

The second step aggregates the daily time series over multiple days. The weighted frequency indicators are computed as $\boldsymbol{f}'_{[k,u]} \boldsymbol{B}_{[k,u]} \boldsymbol{W}_{[k,u]}$, with $\boldsymbol{B}_{[k,u]}$ a $(u - k + 1) \times (u - k + 1)$ diagonal matrix with the time weights $\boldsymbol{b}_{[k,u]} = (b_k, \ldots, b_t, \ldots, b_u)'$ on the diagonal, and $\boldsymbol{W}_{[k,u]}$ a $(u-k+1) \times 7$ metadata weights matrix defined as:

$$\boldsymbol{W}_{[k,u]} = \begin{bmatrix} p_k & g(d_k) & h(n_k) & p_k g(d_k) & p_k h(n_k) & g(d_k)h(n_k) & p_k g(d_k)h(n_k) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_t & g(d_t) & h(n_t) & p_t g(d_t) & p_t h(n_t) & g(d_t)h(n_t) & p_t g(d_t)h(n_t) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_u & g(d_u) & h(n_u) & p_u g(d_u) & p_u h(n_u) & g(d_u)h(n_u) & p_u g(d_u)h(n_u) \end{bmatrix}, \tag{3}$$

where $g(x) = \ln(1 + x)$ and $h(x) = x$. In our application, we choose to multiplicatively emphasize the number of keywords and entity mentions but alleviate the effect of the first, as in rare cases disproportionately many keywords pop up. The value p_t is a proportion between 0 and 1 and requires no transformation. The aggregate for the last column is $\sum_{t=k}^{u} f_t b_t p_t \ln(1 + d_t) n_t$, for instance.

The aggregations repeated for $u = \tau, \ldots, T$, where τ pinpoints the size of the first aggregation window, give the time series. They are assembled in a $U \times 7$ matrix of column vectors. Every vector represents a different weighting of the obtained information in the text mining step.

We opt for a daily moving fixed aggregation window $[k, u]$ with $k \equiv u - \tau + 1$. As a time weighting parameter, we take $b_t = \alpha_t / \sum_{t=k}^{u} \alpha_t$, with $\alpha_t = \exp\left(0.3\left(\frac{t}{\tau} - 1\right)\right)$. We set τ to 30 days. The chosen exponential time weighting scheme distributes half of the weight to the last 7 days in the 30-day period, therefore ensuring that peaks are not averaged away. To omit any time dynamic, it is sufficient to set $b_t = 1$.

The non-weighted frequency measures for time u are computed as $\boldsymbol{b}'_{[k,u]}\boldsymbol{A}_{[k,u]}$, where $\boldsymbol{A}_{[k,u]}$ is a $(u-k+1)\times 4$ weights matrix defined as:

$$\boldsymbol{A}_{[k,u]} = \left[\boldsymbol{f}_{[k,u]}\ \boldsymbol{p}_{[k,u]}\ \boldsymbol{d}_{[k,u]}\ \boldsymbol{n}_{[k,u]}\right]. \tag{4}$$

The frequency-based time series indicators are all stored into a $U \times 11$ matrix.

The computation of the (weighted) sentiment values follows the same logic as described and results in a $U \times 8$ matrix. The final indices combined are in a $U \times 19$ matrix $\boldsymbol{I}_{e,d}$. We do this for the 3 ESG dimensions, for a total of 57 unique text-based sustainability indicators, for each of the 291 companies.

3.2.4 Validation

We first present a couple of sustainability crisis cases and how they are reflected in our indicators relative to the scores from Sustainalytics. Figure 3 shows the evolution of the indicators for the selected cases.

Figure 3a displays Lonmin, a British producer of metals active in South Africa, whose mine workers and security were at the center of strikes mid-August 2012 leading to unfortunate killings. This is a clear example of a news-driven sustainability downgrade. It was picked up by our constructed news indicators, in that news coverage went up and news sentiment went down, and later reflected in a severe downgrade by Sustainalytics in their social score. Similar patterns are visible for the Volkswagen Dieselgate case (Fig. 3b), for the Libor manipulation scandal (Fig. 3c, which besides Barclays, also other financial institutions are impacted), and for a corruption lawsuit at Finmeccanica (Fig. 3d).

The main conclusions are the following. First, not all Sustainalytics downgrades (or sustainability changes in general) are covered in the press. Second, our indicators pick up severe cases faster, avoiding the lag of a few weeks or longer before adjustments in Sustainalytics scores are observed. The fact that media analysis does not pick up all events, but when it does, it does so fast(er), is a clear argument in favor of combining news-based ESG data with traditional ESG data.

In these illustrations, the general pattern is that the peak starts to wear out before the change in Sustainalytics score is published. Smoother time scaling would result in peaks occurring later, sometimes after the Sustainalytics reporting date, as well as phasing out slower (i.e., more persistence). This is because the news reporting is often clustered and spread out over several days. Likewise, an analysis run without the strict relevance filtering revealed less obvious peaks. Therefore, for (abnormal) peak detection, we recommend short-term focused time weighting and strict filtering.

In addition to the qualitative validation of the indicators, we present one possible way to quantitatively measure their ability to send early warnings for further investigation. We perform an ex-post analysis. Early warnings coming from the news-based indicators are defined as follows. We first split the period prior to a downward re-evaluation by Sustainalytics (a drop larger than 5) into two blocks

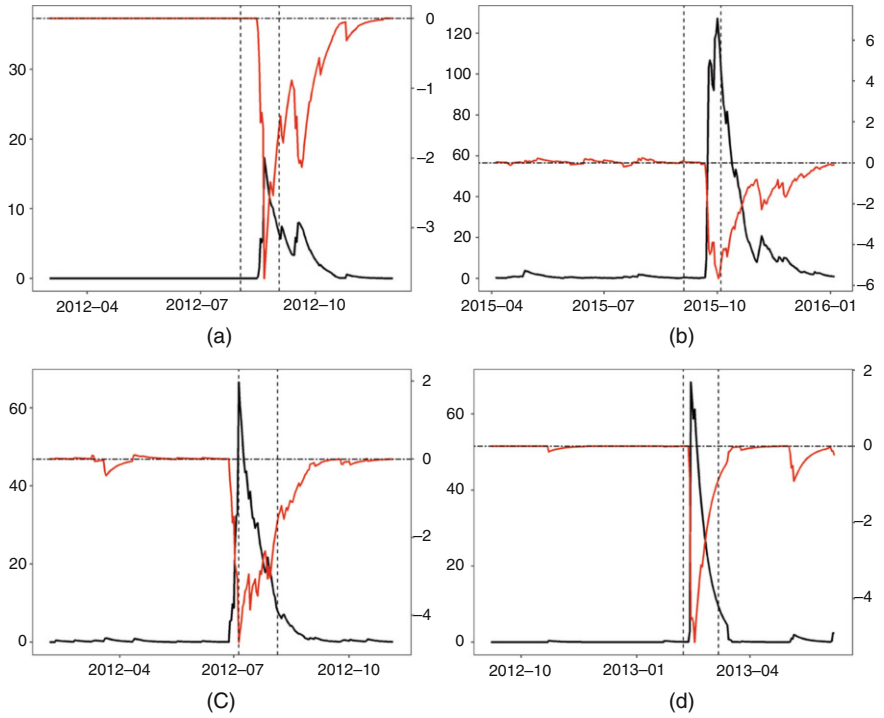

Fig. 3 News-based indicators around a selection of severe Sustainalytics downgrades (a drop larger than 5 on their 0–100 scale). The vertical bars indicate the release date of the downgraded score and 1 month before. The time frame shown is 6 months prior and 3 months after the release date. In black the average of the 11 frequency-based indicators (left axis) and in red of the 8 sentiment-based measures (right axis, with a horizontal line through zero). (**a**) Lonmin (Social). (**b**) Volkswagen (ESG). (**c**) Barclays (Governance). (**d**) Finmeccanica (Governance)

of 3 months. The first 3-month block is the reference period. The indicator values in the second 3-month block are continuously benchmarked against an extreme outcome of the previous block. For the frequency-based indicators, a hypothetical early warning signal is sent when the indicator surpasses the 99% quantile of the daily values in the reference block. For the sentiment-based indicators, a signal is sent if the indicator dips below the 1% reference quantile. Less signals will be passed on if the cut-offs are more extreme, but they will more likely be relevant.

Table 2 displays the results of the analysis for the averaged frequency-based and sentiment-based indicators. Between 11% and 34% of downgrades correspond with more abnormal news dynamics as defined. When so, on average about 50 days ahead of a realized downgrade, an initial news-based early warning is sent. Note that these early warnings should be interpreted as reasonable *first* signals, not necessarily the optimal ones, nor the only ones. There is ample room to fine-tune these metrics, and especially the amplitude of the signals generated in line with investment needs, as hinted to in Sect. 2.2.

Table 2 Ex-post early warning ability of news-based indicators

		Detected		Time gain (days)	
	Events	*f*	*s*	*f*	*s*
E	53%	19%	11%	48	48
S	53%	34%	24%	52	52
G	63%	25%	19%	51	46
ESG	24%	28%	18%	52	47

This table shows ex-post early warning performance statistics. The "events" column is the proportion of the 291 companies analyzed that faced at least one substantial Sustainalytics downgrade released at a day t_D. The "detected" column is the proportion of downgrades for which minimum one early warning was generated within 3 months before t_D. The "time gain (days)" column is the average number of days the first early warning precedes t_D. The analysis is done for the average of the 11 frequency-based indicators (*f*) and of the 8 sentiment-based measures (*s*)

3.3 Stock and Sector Screening

Another test of the usefulness of the created indices is to input them in a sustainable portfolio construction strategy. This allows studying the information content of the indices in general, of the different types of indices (mainly frequency-based against sentiment-based), and of the three ESG dimensions. The analysis should be conceived as a way to gauge the value of using textual data science to complement standard ESG data, not as a case in favor of ESG investing in itself.

We run a small horse race between three straightforward monthly screening strategies. The investable universe consists of the 291 analyzed companies. The strategies employed are the following:

- Invest in the 100 top-performing companies. [S1]
- Invest in the companies excluding the 100 worst-performing ones. [S2]
- Invest in the companies in the 10 top-performing sectors. [S3]

All strategies equally weight the monthly rebalanced selection of companies. We include 24 sectors formed by combining the over 40 peer groups defined in the Sustainalytics dataset. The notion of top-performing companies (resp. worst-performing) means having, at rebalancing date, the lowest (resp. the highest) news coverage or the most positive (resp. the most negative) news sentiment. The strategies are run with the indicators individually for each ESG dimension. To benchmark, we run the strategies using the scores from Sustainalytics and also compare with a portfolio equally invested in the total universe.

We take the screening one step further by imposing for all three strategies that companies should perform among the best both according to the news-based indicators and according to the ratings from Sustainalytics. We slightly modify the strategies per approach to avoid retaining a too limited group of companies; strategy S1 looks at the 150 top-performing companies, strategy S2 excludes the 50 worst-performing companies, and strategy S3 picks the 15 top-performing sectors. The

total investment portfolio consists of the intersection of the selected companies by the two approaches.

We split the screening exercise in two out-of-sample time periods. The first period covers February 1999 to December 2009 (131 months), and the second period covers January 2010 to August 2018 (104 months). The rebalancing dates are at every end of the month and range from January 1999 to July 2018.[6] To screen based on our news-based indicators, we take the daily value at rebalancing date. For the Sustainalytics strategy, we take the most recently available monthly score, typically dating from 2 to 3 weeks earlier.

An important remark is that to estimate the word embeddings, we use a dataset whose range (i.e., January 1996–November 2019) is greater than that of the portfolio analysis. This poses a threat of lookahead bias—meaning, at a given point in time, we will have effectively already considered news data beyond that time point. This would be no problem if news reporting style is fixed over time, yet word use in news and thus its relationships in a high-dimensional vector space are subject to change.[7]It would be more correct (but also more compute intensive) to update the word embeddings rolling forward through time, for example, once a year. The advantage of a large dataset is an improved overall grasp of the word-to-word semantic relationships. Assuming the style changes are minor, and given the wide scope of our dataset, the impact on the outcome of the analysis is expected to be small.

3.3.1 Aggregate Portfolio Performance Analysis

We analyze the strategies through aggregate comparisons.[8] The results are summarized in Table 3. We draw several conclusions.

First, in both subsamples, we notice a comparable or better performance for the S2 and S3 investment strategies versus the equally weighted portfolio. The sector screening procedure seems especially effective. Similarly, we find that our news indicators, both the news coverage and the sentiment ones, are a more valuable screening tool, in terms of annualized Sharpe ratio, than using Sustainalytics scores. The approach of combining the news-based signals with the Sustainalytics ratings leads for strategies S1 and S2 to better outcomes compared to relying on the Sustainalytics ratings only. Most of the Sharpe ratios across ESG dimensions for the combination approach are close to the unscreened portfolio Sharpe ratio. The worst-

[6]Within this first period, the effective corpus size is 87611 articles. Within the second period, it is 60,977 articles. The two periods have a similar monthly average number of articles.

[7]An interesting example is *The Guardian* who declared in May 2019 to start using more often "climate emergency" or "climate crisis" instead of "climate change."

[8]As a general remark, due to the uncertainty in the expected return estimation, the impact of any sustainability filter on the portfolio performance (e.g., the slope of the linear function; Boudt et al. [8] derive to characterize the relationship between a sustainability constraint and the return of mean-tracking error efficient portfolios) is hard to evaluate accurately.

Table 3 Sustainable portfolio screening (across strategies)

(a) News engine

		E		S		G		ESG	
		f	*s*	*f*	*s*	*f*	*s*	*f*	*s*
P1	S1	0.50	0.46	**0.59**	0.40	**0.57**	0.50	**0.55**	0.46
	S2	**0.53**	0.49	**0.61**	0.47	**0.60**	**0.54**	**0.58**	0.50
	S3	**0.65**	0.46	**0.76**	0.44	**0.62**	**0.59**	**0.69**	0.50
P2	S1	0.88	0.81	0.93	0.85	0.91	0.86	0.91	0.84
	S2	**1.03**	0.99	0.99	**1.02**	**1.04**	**1.03**	**1.02**	**1.01**
	S3	**1.11**	**1.02**	**1.02**	0.98	0.99	**1.17**	**1.06**	**1.08**
All	S1	0.64	0.59	**0.72**	0.56	**0.70**	0.63	0.69	0.60
	S2	**0.71**	0.68	**0.76**	0.67	**0.77**	**0.73**	**0.75**	0.69
	S3	**0.82**	0.66	**0.86**	0.64	**0.76**	**0.81**	**0.82**	**0.71**

(b) Sustainalytics

		E	S	G	ESG
P2	S1	0.81	0.82	0.98	0.88
	S2	0.92	0.91	0.98	0.94
	S3	**1.07**	0.89	0.98	**1.00**

(c) News engine + Sustainalytics

		E		S		G		ESG	
		f	*s*	*f*	*s*	*f*	*s*	*f*	*s*
P2	S1	0.93	0.86	0.84	0.79	**1.09**	**1.08**	0.96	0.92
	S2	0.97	0.94	0.99	0.98	0.99	0.98	0.99	0.97
	S3	0.41	0.93	0.33	**1.03**	0.46	0.85	0.41	0.98

Table 3a shows the annualized Sharpe ratios for all strategies (S1–S3), averaged across the strategies on the 11 frequency-based indicators (*f*) and on the 8 sentiment-based indicators (*s*). The ESG column invests equally in the related E, S, and G portfolios. Table 3b shows the Sharpe ratios for all strategies using Sustainalytics scores. Table 3c refers to the strategies based on the combination of both signals. P1 designates the first out-of-sample period (February 1999 to December 2009), P2 the second out-of-sample period (January 2010 to August 2018), and All the entire out-of-sample period. An equally weighted benchmark portfolio consisting of all 291 assets obtains a Sharpe ratio of 0.52 (annualized return of 8.4%), of 1.00 (annualized return of 12.4%), and of 0.70 (annualized return of 10.1%) over P1, P2, and All, respectively. The screening approaches performing at least as good as the unscreened portfolio are indicated in bold

in-class exclusion screening (strategy S2) performs better than the best-in-class inclusion screening (strategy S1), of which only a part is explained by diversification benefits.

There seems to be no performance loss when applying news-based sustainability screening. It is encouraging to find that the portfolios based on simple universe screening procedures contingent on news analysis are competitive with

an unscreened portfolio and with screenings based on ratings from a reputed data provider.

Second, the indicators adjusted for sentiment are not particularly more informative than the frequency-based indicators. On the contrary, in the first subsample, the news coverage indicators result in higher Sharpe ratios. Not being covered (extensively) in the news is thus a valid screening criterion. In general, however, there is little variability in the composed portfolios across the news-based indicators, as many included companies simply do not appear in the news, and thus the differently weighted indices are the same.

Third, news has in both time periods satisfactory relative value. The Sharpe ratios are low in the first subsample due to the presence of the global financial crisis. The good performance in the second subperiod confirms the universally growing importance and value of sustainability screening. It is also consistent with the study of Drei et al. [10], who find that, between 2014 and 2019, ESG investing in Europe led to outperformance.

Fourth, the utility of each dimension is not uniform across time or screening approach. In the first subperiod, the social dimension is best. In the second period, the governance dimension seems most investment worthy, but closely followed by the other dimensions. Drei et al. [10] observe an increased relevance of the environmental and social dimensions since 2016, whereas the governance dimension has been the most rewarding driver overall [5]. An average across the three dimension-specific portfolios also performs well, but not better.

The conclusions stay intact when looking at the entire out-of-sample period, which covers almost 20 years.

3.3.2 Additional Analysis

We also assess the value of the different weighting schemes. Table 4 shows the results for strategy S3 across the 8 sentiment indices, in the second period. It illustrates that the performance discrepancy between various weighting schemes for the sentiment indicators is not clear-cut. More complex weighting schemes, in this application, do not clearly beat the simpler weighting schemes.

Table 4 Sustainable portfolio screening (across sentiment indicators)

	s1	*s2*	*s3*	*s4*	*s5*	*s6*	*s7*	*s8*
E	1.01	1.04	0.99	0.99	1.04	1.02	1.04	1.02
S	0.98	0.99	0.98	1.00	0.98	0.95	0.95	1.01
G	1.14	1.17	1.20	1.20	1.16	1.19	1.09	1.17
ESG	1.06	1.09	1.08	1.08	1.08	1.08	1.05	1.08

This table shows the annualized Sharpe ratios in P2 for the screening strategy S3, built on the sentiment-based indicators, being *s1* and *s2*–*s8* as defined through the weighting matrix in (3)

An alternative approach for the strategies on the frequency-based indicators is to invert the ranking logic, so that companies with a high news coverage benefit and low or no news coverage are penalized. We run this analysis but find that the results worsen markedly, indicating that attention in the news around sustainability topics is not a good screening metric.

To test the sensitivity to the strict filtering choice of leaving out articles not having at least three keywords and more than half of all keywords related to one dimension, we rerun the analysis keeping those articles in. Surprisingly, some strategies improve slightly, but not all. We did not examine other filtering choices.

We also tested a long/short strategy but the results were poor. The long leg performed better than the short leg, as expected, but there was no reversal effect for the worst-performing stocks.

Other time lag structures (different values for τ or different functions in $\boldsymbol{B}$) are not tested, given this would make the analysis more a concern of market timing than of assessing the lag structure. A short-term indicator catches changes earlier, but they may have already worn out by the rebalancing date, whereas long-term indicators might still be around peak level or not yet. We believe fine-tuning the time lag structure is more crucial for peak detection and visualization.

4 Conclusion

This chapter presents a methodology to create frequency-based and sentiment-based indicators to monitor news about the given topics and entities. We apply the methodology to extract company-specific news indicators relevant to environmental, social, and governance matters. These indicators can be used to timely detect abnormal dynamics in the ESG performance of companies, as an input in risk management and investment screening processes. They are not calibrated to automatically make investment decisions. Rather, the indicators should be seen as an additional source of information to the asset manager or other decision makers.

We find that the indicators often anticipate substantial negative changes in the scores of the external ESG research provider Sustainalytics. Moreover, we also find that the news indices can be used as a sole input to screen a universe of stocks and construct simple but well-performing investment portfolios. In light of the active sustainable investment manager being an "ESG ratings aggregator," we show that combining the news signals with the scores from Sustainalytics leads to a portfolio selection that performs equally well as the entire universe.

Given the limited reach of our data (we use Flemish and Dutch news to cover a wide number of European stocks), better results are expected with geographically more representative news data as well as a larger universe of stocks. Hence, the information potential is promising. It would be useful to investigate the benefits local news data bring for monitoring companies with strong local ties.

Additional value to explore lies in more meaningful text selection and index weighting. Furthermore, it would be of interest to study the impact of more

fine-grained sentiment calculation methods. Summarization techniques and topic modeling are interesting text mining tools to obtain a drill down of sustainability subjects or for automatic peak labeling.

Acknowledgments We are grateful to the book editors (Sergio Consoli, Diego Reforgiato Recupero, and Michaela Saisana) and three anonymous referees, seminar participants at the CFE (London, 2019) conference, Andres Algaba, David Ardia, Keven Bluteau, Maxime De Bruyn, Tim Kroencke, Marie Lambert, Steven Vanduffel, Jeroen Van Pelt, Tim Verdonck, and the Degroof Petercam Asset Management division for stimulating discussions and helpful feedback. Many thanks to Sustainalytics (https://www.sustainalytics.com) for providing us with their historical dataset, and to Belga for giving us access to their news archive. This project received financial support from Innoviris, swissuniversities (https://www.swissuniversities.ch), and the Swiss National Science Foundation (http://www.snf.ch, grant #179281).

References

1. Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys, 34*(3), 512–547. https://doi.org/10.1111/joes.12370
2. Amel-Zadeh, A., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal, 74*(3), 87–103. https://doi.org/10.2469/faj.v74.n3.2
3. Antweiler, W., & Frank, M. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance, 59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x
4. Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). The R package **sentometrics** to compute, aggregate and predict with textual sentiment. *Forthcoming in Journal of Statistical Software*. https://doi.org/10.2139/ssrn.3067734
5. Bennani, L., Le Guenedal, T., Lepetit, F., Ly, L., & Mortier, V. (2018). *The alpha and beta of ESG investing*. Amundi working paper 76. http://research-center.amundi.com
6. Berg, F., Koelbel, J., & Rigobon, R. (2019). *Aggregate confusion: The divergence of ESG ratings*. MIT Sloan School working paper 5822–19. https://doi.org/10.2139/ssrn.3438533
7. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051
8. Boudt, K., Cornelissen, J., & Croux, C. (2013). The impact of a sustainability constraint on the mean–tracking error efficient frontier. *Economics Letters, 119*, 255–260. https://doi.org/10.1016/j.econlet.2013.03.020
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre–training of deep bidirectional transformers for language understanding*. Working paper, arXiv:1810.04805, https://arxiv.org/abs/1810.04805v2
10. Drei, A., Le Guenedal, T., Lepetit, F., Mortier, V., Roncalli, T., & Sekine, T. (2019). *ESG investing in recent years: New insights from old challenges*. Amundi discussion paper 42. http://research-center.amundi.com
11. Engle, R., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. *Review of Financial Studies, 33*(3), 1184–1216. https://doi.org/10.1093/rfs/hhz072
12. Escrig-Olmedo, E., Muñoz-Torres, M. J., & Fernandez-Izquierdo, M. A. (2010). Socially responsible investing: Sustainability indices, ESG rating and information provider agencies. *International Journal of Sustainable Economy, 2*, 442–461.

13. Heston, S., & Sinha, N. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal, 73*(3), 67–83. https://doi.org/10.2469/faj.v73.n3.3
14. Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics, 110*, 712–729. https://doi.org/10.1016/j.jfineco.2013.08.018
15. Krijthe, J., van der Maaten, L. (2018). *Rtsne: T-distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation*. R Package Version 0.15. https://CRAN.R-project.org/package=Rtsne
16. Leskovec, J., Rajaraman, A., & Ullman, J. (2014). *Mining of massive datasets*. Chapter Finding Similar Items (pp. 72–134). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139924801
17. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10–Ks. *Journal of Finance, 66*, 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 3111–3119). http://dl.acm.org/citation.cfm?id=2999792.2999959
19. Moniz, A. (2016). *Inferring the financial materiality of corporate social responsibility news*. Working paper, SSRN 2761905. https://doi.org/10.2139/ssrn.2761905
20. Mullen, L. (2016). *textreuse: Detect Text Reuse and Document Similarity*. R Package Version 0.1.4. https://CRAN.R-project.org/package=textreuse
21. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). New York, NY, USA: ACM. https://doi.org/10.3115/v1/D14-1162
22. Riedl, A., & Smeets, P. (2017). Why do investors hold socially responsible mutual funds? *Journal of Finance, 72*(6), 2505–2550. https://doi.org/10.1111/jofi.12547
23. Selivanov, D., & Wang, Q. (2018). *text2vec: Modern Text Mining Framework for R*. R Package Version 0.5.1. https://CRAN.R-project.org/package=text2vec
24. Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance, 63*(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x
25. Theil, C. K., Štajner, S., & Stuckenschmidt, H. (2018). Word embeddings–based uncertainty detection in financial disclosures, In *Proceedings of the First Workshop on Economics and Natural Language Processing* (pp. 32–37). https://doi.org/10.18653/v1/W18-3104