Using the GDELT Dataset to Analyse the Italian Sovereign Bond Market

Sergio Consoli^(✉), Luca Tiozzo Pezzoli, and Elisa Tosetti

European Commission, Joint Research Centre, Directorate A-Strategy, Work Programme and Resources, Scientific Development Unit, Via E. Fermi 2749, 21027 Ispira, VA, Italy
`sergio.consoli@ec.europa.eu`

Abstract. The Global Data on Events, Location, and Tone (GDELT) is a real time large scale database of global human society for open research which monitors worlds broadcast, print, and web news, creating a free open platform for computing on the entire world's media. In this work, we first describe a data crawler, which collects metadata of the GDELT database in real-time and stores them in a big data management system based on Elasticsearch, a popular and efficient search engine relying on the Lucene library. Then, by exploiting and engineering the detailed information of each news encoded in GDELT, we build indicators capturing investor's emotions which are useful to analyse the sovereign bond market in Italy. By using regression analysis and by exploiting the power of Gradient Boosting models from machine learning, we find that the features extracted from GDELT improve the forecast of country government yield spread, relative that of a baseline regression where only conventional regressors are included. The improvement in the fitting is particularly relevant during the period government crisis in May-December 2018.

Keywords: Government yield spread · Machine learning · Big data management · Quantile regression · Feature Engineering · GDELT

1 Introduction

The explosion in computation and information technology experienced in the past decade has made available vast amounts of data in various domains, that has been referred to as *Big Data*. In Economics and Finance in particular, tapping into these data brings research and business closer together, as data generated in ordinary economic activity can be used towards rapid-learning economic systems, continuously improving and personalizing models. In this context, the recent use of Data Science technologies for Economics and Finance is providing mutual benefits to both scientists and professionals, improving forecasting and nowcasting for several types of applications.

In particular, the recent surge in the government yield spreads in countries within the Euro area has originated an intense debate about the determinants and sources of risk of sovereign spreads. Traditionally, factors such as the credit-worthiness, the sovereign bond liquidity risk, and global risk aversion have been identified as the main factors having an impact on government yield spreads [2, 20]. However, a recent literature has pointed at the important role of financial investor's sentiment in anticipating interest rates dynamics [17, 25].

This paper exploits a novel, open source, news database known as *Global Database of Events, Language and Tone (GDELT)*¹ [15] to construct news-based financial indicators related to economic and political events for a set of Euro area countries. As described in Sect. 3.1, since the dimensions of the GDELT dataset make unfeasible the use of any relational database to perform an analysis in reasonable time, in Sect. 4.1 it is discussed the big data management infrastructure that we have used to host and interact with the data. Once GDELT data are crawled from the Web by the means of custom REST APIs², we efficiently transform and store them on our big data management system based on Elasticsearch, a popular and efficient NO-SQL search engine (Sect. 4.1).

Afterwards a feature engineering process is applied on the detailed information encoded in GDELT (Sect. 4.2) to select the most profitable variables which capture, among others, investor's emotions and popularity of news thematics, and that are useful to analyse the Italian sovereign bond market. In Sect. 4.3 we describe the Gradient Boosting machine we have used to analyse the Italian sovereign bond market. Our experimental analysis reported in Sect. 4.4 shows that the implemented machine learning model using the constructed GDELT indicators is useful to predict country government yield spread and financial instability, well aligned with previous studies in the literature.

2 Related Work

News articles represent a recent addition to the standard information used to model economic and financial variables. An early paper is [25] that uses sentiment from a column in the Wall Street Journal to show that high levels of pessimism are a relevant predictor of convergence of the stock prices towards their fundamental values. Following this early work, several other papers have tried to understand the role that news play in predicting, for instance, company news announcements, stock returns and volatility. For example recent works in finance exist on the application of semantic sentiment analysis from social media, financial microblogs, and news to improve predictions of the stock market (e.g. [1, 7]). However these approaches generally suffer from a limited scope of the historical financial sources available. Recently, news have been also used in macroeconomics. For example, [13] looks at the informational content of the Federal Reserve statements and the guidance that these statements provide about the future evolution of monetary policy. Other papers ([26, 27] and [23] among

¹ GDELT website: <https://blog.gdeltproject.org/>.

² See <https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>.

others) use Latent Dirichlet allocation (LDA) to classify articles in topics and calculate simple measures of sentiment based on the topic classification. The goal of these papers is to extract a signal that could have some predictive content for measures of economic activity, such as GDP, unemployment and inflation [11]. Their results show that economic sentiment is a useful addition to the predictors that are commonly used to monitor and forecast the business cycle [7].

Machine learning approaches in the existing literature for controlling financial indexes measuring credit risk, liquidity risk and risk aversion include the works in [2, 4, 8, 9, 18], among others. Efforts to make machine learning models accepted within the economic modeling space have increased exponentially in recent years [19, 24]. Among popular machine learning approaches, Gradient Boosting machines have been shown to be successful in various forecasting problems in Economics and Finance (see e.g. [5, 6, 16, 28] among others).

3 Data

3.1 About GDELT

GDELT is the global database of events, location and tone that is maintained by Google [15]. It is an open Big Data platform on news collected at worldwide level, containing structured data mined from broadcast, print and web news sources in more than 65 languages. It connects people, organizations, quotes, locations, themes, and emotions associated with events happening across the world. It describes societal behavior through eye of the media, making it an ideal data source for measuring social factors and for testing our hypotheses. In terms of volume, GDELT analyses over 88 million articles a year and more than 150,000 news outlets. Its dimension is around 8 TB, growing 2TB each year. GDELT consists of two main datasets, the “Global Knowledge Graph (GKG)” and the “Events Table”. For our study we have relied on the first dataset. GDELT’s GKG captures what’s happening around the world, what its context is, who’s involved, and how the world is feeling about it; every single day. It provides English translation from the supported languages of the encoded information. In addition, the included themes are mapped into commonly used practitioners’ topical taxonomies, such as the “World Bank (WB) Topical Ontology”³, or into the GDELT built-in topical taxonomy. GDELT also measures thousands of emotional dimensions expressed by means of popular dictionaries in the literature, such as the “Harvard IV-4 Psychosocial Dictionary”⁴, the “WordNet-Affect dictionary”⁵, and the “Loughran and McDonald Sentiment Word Lists dictionary”⁶, among others. For this application we use the GDELT GKG fields from

³ <https://vocabulary.worldbank.org/taxonomy.html>.

⁴ Harvard IV-4 Psychosocial Dictionary: <http://www.wjh.harvard.edu/~inquirer/homecat.htm>.

⁵ WordNet-Affect dictionary: <http://wndomains.fbk.eu/wnaffect.html>.

⁶ Loughran and McDonald Sentiment Word Lists: <https://sraf.nd.edu/textual-analysis/resources/>.

the World Bank Topical Ontology (i.e. WB themes), all emotional dimensions (GCAM), and the name of the journal outlet.

3.2 Yield Spread

We have extracted data from Bloomberg on the term-structure of government bond yields for Italy over the period 2 March 2015 to 31 August 2019. We calculate the sovereign spread for Italy against Germany as the difference between the Italian 10 year maturity bond yield minus the German counterpart. We also extract the standard level, slope and curvature factors of the term-structure using the Nelson and Siegel [21] procedure.

We estimate a model of credit spread forecast using conventional yield curve factors and the GDELT selected features, and compare with a classical model of credit spread forecast using only the level, slope and curvature as regressors.

4 Methods

4.1 Big Data Management

Massive unstructured datasets like GDELT require to be stored in specialized distributed file systems (DFS), joining together many computational nodes over a network, that are essential for building the data pipes that slice and aggregate this large amount of information. Among the most popular DFS platforms today, considering the huge number of unstructured documents coming from GDELT, we have used Elasticsearch [12,22] to store the data and interact with them. Elasticsearch is a popular and efficient document-store which, instead of storing information as rows of columnar data like in classical relational databases, stores complex data structures that have been serialized as JSON documents. Being built on the Apache Lucene search library⁷, it then provides real-time search and analytics for different types of structured or unstructured data.

Elasticsearch is built upon a distributed setting, allowing connecting multiple Elasticsearch nodes in a unique cluster. At the moment a document is stored, it is indexed and fully searchable in near real-time. An Elasticsearch index can be thought of as an optimized collection of documents and each document is a collection of fields, which are the key-value pairs that contain the stored data. An index is really just a logical grouping of one or more physical shards, where each shard is actually a self-contained index.

Elasticsearch is also schema-less, which means that documents can be indexed without explicitly specifying how to handle each of the different fields that might occur in a document. Elasticsearch provides a simple REST API for managing the created cluster and interacting with the stored documents. It is possible to submit API requests directly from the command line or through the Developer Console within the user web interface, referred to as *Kibana*⁸. The Elasticsearch

⁷ <https://lucene.apache.org/>.

⁸ Kibana, version 7.4: <https://www.elastic.co/guide/en/kibana/7.4/>.

REST APIs support structured queries, full text queries, and complex queries that combine the two, using Elasticsearch's JSON-style query language, referred to as *Query DSL*⁹.

4.2 Feature Engineering

As GDELT adds news articles every fifteen minutes, each article is concisely represented as a row of a GDELT table in .csv format. The GKG table contains around 10 TB of data that need to be integrated and ingested as serialized JSON documents into our Elasticsearch framework. This involves applying the three usual steps of *Extract, Transform* and *Load* (ETL), that have to identify and overcome structural, syntactic, and semantic heterogeneity across the data. For this reason we have used the available World Bank Topical Ontology to understand the primary focus (theme) of each article and select the relevant news whose main themes are related to events concerning bond market investors. Such taxonomy is a classification schema for describing the World Bank's areas of expertise and knowledge domains representing the language used by domain experts. GDELT contains all themes discussed in an article as an entry in a single row. We separate these themes into separate entries.

We have extracted news information from GKG from a set of around 20 newspapers for Italy, published over the period March 2015 until end of August 2019. We rely on the Geographic Source Lookup file available on GDELT blog¹⁰ in order to chose both generalist national newspapers with the widest circulation in that country, as well as specialized financial and economic outlets. Once collected the news data, we have mapped these to the relevant trading day. Specifically, we assign to a given trading day all the articles published during the opening hours of the bond market, namely between 9.00 am and 5.30 pm. Articles that have been published after the closure of the bond market or overnight are assigned to the following trading day.¹¹ Following [10], we assign the news published during weekends to Monday trading days, and omit articles published during holidays or in weekends preceding holidays.

Hence, we selected only articles such that the topics extracted by GDELT fall into one of the following WB themes of interest: *Macroeconomic Vulnerability and Debt*, and *Macroeconomic and Structural Policies*. We observe that articles can mention only briefly one of the selected topics and then focus on a totally different theme. To make sure that the main focus of the article is one of the selected WB topics, we have retained only news that contain in their text at least three keywords belonging to these themes. The aim is to select news that focus on topics relevant to the bond market, while excluding news that only briefly

⁹ <https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html>.

¹⁰ See <https://blog.gdeltproject.org/mapping-the-media-a-geographic-lookup-pf-gdelt-sources/>.

¹¹ Since the GKG operates on the UTC time, <https://blog.gdeltproject.org/new-gkg-2-0-article-metadata-fields/>, we made a one-hour lag adjustment according to Italian time zone.

report macroeconomic, debt and structural policies issues. Finally, to obtain a pool of news that are not too heterogeneous in length, we have retained only articles that are at least 100 words long. After this selection procedure we obtain a total of 18,986 articles. From this large amount of information, we construct features counting the total number of words belonging to all WB themes and GCAMs detected each day. We also created the variables “Number of mentions” denoting the word count of each location mentioned in the selected news. By doing this we obtain a total of 2,978 GCAM, 1,996 Themes and 155 locations. Notice that, all of our features are expressed in terms of daily word count with the exception of the ANEW dictionary (v19) and the Hedenometer measure of happiness (v21) which are already provided as score values.

Once extracted the data from GKG, we have adopted a five step procedure to filter out features from news stories. In the first step we have applied a domain knowledge criteria and have retained a subset of 413 GCAM dictionaries that are potentially relevant for our analysis. Specifically we have extracted 31 dimensions of the General Inquirer Harvard IV psychosocial Dictionary, 61 dimensions of Roget’s Thesaurus, 7 dimensions of the Martindale Regressive Imagery and 3 dimensions of the Affective Norms for English Words (ANEW) dictionary. The second step concerns the variability of the extracted features. In particular, we have retained variables with a standard deviation calculated over the full sample that is greater than 5 words and allowed a 10% of missing values on the total number of days. In addition, features that are missing at the beginning of the sample (more than 33% of the sample) have been excluded. In the final step we have performed a correlation analysis across the selected variables. We have normalized at first all features by number of daily articles. If the correlation between any two features is above 80% we give preference to the variable with less missing values, while if the number of missing values is identical and the two variables belong to the same category (i.e. both are themes or GCAMs), we randomly pick one of them. Finally, if the number of missing values is identical but the two variables belong to the same category, we consider the following order of priority: GCAM, WB themes, GDELT themes, locations. After this Feature Engineering procedure, we are left with a total of 45 variables, of which 9 are themes, 34 are GCAM, 2 are locations. A careful inspection among the selected topics reveals that WB themes such as Inflation, Government, Central Banks, Taxation and Policy have been selected by our procedure. These are important topics discussed in the news when considering interest rates issues. Moreover, features constructed and selected from GCAM dimensions such as optimism, pessimism or arousal are also inculed and allow us to explore the emotional state of the market.

4.3 Big Data Analytics

Classical economic models are not dynamically scalable to manage and maintain Big Data structures, like the GDELT one we are dealing with. A whole new set of big data analytics models and tools that are robust in high dimensions, like the ones from machine learning, are required [19,24]. In particular

in our computational study we have chosen to rely on *Gradient Boosting* (GB) [14], a well-known machine-learning approach which has been shown to be successful in various modelling problems in Economics and Finance [5, 6, 16, 28]. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction.

In particular for our implementation we have used the *H2O*¹² library available for the R programming language. H2O is a scalable open-source machine learning platform that offers parallelized implementations of many supervised and unsupervised machine learning algorithms, including Gradient Boosting Machines.

In addition, to determine the optimal parameter values of our GB model, we have used a 10-fold cross-validation together with a grid search (or parameter sweep) procedure [3]. Grid search involves an exhaustive searching through a manually specified subset of the hyperparameter space of the learning algorithm, guided by some performance metric (like in our case minimizing the mean squared error mean). The main parameters to optimize in our GB model are the *maximum tree depth*, which indicates the the maximum possible depth of a tree in the model and is used to control over-fitting, as higher depth will allow model to learn relations very specific to a particular sample; and the *learning rate*, which determines the impact of each tree on the final outcome of the GB model. GB works by starting with an initial estimate which is updated using the output of each tree; the learning rate parameter controls then the magnitude of this change in the estimates. Therefore, lower values are generally preferred as they make the model robust to the specific characteristics of the tree and thus allowing it to generalize well. However lower values would require higher number of trees to model all the relations and will be computationally expensive.

To explore the hyperparameters space looking for optimal values of these parameters, the grid search procedure tests the GB model with values going from 1 to 10 (by steps of 1) for the maximum tree depth parameter and from 0.01 to 0.99 (by steps of 0.10) for the learning rate parameter, the best parameter values with respect to the mean squared error are produced as output. In the case that one of the produced parameter values reaches one of the related upper or lower bound, i.e. corner solutions, a greedy approach is iterated. The search boundaries of the specific parameter are perturbed, and the grid search procedure is restarted from the sub-optimal parameters coming from the previous estimation. The procedure halts when both produced parameter values fall inside the related search boundaries, giving these parameter values as output.

Although grid search does not provide an absolute guarantee that it will find the global optimal parameter values, in practice we have found it to work quite

¹² R Interface for the “H2O” Scalable Machine Learning Platform: <https://cran.r-project.org/web/packages/h2o/index.html>.

well, despite to be quite computationally expensive. In general grid searching is a widely adopted and accepted procedure for this kind of tuning tasks [3].

4.4 Experimental Analysis

The main objective of our empirical exercise is to assess the predictive power of GDELT selected features over and above the classical determinants for government credit spreads during stressed periods. We explore predictability for the 90th percentile of the credit spread distribution, since this is usually classified as a situation of financial distress (see [4] among others). Several studies in the literature have shown that during these periods, complex non-linear relationships among explanatory variables affect the behaviour of the output target which simple linear models are not able to capture. We account for that by using a GB model with a quantile distribution and we adopt a rolling window estimation technique where the first estimation sample starts at the beginning of March and ending in May 2017. We compare the forecasting performance of a GB model where classical determinants as well as selected GDELT features are included, with that of a GB model where only classical factors are considered. We measure the forecasting performance by calculating the absolute error for each forecast. We also assess the explanatory power of GDELT features over time by calculating the variable importance at each estimation, and explore the five most important variables at each rolling window estimation. This is in line with standard term structure literature, stating that from three to five factors are sufficient to explain the dynamics of yield spreads. We estimate the model on half of the sample and adopt a rolling window to generate one-step ahead forecasts. We use a 10-fold cross-validation for each estimation window, and apply the grid search procedure previously described in Sect. 4.3 to optimally find the hyperparameters of the GB model.

Figure 1 displays the time series of the Italian spread and the ratio between absolute forecast errors of the model augmented with GDELT features and that of the model with classical regressors only. Notice that, when the value of this ratio is below one, our augmented model performs better than the benchmark. It is interesting to observe that the performance of our augmented model improves considerably starting from the end of May 2018 when a period of political turmoil started in Italy. On the 29th of May, the Italian spread sharply rose reaching 250 basis point. Investors were particularly worried about the possibility of anti-euro government and not confident on the formation of a stable government. During these stressed events our GDELT features augmented model strongly outperforms the benchmark model with absolute ratios values well below one, i.e. 0.93 on May the 24th that is the minimum value across all the sample under analysis. This result emphasises the value added of news articles stories in forecasting the yield spreads dynamics during periods of financial distress. From June until November 2018, a series of discussions about deficit spending engagements and possible conflicts with European fiscal rules continued to worry the markets. The spread strongly increased in October and November with values

Fig. 1. Absolute errors ratios and spread

around 300 basis point. During this period our augmented model agains performs particularly well, with ratio around 0.95 with a minimum value of 0.94 on November the 9th. Since 2019, the italian political situation started to improve and the spread smothly decline, especially after the agreement with Brussels on buget deficit in December 2018. However, some events hit the Italian economy afterwards, such as the EU negative outlook and the European parliament elections which contributed to a temporary increase on interest rates.

Although in 2019 our aumented model did not perform as well as in 2018 in terms of absolute error ratios, it still consistently outperforms the benchmark model. From the analysis above we clearly observe three main sub-periods. The first pre-crisis period ranging form July 2017 till May 2018, the second one is the crisis period from June till December 2018, the third period from January 2019 till the end of the sample. We next analyse the contribution of each selected GDELT feature in the prediction of interest rates spread during periods of political stress, splitting the sample in the three identified subperiods.

Figure 2 shows the frequency of each variables appearing on the top five positions according to the Gradient Boosting Machine variable importance during (a) the pre-crisis period, (b) the crisis period, and (c) the post crisis period. It is interesting to observe that classical factors such as slope and curvature of spread yield curve are the most important variables during the pre-crisis period. However, we also observe that amongst the classical regressors, the level factor, which is pointed by the literature as the most important variable in explaining interest rates dynamics, is less important than two GCAM sentiment measures, namely the arousal form ANEW dictionary and Hate of Thesauris. Figure 2(b) shows that, during crisis period, classical yield spread factors reduces considerably their predictive contribution. The most important variable is the Arousal

(a): Pre-Government crisis

(b): Government crisis

(c): Post-Government crisis

Fig. 2. Variable importance: percentage of times variables are top 5 classified

index of the ANEW dictionary. Interestingly, mentions of German locations and “Government” occupy the second and the third position, respectively. Classical slope factor is only classified as fifth. Finally, Fig. 2(c) shows that in 2019, the importance of the level factor strongly increase with also the mentions of Monetary Policies issues. ANEW Arousal and Government mentions are still on the top five list meaning that sentiment charged discussion about governmental issues were still important during the post-stressed period. It is important to underline that again, only one out of three classical predictors are out of the top-five list as well as in the post-crisis period.

5 Conclusions

Our analysis is one of the first to study the behaviour of government yield spreads and financial portfolio decisions in the presence of classical yield curve factors and financial sentiment measures extracted from news. We believe that these new measures are able to capture and predict changes in interest rates dynamics especially in period of turmoil. We empirically show that the contribution of our sentiment dimensions is substantial in such periods even more than classical interest rates factors. Interestingly, the importance of our measures also extends in post-crisis periods meaning that our features are able to capture charged sentiment narratives about governmental and monetary policy issues that were the focus of stories also after the crisis. Overall, the paper shows how to use a large scale database as GDELT to derive financial indicators in order to capture future intentions of agents in financial bond markets. In future work we will adopt additional machine learning techniques to better exploit non-linear effects on the dependent variables.

References

1. Agrawal, S., Azar, P., Lo, A.W., Singh, T.: Momentum, mean-reversion and social media: evidence from StockTwits and Twitter. *J. Portfolio Manag.* **44**, 85–95 (2018)
2. Beber, A., Brandt, M.W., Kavajecz, K.A.: Flight-to-quality or flight-to-liquidity? Evidence from the Euro-area bond market. *Rev. Finan. Stud.* **22**(3), 925–957 (2009)
3. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**(1), 281–305 (2012)
4. Bernal, O., Gnabo, J.-Y., Guilmin, G.: Economic policy uncertainty and risk spillover in the Eurozone. *J. Int. Money Finan.* **65**(C), 24–451 (2016)
5. Chang, Y.-C., Chang, K.-H., Wu, G.-J.: Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft Comput. J.* **73**, 914–920 (2018)
6. Deng, S., Wang, C., Wang, M., Sun, Z.: A gradient boosting decision tree approach for insider trading identification: an empirical model evaluation of china stock market. *Appl. Soft Comput. J.* **83**, 105652 (2019)

7. Dridi, A., Atzeni, M., Reforgiato Recupero, D.: FineNews: fine-grained semantic sentiment analysis on financial microblogs and news. *Int. J. Mach. Learn. Cybernet.* **10**(8), 2199–2207 (2018). <https://doi.org/10.1007/s13042-018-0805-x>
8. Favero, C., Pagano, M., von Thadden, E.-L.: How does liquidity affect government bond yields? *J. Finan. Quant. Anal.* **45**(1), 107–134 (2010)
9. Garcia, A.J., Gimeno, R.: Flight-to-liquidity flows in the Euro area sovereign debt crisis. Technical report, Banco de Espana Working Papers (2014)
10. Garcia, D.: Sentiment during recessions. *J. Finan.* **68**(3), 1267–1300 (2013)
11. Gentzkow, M., Kelly, B., Taddy, M.: Text as data. *Journal of Economic Literature* (2019, to appear)
12. Gormley, C., Tong, Z.: *Elasticsearch: The definitive guide*. O’ Reilly Media, US (2015)
13. Hansen, S., McMahon, M.: Shocking language: understanding the macroeconomic effects of central bank communication. *J. Int. Econ.* **99**, S114–S133 (2016)
14. Hastie, T., Tibshirani, R., Friedman, J.: Additive models, trees, and related methods. *The Elements of Statistical Learning*. SSS, pp. 295–336. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7_9
15. Leetaru, K., Schrod, P.A.: Gdelt: global data on events, location and tone, 1979–2012. Technical report, KOF Working Papers (2013)
16. Liu, J., Wu, C., Li, Y.: Improving financial distress prediction using financial network-based information and GA-based gradient boosting method. *Comput. Econ.* **53**(2), 851–872 (2019)
17. Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries and 10-ks. *J. Finan.* **66**(1), 35–65 (2011)
18. Manganello, S., Wolswijk, G.: What drives spreads in the Euro area government bond markets? *Econ. Policy* **24**(58), 191–240 (2009)
19. Marwala, T.: *Economic Modeling using Artificial Intelligence Methods*. Springer-Verlag, London (2013). <https://doi.org/10.1007/978-1-4471-5010-7>
20. Monfort, A., Renne, J.-P.: Decomposing Euro-area sovereign spreads: credit and liquidity risks. *Rev. Finan.* **18**(6), 2103–2151 (2013)
21. Nelson, C., Siegel, A.F.: Parsimonious modeling of yield curves. *J. Bus.* **60**(4), 473–489 (1987)
22. Shah, N., Willick, D., Mago, V.: A framework for social media data analytics using elastic search and kibana. *Wireless Networks* (2018, in press)
23. Shapiro, A.H., Sudhof, M., Wilson, D.: Measuring news sentiment. Federal Reserve Bank of San Francisco Working Paper (2018)
24. Taddy, M.: *Business Data Science: Combining Machine Learning and Economics to optimize, automate, and accelerate business decisions*. McGraw-Hill, US (2019)
25. Tetlock, P.C.: Giving content to investor sentiment: the role of media in the stock market. *J. Finan.* **62**(3), 1139–1168 (2007)
26. Thorsrud, L.A.: Nowcasting using news topics. big data versus big bank. Norges Bank Working Paper (2016)
27. Thorsrud, L.A.: Words are the new numbers: a newsy coincident index of the business cycle. *J. Bus. Econ. Stat.* **38**(2), 1–17 (2018)
28. Yang, X., He, J., Lin, H., Zhang, Y.: Boosting exponential gradient strategy for online portfolio selection: an aggregating experts’ advice method. *Comput. Econ.* **55**(1), 231–251 (2020)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.