IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 2, 2021, accepted February 14, 2021, date of publication February 16, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059960

Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting

SALVATORE M. CARTA[1], (Member, IEEE), SERGIO CONSOLI[2], LUCA PIRAS[1], ALESSANDRO SEBASTIAN PODDA[1], AND DIEGO REFORGIATO RECUPERO[1]

[1]Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
[2]European Commission, Joint Research Centre (DG-JRC), I-21027 Ispra, Italy

Corresponding author: Sergio Consoli (sergio.consoli@ec.europa.eu)

We would like to thank the Centre for Advanced Studies at the Joint Research Centre of the European Commission for guidance and support during the development of this research work. This work was also partially supported by the POR FESR 2014-2020 project: "AlmostAnOracle - AI and Big Data Algorithms for Financial Time Series Forecasting."

ABSTRACT In this manuscript, we propose a Machine Learning approach to tackle a binary classification problem whose goal is to predict the magnitude (high or low) of future stock price variations for individual companies of the S&P 500 index. Sets of lexicons are generated from globally published articles with the goal of identifying the most impactful words on the market in a specific time interval and within a certain business sector. A feature engineering process is then performed out of the generated lexicons, and the obtained features are fed to a Decision Tree classifier. The predicted label (high or low) represents the underlying company's stock price variation on the next day, being either higher or lower than a certain threshold. The performance evaluation we have carried out through a walk-forward strategy, and against a set of solid baselines, shows that our approach clearly outperforms the competitors. Moreover, the devised Artificial Intelligence (AI) approach is explainable, in the sense that we analyze the white-box behind the classifier and provide a set of explanations on the obtained results.

INDEX TERMS Stock market forecasting, machine learning, natural language processing, financial technology, explainable artificial intelligence.

I. INTRODUCTION

It has been proved in literature that stock prices of financial markets are heavily affected by exogenous factors, i.e., information such as events reported in the news, social media, etc., [1], [2]. In fact, the *Adaptive Market Hypothesis* [3] observed that excess return in stock market is ascribed to information asymmetry. This had implications on a new vision of the *Efficient Market Hypothesis* [1], which has been reconsidered in light of the behavioural economics. Basically, traders who can access, mine and analyse heterogeneous information will have a competitive advantage. Besides the stock prices information, there are several other information freely available today like newspapers, social media, etc., that can be used to improve the forecasting or classification problems existing within the financial domain [4]. What results to be still challenging is to take advantage of all this heterogeneous information and creating useful indicators or lexical resources, specifically targeted for the economic and financial domains and able to improve existing forecasting systems.

Therefore, Information Extraction, Text Mining and other techniques within the Natural Language Processing (NLP) sphere, have been leveraged by researchers with the goal of improving their stock price forecasting systems. The first approaches employed text representation models such as bag-of-words combined with simple statistical measures [5]. Then with the advancements in hardware (e.g. GPUs) and middlewares (e.g. CUDA[1]), scientists started using new methods based on evolved Machine Learning approaches (e.g. Deep Learning [6]–[9]). This led to an increasing trend of publications within the NLP-based financial forecasting domain. To further increase this trend, in 2010, social media platforms such as Twitter[2] and StockTwits[3] started generating a huge

The associate editor coordinating the review of this manuscript and approving it for publication was Bohui Wang.

[1]https://developer.nvidia.com/cuda-toolkit
[2]http://www.twitter.com
[3]http://www.stocktwits.com

amount of textual content that was employed within current Machine Learning methods to improve existing classification techniques [10]. Other domains benefited from these social media platforms as well: for example, within the Sentiment Analysis domain, one recurrent task has been to automatically extract moods and opinions to analyse their impact on the market [11], [12]. One more consideration is related to price stocks variations, known as the magnitude of the differences between the highest and lowest values on a given trading day. Daily price variation is a measure of volatility, or how much a stock's value changes. In finance, volatility is the degree of variation of a trading price series over time, usually measured by the standard deviation of logarithmic returns.[4] Recent studies [13] show strong empirical performance for volatility-managed version of popular trading strategies, including the market momentum, betting-against-beta, and financial distress factors. Similar to volatility forecasting, the prediction of daily price magnitude variations has strong potentials in devising high-performing trading algorithms and policies. To the best of our knowledge, no work in literature today has ever investigated the forecast of the magnitude of stock price variations for a certain company in a given task. We believe that by providing an efficient method able to solve such a challenging problem might provide further benefits even to stock price forecasting systems.

In this paper we present a novel NLP-based approach whose goal is to predict the magnitude of future stock price variations for individual companies of the S&P 500 index. The stock price information we have handled is related to the S&P500 dataset where we consider relative increment or decrement of the close price registered on a given day with respect to the close price of its previous day. More in detail, through a walk-forward strategy, we first automatically generate a number of lexicons from articles published by international newspapers (collected in the *Dow Jones DNA dataset*[5]) with the goal of identifying the words that have an impact on the market in a specific time-span. Then, from the original news and by exploiting the generated lexicons, we extract a set of relevant features to capture statistical indicators associated with the company and its industry (a certain business area where a set of companies are grouped together) in a given interval time. A Decision Tree is hence trained on the created features, using as labels the company's stock price variation on the next day.

The results and comparisons against baseline methods show that our classification is effective and, therefore, that our feature engineering step is able to correctly identify peculiarities of the market hidden among the news. Besides, after the prediction, our algorithm provides an explanation of it, by extrapolating a set of rules from the Decision Tree model and the most important words linked to high stock price variations. This has been performed according to the mission of the new field of *Explainable Artificial Intelligence* (XAI) [14], whose aim is to address how AI systems undertake their decisions.

Our approach extends our initial study [15], where we had a set of manual rules based on statistical measures and came up with simple heuristics to find appropriate match-thresholds to perform the forecast.

The innovations we bring with respect to the literature are therefore:

- We propose a feature engineering process where we create an extended set of features extracted using generated lexicons and news from DNA;
- We make use of Machine Learning-based predictive algorithms that take into account the generated features and the stock price variations of the next day as labels to be forecasted;
- Concerning the explainability of the model, we inspect the inferred Decision Tree and provide explanation examples and the list of words associated with high stock price variations;
- We confirm the correlation between our lexicons and stock price variations of individual companies, through an experimental study on industries of the S&P 500 index;
- We show that our approach is general and can be easily extended to other stock markets and news sources (e.g. social media), or by adopting other different kinds of classifiers and their combinations (e.g. *ensemble*).

The remainder of this paper is organized as it follows. Section II describes the background work on financial forecasting based on Natural Language Processing techniques. Section III defines the forecasting problem we target in this paper. The pipeline of our proposed approach is depicted in Section IV, where we show how we create the industry-specific lexicons, how we perform the feature engineering process, how we employ the generated features for the Machine Learning algorithms we have employed, and mention the explainability of the model reasoning on its underneath white-box thus showing how certain outputs are generated. In Section V we evaluate the performances of our method detailing the used dataset, the adopted baselines, and indicating the effectiveness of the approach. Finally, in Section VI, we conclude the paper with final remarks and discuss possible future directions where we are headed.

II. RELATED WORK

Natural Language Processing, Text Mining and Sentiment Analysis have been widely applied in the financial sphere to provide more and more insightful tools for supporting decision making [5]. The increasing number of studies in the last couple of decades can be attributed to the development of techniques that allow an effective automatic processing of textual information, such as probabilistic topic models [16] and word-embeddings [17]. Furthermore, this research branch has been fostered by the birth and fast spread of social media platforms and micro-blogging websites such

[4]https://en.wikipedia.org/wiki/Volatility_(finance)
[5]https://developer.dowjones.com/site/global/home/index.gsp

as Twitter and StockTwits. This has led to a dramatic growth of the amount of user content, which constitutes a potentially valuable source of information for financial applications [18]. NLP within financial services is quickly expanding to beyond its current usage in banking, insurance and hedge funds. NLP technology has been a core component in chatbots, voice assistants, text analytics: today it is considered the next disruptor in the financial sector. For example, instead of logging into individual accounts for balance checking, users may simply use chatbots and voice assistants to check their account details. Authors in [19] have extracted financial events from financial announcements by constructing an end-to-end model with transformer encoder and the BiLSTM-CRF event recognizer. Others (see [20], [21]) have proposed an algorithm to construct automatically the relation graph from banking orders and, by using both news and bank contact histories, to capture the relations between corporate customers with Granger causality analysis. The construction of the personal knowledge graph can be considered a future research direction [22]. It retrieves extra features from the customers' daily lifelogs and can be used for several tasks, such as the risk evaluation of insurance companies, the measurement of default possibility of commercial banks, and personalized precision marketing.

A good example of NLP within the financial domain is Cleo,[6] a personal financial assistant that provides financial advice and helps clients meeting their financial goals. LenddoScore[7] is a system that uses advanced NLP and machine learning algorithms to assess creditworthiness of borrowers. The platform checks customer's social data to come up with a score that measures the creditworthiness of an individual.

While a remarkable amount of investigation has been conducted on textual data coming from social media [23]–[25], we hereby focus our attention on studies that, similarly to ours, analyze data coming from press releases or company disclosures. Several works in the financial literature, drawing inspiration from the aforementioned Adaptive Market Hypothesis, demonstrate that public news have an impact on the stock markets, partly accounting for the variance of returns [26]. This explains the effort to create automatic tools that are able to extract insights from financial news, with the objective of supporting companies, traders and all the other actors involved in the market [27].

A promising branch of research employs event detection techniques to extract relevant topics from news documents. Authors in [28] propose a supervised algorithm that automatically identifies pre-determined economic event categories in a sentence of a news article, by means of a sentence-level multilabel classifier. Others [29] develop a clustering-based method to detect events in news stories related to specific stocks, improving the original hierarchical algorithm based on average-linkage; subsequently, the single-pass clustering algorithm is used to accomplish the tracking of the identified events and the relevant topics are shown to the final user in chronological order. Research work mentioned in [30] exploits the Open Information Extraction tool developed by [31] to identify events in financial news and represents them as tuples. A neural tensor network is trained to learn event embeddings, which are then fed to a deep learning model to forecast short-term and long-term stock price movements on S&P 500.

Similarly to [30], many researchers have attempted to exploit the information extracted from the news in order to predict the future movements of the stock prices. Work performed in [32] combines news textual data and S&P 500 price time series to estimate a discrete stock price twenty minutes after a news article was released, using Support Vector Machines [33]. Authors in [34] analyze the effect of news sentiment and different levels of aggregation on the time horizon of the stock return predictability. Specifically, they use a neural network-based method to demonstrate that daily news can predict returns within one or two days, whereas aggregating news over one week provides predictability for up to 3 months; furthermore, the authors show that stock returns react quickly to positive news stories, while they absorb the influence of negative stories throughout a longer time span. Authors in [35] used the Harvard psychological dictionary and Loughran–McDonald financial sentiment dictionary to construct a sentiment space proving that, at individual stock, sector and index levels, the stock price prediction is improved. These findings particularly inspired us in the design of the feature extraction stage presented in this paper, which takes into account different time horizons in the aggregation of the news. In fact, differently from the work carried out in the past, here we leverage a huge amount of news data to create ad-hoc lexicons using the walk-forward strategy with the goal of predicting the magnitude of stock price variations. To the best of our knowledge, no prior work has ever addressed this specific problem. More in detail, we leverage the news data to create a lexicon that is intrinsically designed and created to support our task. We employ classical machine learning approaches (a Decision Tree) and indicate the most relevant features which affect the stock price variations. Adopting more advanced tools such as Deep Learning approaches (e.g. transformers) was out of the scope of the paper because we just wanted to confirm that the proposed approach beats the random classification for the proposed task.

III. PROBLEM FORMULATION

The input of our problem consists of a set of companies, grouped into industries or business sectors. For example, the business sector: ''Information Technology'', would contain companies such as Apple, Google, Microsoft, Accenture, etc.[8] These companies are associated, on the one hand, with a collection of news articles delivered by authoritative press sources and, on the other hand, with the stock price time series

[6]https://www.meetcleo.com/

[7]https://lenddo.com/

[8]At the following url https://en.wikipedia.org/wiki/List_of_S%26P_500_companies, it is possible to see the business sectors and the list of considered companies.

within the S&P 500 Index. In this work we set out to tackle the problem of predicting the magnitude of the variation, on a given day d, of the stock price of a company c; this quantity is denoted as $\Delta_{d,c}$ and is defined as the relative increment – or decrement – of the *close* price (i.e., the price of the stock at the closing time of the market), registered on the day d with respect to the close price of the previous day $(d-1)$:

$$\Delta_{d,c} = \frac{|\ close_d - close_{(d-1)}\ |}{close_{(d-1)}} \tag{1}$$

This formula considers the absolute value because, as previously mentioned, we intend to determine the magnitude of the variation, regardless of its direction (positive or negative). Indeed, an estimation of this value can still be precious piece of information for the trader, as it represents an indicator associated to the volatility of the stock price, thus conveying a measure of risk. In this respect, it is comparable to the CBOE Volatility Index (VIX), a real-time market index that aims to estimate the volatility expectation of the S&P 500 stock index in the following 30 days [36], although the latter's effective ability to predict the future volatility is debated [37].

More specifically, we define our problem as a classification task, in which:

- each sample corresponds to a day-company pair (d, c) for which we have a set of news articles $N_{d-1,c}$ published on the day $(d-1)$ and involving the company c;
- the two classes are `high` and `low`, defined as follows:

$$Class(d, c) = \begin{cases} \texttt{high} & \text{if } |\Delta_{d,c}| > class\_threshold \\ \texttt{low} & \text{otherwise} \end{cases} \tag{2}$$

where *class_threshold* is a fixed value above which we consider the market variation as significant.

IV. PROPOSED APPROACH

From a general perspective, the pipeline of the proposed approach (Figure 1) can be split into four different stages: (i) lexicon creation; (ii) feature extraction; (iii) prediction algorithm; and (iv) model explanation.

The main idea behind the first step is to capture the correlation between words that appear in news stories and stock price movements. This is achieved by means of a series of lexicons that contain the most impactful words in a given period for a specific industry, following the method described in [15]. Secondly, the generated lexicons are used to extract a set of features that characterize the news associated with the industry and specific firms, aggregated on a monthly, weekly and daily level. Then, the feature vectors are fed to a Decision Tree classifier that predicts whether the daily stock price variation is labelled as `high` or `low`, according to Equation 2. Finally, the rules that determine the prediction are retrieved from the Decision Tree and presented to the user as a form of explanation, together with the lists of relevant sentences that contain the lexicon words, selected from the groups of news considered for the feature extraction.

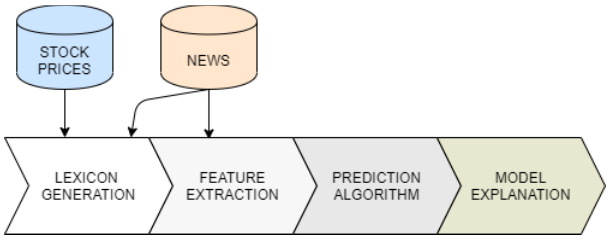


FIGURE 1. Pipeline of the proposed approach.

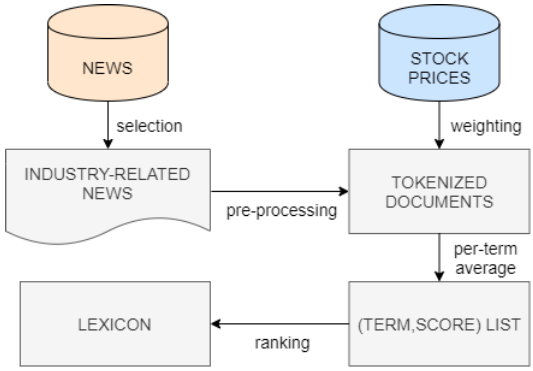


FIGURE 2. Flow chart of the lexicon creation algorithm.

The four steps of the pipeline are executed on each day on which the stock market is open. In this way, the lexicons are dynamically updated to incorporate the new information delivered on the press about the associated business sector. Grouping the companies by industry allows to reduce the ambiguity deriving from linguistic properties such as polysemy or heterosemy, as words that appear multiple times in the same context tend to be used with the same meaning [38], [39]. Furthermore, it is reasonable to assume that stock prices of companies pertaining to the same industry have comparable reactions to current affairs. Therefore, the lexicons can be defined as *dynamic* and *industry-specific*. The following sections describe the stages of the algorithm in further detail.

A. LEXICON GENERATION

The first stage of the pipeline (Figure 2) consists of the creation of the industry-specific lexicons [15]). For the sake of completeness, we hereby set out to summarize the steps of the algorithm and provide details about the employed text pre-processing techniques.

Given an industry I, the first source of input of the lexicon generation algorithm is a collection of news articles that are relevant, to some extent, for at least one company $c \in I$, published during a time frame $[d-\ell; d-1]$, with $\ell \geq 1$. To obtain a textual representation of the article, we concatenate the text contained in its title, snippet and full body. A news document published in a certain date is considered relevant for a company c whether either its full name or its acronym appears somewhere in its textual representation. From now on, we will therefore assume each document associated to one or more companies. Every document undergoes a series of standard pre-processing techniques. First of all, tokenization is necessary to get rid of punctuation and to

represent the text to a simple list of words. One drawback of this type of representation is that the semantic information given by compound terms and expressions (e.g. *Wall Street*, *Barack Obama*, *high interest rate*) tends to get lost. One technique to overcome this is to consider n-grams, with n that can typically range between 2 and 4. Subsequently, stopwords removal is applied to the tokenized documents, to guarantee that extremely common words of the language (English in our case), such as prepositions and conjunctions, are not accounted for in the estimation of the impact of the words, since they are devoid of semantic value. In addition, we remove from the corpus all the words that appear too frequently and too infrequently, according to given tolerance thresholds, which were set through an experimental optimization (more details are given in Section V-B). In fact, it is uncertain to estimate the impact of words that appear very few times, as their correlation with stock prices might be subject to a higher degree of randomness. On the other hand, terms that appear too frequently are likely to be commonly used words of the language, with a low relevance for the topic discussed in the document (similarly to stop-words). In the next step, stemming is used to reduce each term, that normally appears in some inflected form (plurals, verbs in different tenses, etc.), to its root form (e.g. the word *cancelled* becomes *cancel* after stemming). This is useful because it allows to ignore small differences in the inflections of words that, nonetheless, correspond to the same semantics.

After the pre-processing is complete, the terms of each news document are weighted according to the stock price variation registered by the company associated with the article on the day following its publication. To note that we do not make use of the frequency information of each word but this will be investigated in future works to check whether it brings benefits to the overall lexicon creation process. Specifically, each word in the document receives a score s that corresponds to the absolute value of the stock price variation, as expressed in Equation 1. Finally, for every term we compute a unique score s', given by the average of all the scores s associated to each occurrence of that term throughout the collection.

B. FEATURE EXTRACTION

Once the specialized lexicon is created as described in previous Section IV-A, it is used to extract a set of features associated to each (company, day) pair for which we want to produce a prediction. In order to take into account the effect of words in the long-term, in the mid-term and in the short-term, we decided to compute the features on groups of news articles aggregated using different time intervals. Furthermore, it is important to capture the difference between news stories associated with high variations and those associated with low variations. We define a news story as *associated with a high (low) variation* if the stock price variation (Equation 1) on the day following the publication of the article is bigger (smaller) than a pre-defined *class_threshold*.

More in detail, given an industry I and a (company, day) pair (c, d) where $c \in I$, we extract from the news dataset all the articles published in the last 30 days about the company c for which we want to predict the variation and about the industry I. These articles are divided between company-related and industry-related. This latter group contains articles relevant for any company of the industry, including c; for this reason, there is a small overlapping between the two groups. Both company-related and industry-related news, respectively, are further split into three groups: documents published in the previous month, in the previous week and in the previous day (also in this case there is a small overlapping, since the daily news are a subset of the weekly news and the weekly news are a subset of the monthly news). At this point, previous month and previous week news, respectively, are further split into two mutually exclusive groups: articles associated with high variations and articles associated with low variations. This cannot be applied to previous day news because, at the moment of the prediction, the stock price variation on the following day is still unknown (indeed, this is the target of our predictor). Now the lexicon matching phase takes place. For each of the ten final groups of news described above, after applying the same aforementioned text pre-processing techniques, we count the percentage of lexicon-words that each of the documents in the group contains, with respect to the total number of words in that document. Then, we calculate the average percentage of lexicon-words for each group: these ten real values, expressed in a [0-100] scale, are precisely the features that constitute the samples fed into the classifier. Intuitively, each feature represents a statistical indicator about the behavior of the industry or the company in a given time interval. We invite the reader to follow the steps of the algorithm with the help of Figure 3.

C. PREDICTION ALGORITHM

The feature vectors described in previous Section IV-B constitute the input for the prediction algorithm, that relies on a Decision Tree classifier (4). The target of the classification task is the class assigned to each (day, company) pair, following Equation 2.

Decision Trees belong to the category of supervised learning methods for classification and regression [40]. They learn simple decision rules inferred from the data features in order to create a model that predicts the value of a target variable. One of the main advantages of Decision Trees is their explainability, as they use a *white box* model: in fact, the trees can be visualized and the predictions can be easily explained in terms of conditions expressed in boolean logic. By contrast, *black box* models (e.g., in an artificial neural network), are typically more difficult to explain, even though they may lead to better performance in terms of accuracy. For further details on Decision Trees the reader is referred to [41] and [40], among others.

D. MODEL EXPLANATION

The explanation has the goal to give users awareness on the main working mechanisms of the model, providing details on

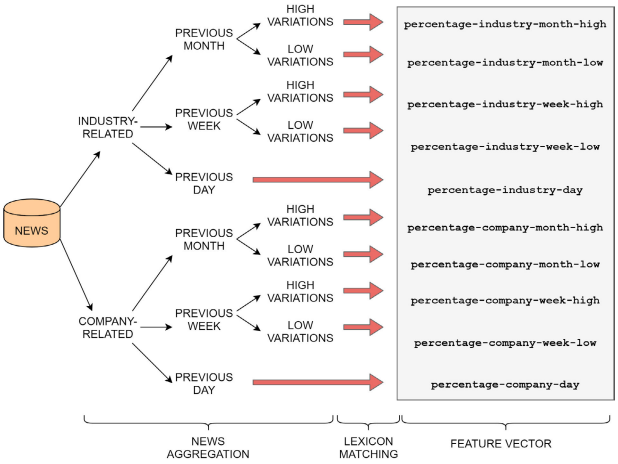


FIGURE 3. Flow chart of the feature extraction algorithm. For the feature vector of a given day *d*, we collect and group up all the articles about *c* and all other companies in *I* published in the previous month, in the previous week and in the previous day, respectively. For the monthly and the weekly sets, we further split the articles into two groups depending on the associated stock price variation on the next day (high or low variation, respectively). As a result, we obtain ten sets of news in total, which are summarized on the left side of Figure 3. Please note that it is not possible to associate a stock price variation to the news published on the previous day, since this is exactly what we intend to predict. For each group of news, we calculate the average percentage of words that the articles in the group share with the lexicon. In this way, we obtain a vector of ten real-valued features for each sample included in the classification task.

the specific input conditions that determine a certain prediction as output. In particular, two pieces of information are provided:

1) Sequence of rules followed by the model to produce the prediction;
2) Lists of sentences extracted from the news articles that generated the features.

In the following we better describe these steps.

Sequence of Rules: The rules are easily extracted from the Decision Tree induced on the training data. By construction, each internal node of the tree, including the root, is associated to a rule, expressed as an inequality of the form *feature* $\leq$ *value*; the leaves are associated to a target label. During the prediction step, the feature vector is matched against the rules from the root of the tree to one of the leaves. Therefore, to generate the explanation it is sufficient to present the user with the sequence of boolean conditions that were satisfied by the feature vector along this path. In order to make the explanation easily readable by the user, it is recommendable to fix the depth of the tree to a reasonable small value, since the number of levels in the tree corresponds to the number of rules in the explanation.

List of Sentences: The features involved in the boolean conditions correspond to the percentages of lexicon words present in some group of news articles, as described in Section IV-B. We can exploit this fact to extract sentences from the articles where such words appear and provide them to the user as a complement of the explanation. For instance, given the rules in the previous example, the algorithm

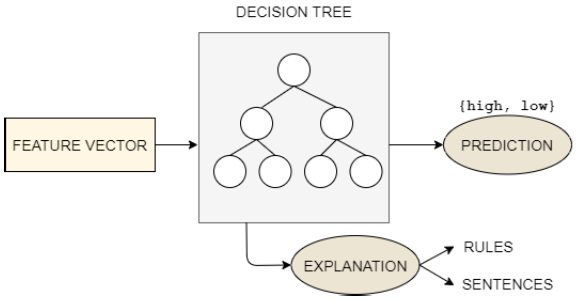


FIGURE 4. Flow chart of the prediction algorithm, including the generation of the model explanation.

would extract three lists of sentences, respectively from the monthly industry news associated with low variations, from the weekly industry news associated with high variations and from the company news published on the previous day. An example of explanation with rules and selected sentences will be illustrated in Section V-C2.

This kind of approach meets the recent and exponentially increasing trend of *Explainable AI* [42], [43], consisting of opening up the given black-box model by explaining how the method arrives to a certain final output. In this way predictions are more transparent and even allow for statistical inference. Explanation of the resulting analysis also allows easing algorithm reuse and extension [42].

V. EXPERIMENTAL EVALUATION

In this section, we illustrate the experiments carried out to assess the effectiveness and soundness of our approach. In particular: (i) we describe the datasets used for the evaluation; (ii) we introduce the adopted methodology and setting; and, finally, (iii) we present and discuss the obtained results.

A. DATASETS

Dow Jones DNA: The Dow Jones "Data, News and Analytics" dataset provides documents from more than 33, 000 globally renowned newspapers, including e.g. *The Wall Street Journal*, the *Dow Jones Newswires* and *The Washington Post*. The publications are both in print and online format and cover a wide variety of topics, such as finance, business, current affairs and lifestyle. The delivery frequency ranges from ultra-low latency newswires to daily, weekly, or monthly editions. For every article in the dataset, the headline, the snippet and the full body are available. Furthermore, every item is enriched with a set of metadata providing information about the source, the time and place of the publication, the relevant companies and the topics, among others.

Content usage rights vary based on the specific content, API, or feed combination. These rights include the display for human consumption or text mining for machine consumption and the content retention period. Table 1 includes some statistics about the news data. In particular, for each of the three industrial sectors pertaining to our study, the table shows the number of news employed in the analysis, the average

TABLE 1. Number of documents and statistics for each of the three analysed industrial sectors.

Industrial Sector	#News	Avg. #news per company	Std. Dev. #news per company	Max. #news per company	Min. #news per company
Information Tech.	43331	953.01	2149.31	11705 (APPLE)	1 (DXC Technology)
Financial	40115	976.46	1754.50	7105 (CITYGROUP)	4 (KEYCORP)
Industrials	36233	791.67	1292.01	6189 (BOEING)	1 (FORTIVE)

TABLE 2. List of five of the most important companies for every industry included in our study.

Information Technology	Financial	Industrials
AMD	American Express	American Airlines
Adobe	Bank of America	Boeing
Apple	Goldman Sachs	Delta Airlines
IBM	JPMorgan	FedEx
Microsoft	Wells Fargo	3M Company

number of news and standard deviation per company, and finally the company with the highest and lowest number of news. Note that a news document might belong to more than one industrial sector.

S&P 500 Time Series: The second fundamental data source exploited in our analysis consists of all the stock price *time series* of the companies included in the *Standard & Poor's 500* index (that measures the stock performance of 500 large companies listed on stock exchanges in the United States). Data are collected separately for each individual company at a daily frequency, and they include the following information:

- *open* price (i.e. price of the stock at the opening time of the market);
- *close* price (i.e. price of the stock at the closing time of the market);
- *high* price (i.e. maximum price reached by the stock during the day);
- *low* price (i.e. minimum price reached by the stock during the day);
- *volume* (i.e. number of operations performed on the stock during the day).

The dataset also provides information about the grouping of companies in sectors (e.g. Manufacturing, Healthcare, Information Technology, Communication Services, Finance, etc.). Table 2 shows a list of five among the most relevant companies included in each business sector considered in our study.

B. METHODOLOGY AND SETTING

For our purposes, we selected from DNA all the news articles, in English language, published from 2005 to 2018 and relevant – according to the metadata field *company_codes_about* – for at least one company of one of these three industries: Information Technology, Finance and Industrials. We then grouped the documents by industry, obtaining three groups, which respectively contain 43, 331, 40, 115 and 36, 233 items. To align these documents with stock prices, we restricted the S&P 500 dataset to the same interval, for all the companies included in the aforementioned business sectors.

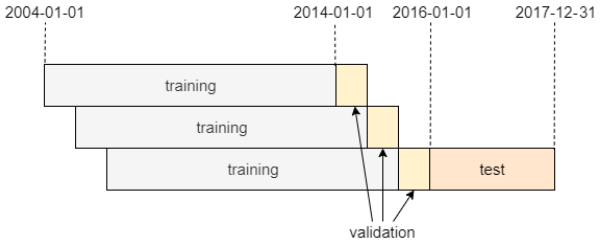


FIGURE 5. Structure of the walk-forward approach used for validation.

The goal of our experimental evaluation is twofold:

- show that the feature-extraction method and the predictive algorithm proposed in this paper capture the correlation between news and stock price movements;
- demonstrate that the Decision Tree classifier can produce explainable predictions, without significant losses in terms of accuracy w.r.t. other state-of-the-art models.

We repeated the experiments on each industry taken alone and we employed a walk-forward validation which iteratively splits the data into training and validation sets, guaranteeing that all the samples in the former temporally precede the ones of the latter (Figure 5). According to the common practice in Machine Learning, the training set was used to infer the Decision Tree (and the other models used for comparison), while the validation set was used to optimize the parameters. On top of that, an independent test set – contiguous and subsequent to the validation set – was isolated to test the model with the selected parameters on unseen samples. The number of walks to iterate in the walk-forward approach was set to 10, with fix-sized portions of training set starting from 2004/01/01 and lasting 10 years (meaning that for the first walk the training set will cover the period: [2004-01-01, 2014-01-01], for the second walk it will be shifted of the validation set size, which is 1 year, and will therefore cover the period: [2005-01-01, 2015-01-01], and so on for the 10 walks). The whole validation set was collected from 2014/01/01 to 2015/12/31, while the test set was collected from 2016/01/01 to 2017/12/31.

The *class_threshold* that determines the class label of each sample (Equation 2) was set empirically to 0.02 (i.e. corresponding to a 2% daily variation, in absolute value). This configuration leads to a 22%, 28% and 26% of samples labeled as `high` in Information Technology, Financial and Industrials, respectively. Given the low percentage of days above the threshold value of 0.02, we have considered such

TABLE 3. Settings used for the evaluation methodology.

Interval time of news articles	2005-2008
Number of walks	10
Starting time for the training set	2004/01/01
Size of the training set	10 years
Interval time for the validation set	2014/01/01-2015/12/31
Interval time for the test set	2016/01/01-2017/12/31
class_threshold	0.02
Interval time for selecting news to include in the lexicon	28 days
Thresholds to filter out words	> 70% or < 10 documents
Percentile for selecting words in the lexicon	upper 95th

variations (i.e. 22%, 28% and 26%) as meaningful for our classification task.

The class imbalance that follows from this setting is taken into account both during the inference of the Decision Trees[9] and in the evaluation phase (by using appropriate metrics such as Balanced Accuracy, Precision, Recall and F1-score, as better defined in the next section). The reader notices that the impact of the choice of different threshold values has been already shown in [15].

For reproducibility purposes, the approach and experimental framework were developed in Python employing a set of open source Machine Learning libraries. The implementation of Decision Tree and the other state-of-the-art classifiers used for comparison was based on the scikit-learn[10] and XGBoost[11] Python libraries. For the text pre-processing tasks we have used the gensim[12] and Natural Language Toolkit (NLTK)[13] Python libraries. The lexicons employed throughout the experiments were generated using the following setting. Only uni-grams were considered. The time interval to select the news to include in the lexicon was set to 28 days. Furthermore, words that appeared in more than 70% or less than 10 documents were filtered out and, finally, words in the upper 95th percentile of the ranking were selected for inclusion in the final lexicon. The values of these parameters, the configuration of the Decision Tree and the other classifiers included in the comparison (Section V-C1) were selected empirically through an experimental validation. Table 3 briefly lists all the parameters previously mentioned. Regarding the Decision Tree implemented with scikit-learn, its hyper parameters were experimentally set as specified in the following. The parameters *max_depth* (i.e. the maximum depth of the tree), *min_samples_split* (i.e. the minimum number of samples required to split an internal node), and *min_samples_leaf* (i.e. the minimum number of samples required for a node to be a leaf node), were set, respectively, to the values 4, 2, and 1, since this has been the resulting combination leading to the best accuracy for all the three considered industries on the validation set; all the other parameters have not shown to influence the final results and therefore have been set to the default scikit-learn values.The developed code has been publicly released in a freely accessible GitHub repository.[14] Furthermore, in order to avoid overfitting the models on a single industry, the model parameters have been chosen by optimizing the weighted average of the Balanced Accuracy on the three industries in the validation set.

[9]For more details, please check https://rb.gy/ki6bdi
[10]http://scikit-learn.org
[11]http://xgboost.readthedocs.io
[12]http://radimrehurek.com/gensim
[13]http://nltk.org

C. RESULTS

1) ALGORITHM COMPARISON

The goal of the first set of experiments is to compare the proposed approach based on Decision Trees against some of the most commonly employed state-of-the-art classifiers, namely Random Forest [44], Gradient Boosting [45], and Multilayer Perceptron, i.e. a basic artificial neural networks (ANNs) model [46]. Our baseline is represented by the method presented in [15]. The latter is an unsupervised predictive algorithm that exploits the input news documents and the generated lexicons to perform the forecasts about the magnitude of future stock prices variations. It does not use any machine learning classifier but just statistical measures and metrics. Basically, we first calculate the percentage of words belonging to articles associated to high variations and low variations that the documents share with the lexicon created within the same time interval of the news. Let $news\_match_{high}$ and $news\_match_{low}$ be these values. Then we select the news articles published on the $d - 1$ day, related to a given company, and calculate the percentage of words contained in the corresponding lexicon. If this value is closer to $news\_match_{high}$, we then assign the class *high* to the sample; otherwise, the class *low* is assigned. Our method was already proven to perform better than a random classifier which predicts both classes with 50% probability and that we include also here for illustrative purposes, hence strongly demonstrating its ability to achieve valuable results also in this difficult forecasting scenario. In fact, the task of predicting the magnitude of price variation (i.e., the volatility) is a very difficult challenge; indeed, in the wider domain of market forecasting, achieving high accuracy values is commonly harder than in classic machine learning tasks [47]. Moreover, other works in literature highlighted that, often, canonical methods for volatility prediction are not able to show a clear effectiveness, and that their performances are highly sensitive to the selected evaluation metric [48].

The performance of the algorithms is measured through the following four standard metrics [49], specifically selected, as already mentioned, to cope with the class imbalance of the dataset described in Section V-B: ''Balanced Accuracy'', ''Precision'' of class `high` and ''Recall'' of class `high` and ''F1-score'' of class `high`.

Balanced accuracy [49] gives a global estimate of the performance of a classifier avoiding inflated estimates on imbalanced datasets. In the binary case, it is equal to the arithmetic mean of sensitivity (true positive rate) and specificity (true

[14]https://github.com/Artificial-Intelligence-Big-Data-Lab/Explainable-ML

negative rate):

$$BalancedAccuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right),$$

where TP is the number of true positive samples (`high` samples correctly classified), FN is the number of false negative (`high` samples wrongly classified as `low`), TN is the number of true negative (`low` samples correctly classified) and FP is the number of false positive (`low` samples wrongly classified as `high`).

Precision and Recall [49] were included in the evaluation to gain a more detailed assessment of the performance of the class `high`. Intuitively, the former provides a measure of the classifier's exactness, whereas the second allows to gauge the classifier's completeness. They are formally defined as follows:

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$

Finally, the F1-score [49] provides a weighted average of the Precision and Recall and is defined as the harmonic mean of the two metrics:

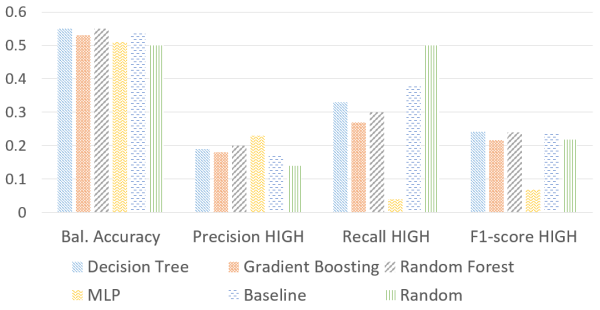
$$F1\_score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The results in Figure 6, grouped by industry, show that the Decision Tree-based approach is competitive against the state-of-the-art classifiers (Gradient Boosting, Random Forest and Multilayer Perceptron), achieving comparable or better values in all the considered metrics. Furthermore, our approach (Decision Tree) leads to a better Balanced Accuracy compared to the baseline of [15] and, consequently, to the random classifier, even though the difference is significant only for Finance. The trade-off between Precision and Recall is a well-known phenomenon in the evaluation of Machine Learning algorithms. In our scenario, Decision Tree obtains higher Precision values than the baseline, but it is outperformed in terms of Recall (especially against the random classifier, which achieves a 50% Recall by construction). For this reason, it is useful to observe the value of the F1-score, which clearly indicates that our approach is globally more effective at predicting the class `high`.

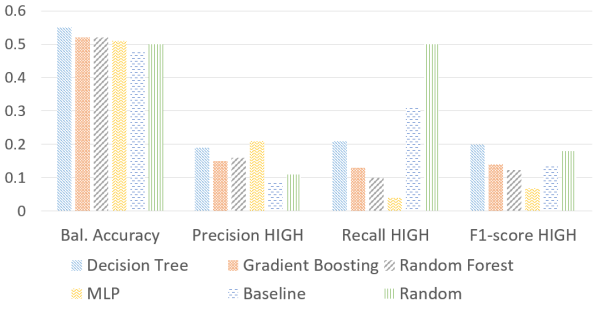
2) MODEL EXPLAINABILITY

In this section, we show a qualitative evaluation of the explainability of the proposed model [50], [51], by inspecting the *white box* of the Decision Tree and by providing examples of explanation.

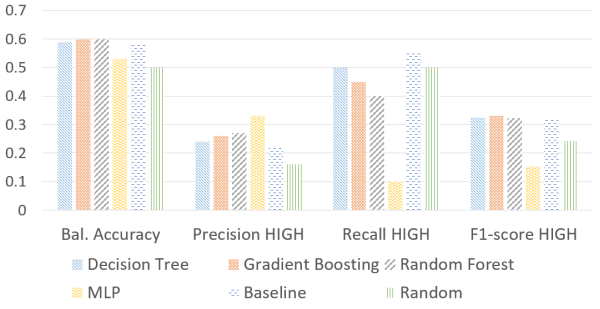
Figure 7 shows an example of Decision Tree inferred on a specific portion of the walk-forward process for the Industrial sector. In this case, the maximum depth of the tree is set to three, in order to allow an easy inspection. Each path from the root to one of the leaves is a possible rule-based explanation of the prediction.



(a) Information Technology



(b) Financial



(c) Industrials

FIGURE 6. Comparison of the algorithms on three different industries.

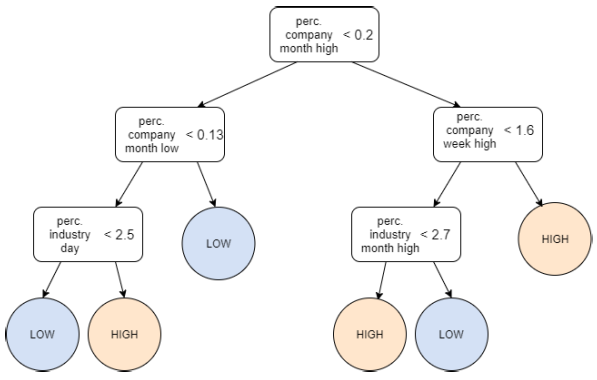


FIGURE 7. Example of Decision Tree inferred on a specific walk for the Industrial business sector. At each node, the left-hand arrow is followed whenever the boolean condition associated with the node is true. Otherwise, the right-hand arrow is followed. Each leaf indicates the predicted class label.

An interesting aspect that can be observed from the Decision Tree models is the importance of each feature, computed as the normalized total reduction of the criterion brought by that feature (in this case, in terms of

TABLE 4. Example of sentences selected from the news published in the 30 days before May 17th, 2016, for the Information Technology sector, as a form of explanation. The words in bold are the terms that belong to the dynamic lexicon created for the related time-span and associated with high stock price variation of companies within the considered sector.

Microsoft's first-quarter results, due Thursday, should help push shares closer to that high-water **mark**.
The software **giant**'s stock price, left for dead during much of the 2000s under former chief Steve Ballmer, has returned to life.
Microsoft's "Intelligent Cloud" made up about 28% of the company's **revenue** in the second half of last year.
Disney's service, part of the U.S. entertainment **giant**'s strategy to chart a digital future, was suspended at the request of Chinese regulators, two people familiar with the matter said.
Apple is grappling with a **slowdown** in the sales of its iPhones and the aftermath of a bruising battle with the F.B.I. over the privacy and security of its devices.
Mr. Daryanani also noted that many companies had been hurt by the **strong** dollar, which makes American products more expensive overseas.
Apple and its supporters argue the government has stretched its use of the All Writs Act **far** beyond What Congress intended, and they argue if the government wants that kind of authority it should convince lawmakers to pass a new law.
Apple still generates tens of billions of dollars of **cash** every year.

TABLE 5. List of the first 10 lexicon words, ranked by absolute frequency, that appear in the news about Information Technology, published in the 30 days before May 17th, 2016, associated with high stock price variations.

word	absolute frequency	document frequency
strong	11	0.37
giant	8	0.37
far	7	0.31
question	6	0.31
slowdown	6	0.26
cash	4	0.1
margins	3	0.15
revenue	2	0.12
happen	2	0.1
book	2	0.13

the Gini coefficient). We computed the feature importance for each walk in the validation for all the three industries, and we calculated the average importance of every feature to get a global estimate. In order to get a value for each feature, the maximum depth of the tree was left unbounded for this specific experiment. Something worth to note is related to the first two ranked features being the `percentage-company-month-high` with a value of 0.124 and `percentage-company-month-low` with a value of 0.12. They have higher values than all the other features implying that news over a month related to single companies have a bigger impact with respect to those related to industries. This also corresponds to the intuition that news related to a certain specific company probably contain important words pertaining to that company and that are more likely to occur within the generated lexicon.

Now, let us also consider the other pairs *(feature, value)* which are positioned later in the ranked list: (`percentage-industry-week-high`, 0.113), (`percentage-industry-month-high`, 0.108), (`percentage-industry-month-low`, 0.105), and (`percentage-industry-week-low`, 0.096). Considering also the first two pairs shown above, in each case, company per month, industry per week and industry per month, the feature corresponding to `high` has always a higher value than its counterpart corresponding to `low`. This indicates that news associated to big stock variations (either positive or negative) contain words with a bigger impact toward the classification with respect to the news associated with low stock variations. This suggests the possibility to go one step further and create a lexicon of all such words. As this analysis goes out of the purpose of this paper (where we generated a lexicon which proved to be efficient for the forecast of stock price variations in absolute value), it is a direction we would like to explore more in detail as future work.

Nevertheless, in Table 4 we report some examples of sentences extracted from news associated with high price stock variations, according to the procedure described in Section IV-D. Table 5, instead, illustrates a list of the 10 most frequent lexicon words that appear in this group of news.

VI. CONCLUSION AND FUTURE WORK

In this work we have proposed an approach based on an explainable Machine Learning model to predict the magnitude of future stock price variations for individual companies of the S&P 500 Index. A series of lexicons are created from articles published by globally renowned newswires, with the goal of identifying the words that have an impact on the market in a specific time-span within a given business sector. Subsequently, these lexicons are exploited to extract a set of features from the same collection of news, to capture certain statistical indicators associated with the industry and the company in a given time. A Decision Tree classifier is trained to predict the class of magnitude associated to the company's stock price variation on the next day. Finally, the algorithm provides an explanation of the prediction, by extrapolating a set of rules from the Decision Tree model, expressed as simple boolean conditions that are satisfied by the feature vector. The explanation is complemented by a list of sentences containing relevant lexicons words, selected from the groups of news articles considered in the feature extraction phase. Note that that by changing either news documents, market or classifier, the entire approach remains still valid. For this reason, in order to stimulate further applications and studies in the subject, we have freely released the developed source code along with examples on the used data. For further improving model explainability, approaches such as LIME [52] might also be exploited to aim at a deeper understanding of the obtained set of features.

Through our experimental validation, we demonstrate that the proposed approach is more accurate with respect to both the best to date baseline in the literature and other considered state-of-the-art classifiers, all having the drawback of employing hardly explainable *black box* models. We carry out both a quantitative evaluation, showing the competitiveness

and superiority of the proposed model with respect to the other methods, and a qualitative evaluation of the explainability of the model, by inspecting the inferred Decision Trees and by providing examples of explanations. The results presented in this paper should be analysed by taking into account the very challenging financial forecasting context considered, where accuracy values of forecasting models in the literature typically reach lower values. For this reason we believe that achieving a Balanced Accuracy between 0.5 and 0.6 scores is a non-trivial and very remarkable result. Our overall objective consists in achieving a methodology both competitive in forecasting performance and explainable in the obtained output, in order to further encourage the development of Machine Learning algorithms allowing for a human-readable understanding of the obtained results.

Although the proposed approach is promising for extracting relevant lexical properties from news sources, one of its limits is that it does not consider the semantic value of their content, which could provide additional information to further improve the forecasting accuracy. As a future improvement, we intend to integrate semantic features into our system by means of state-of-the-art techniques such as topic models, frame semantics and word embeddings, with the goal of identifying words corresponding to features of the classifier with a bigger impact. Another challenge that is worth undertaking is the extraction of events from news stories, which would allow distinguishing more clearly between factors that have an actual impact on the stock markets from others that are in fact irrelevant. In order to achieve this, traditional news wires can be combined with other sources, such as social media platforms, in order to gain a richer perspective of the events and the associated sentiment. Also, we intend to conduct a more fine-grained analysis, observing the behaviour and measuring the performance on individual companies. Specifically, an aspect we want to better investigate is whether the number and frequency of news published about a company, regardless of their content, or the frequency of each words used when creating the lexicon, can be used as effective indicators of higher volatility to improve the forecasting capability of our model. Finally, in the future we also intend to investigate the adoption of neural networks approaches (e.g. transformers) for the problem, performing a detailed comparison against the results obtained by the method proposed here.

REFERENCES

[1] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *J. Finance*, vol. 25, no. 2, pp. 383–417, May 1970.

[2] X. Li, X. Huang, X. Deng, and S. Zhu, "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information," *Neurocomput.*, vol. 142, pp. 228–238, Oct. 2014.

[3] A. W. Lo, "The adaptive markets hypothesis," *J. Portfolio Manage.*, vol. 30, no. 5, pp. 15–29, Jan. 2004.

[4] A. Atkins, M. Niranjan, and E. Gerding, "Financial news predicts stock market volatility better than close price," *J. Finance Data Sci.*, vol. 4, no. 2, pp. 120–137, Jun. 2018.

[5] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: A survey," *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 49–73, Jun. 2018.

[6] M. R. Vargas, C. E. M. dos Anjos, G. L. G. Bichara, and A. G. Evsukoff, "Deep leaming for stock market prediction using technical indicators and financial news articles," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[7] S. Carta, A. Corriga, A. Ferreira, A. S. Podda, and D. R. Recupero, "A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning," *Appl. Intell.*, vol. 51, pp. 1–17, Sep. 2020.

[8] T. Matsubara, R. Akita, and K. Uehara, "Stock price prediction by deep neural generative model of news articles," *IEICE Trans. Inf. Syst.*, vol. E101.D, no. 4, pp. 901–908, 2018.

[9] S. Carta, A. Ferreira, A. S. Podda, D. Reforgiato Recupero, and A. Sanna, "Multi-DQN: An ensemble of deep Q-learning agents for stock market forecasting," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113820.

[10] G. Moro, R. Pasolini, G. Domeniconi, A. Pagliarani, and A. Roli, "Prediction and trading of Dow Jones from Twitter: A boosting text mining method with relevant tweets identification," *Commun. Comput. Inf. Sci.*, vol. 976, pp. 26–42, Nov. 2019.

[11] M. Atzeni, A. Dridi, and D. R. Recupero, "Using frame-based resources for sentiment analysis within the financial domain," *Prog. Artif. Intell.*, vol. 7, no. 4, pp. 273–294, Dec. 2018.

[12] A. Dridi, M. Atzeni, and D. R. Recupero, "FineNews: Fine-grained semantic sentiment analysis on financial microblogs and news," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2199–2207, Aug. 2019.

[13] S. Cederburg, M. S. O'Doherty, F. Wang, and X. Yan, "On the performance of volatility-managed portfolios," *J. Financial Econ.*, vol. 138, no. 1, pp. 95–117, Oct. 2020.

[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[15] S. Carta, S. Consoli, L. Piras, A. Podda, and D. R. Recupero, *Dynamic Industry-Specific Lexicon Generation for Stock Market Forecast*. (Lecture Notes in Computer Science), vol. 12565. Cham, Switzerland: Springer, 2020.

[16] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[18] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media (ICWSM)*, 2010, pp. 58–65.

[19] S. Zheng, W. Cao, W. Xu, and J. Bian, "Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 337–346.

[20] B. Oral, E. Emekligil, S. Arslan, and G. Eryiğit, "Extracting complex relations from banking documents," in *Proc. 2nd Workshop Econ. Natural Lang. Process.* Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 1–9.

[21] H. Sakaji, R. Kuramoto, H. Matsushima, K. Izumi, T. Shimada, and K. Sunakawa, "Financial text data analytics framework for business confidence indices and inter-industry relations," in *Proc. 1st Workshop Financial Technol. Natural Lang. Process.*, Macao, China, Aug. 2019, pp. 40–46.

[22] K. Balog and T. Kenter, "Personal knowledge graphs: A research agenda," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Sep. 2019, p. 217.

[23] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 513–522.

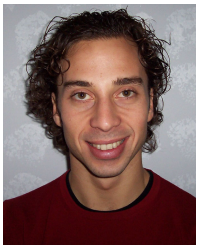
[24] M. Makrehchi, S. Shah, and W. Liao, "Stock prediction using event-based sentiment analysis," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI), Intell. Agent Technol. (IAT)*, vol. 1, Nov. 2013, pp. 337–342.

[25] J. Si, A. Mukherjee, B. Liu, S. J. Pan, Q. Li, and H. Li, "Exploiting social relations and sentiment for stock prediction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1139–1145.

[26] J. Boudoukh, R. Feldman, S. Kogan, and M. Richardson, "Information, trading, and volatility: Evidence from firm-specific news," *Rev. Financial Stud.*, vol. 32, no. 3, pp. 992–1033, Mar. 2019.

[27] A. Hogenboom, F. Hogenboom, F. Frasincar, K. Schouten, and O. Van Der Meer, "Semantics-based information extraction for detecting economic events," *Multimedia Tools Appl.*, vol. 64, no. 1, pp. 27–52, 2013.
[28] G. Jacobs, E. Lefever, and V. Hoste, "Economic event detection in company-specific news text," in *Proc. 1st Workshop Econ. Natural Lang. Process.*, 2018, pp. 1–10.
[29] X.-Y. Dai, Q.-C. Chen, X.-L. Wang, and J. Xu, "Online topic detection and tracking of financial news based on hierarchical clustering," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 6, Jul. 2010, pp. 3341–3346.
[30] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proc. 24th Int. Conf. Artif. Intell. (IJCAI)*. Palo Alto, CA, USA: AAAI Press, 2015, p. 2327–2333.
[31] A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, and S. Soderland, "TextRunner: Open information extraction on the Web," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2007, pp. 25–26.
[32] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 1–19, 2009.
[33] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[34] S. L. Heston and N. R. Sinha, "News vs. sentiment: Predicting stock returns from news stories," *Financial Analysts J.*, vol. 73, no. 3, pp. 67–83, 2017.
[35] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, Oct. 2014.
[36] C. J. Corrado and T. W. Miller, "The forecast quality of CBOE implied volatility indexes," *J. Futures Markets, Futures, Options, Other Derivative Products*, vol. 25, no. 4, pp. 339–373, 2005.
[37] B. K. Adhikari and J. E. Hilliard, "The VIX, VXO and realised volatility: A test of lagged and contemporaneous relationships," *Int. J. Financial Markets Derivatives*, vol. 3, no. 3, pp. 222–240, 2014.
[38] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Lang. Cogn. Process.*, vol. 6, no. 1, pp. 1–28, 1991.
[39] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Annu. Meeting Assoc. Comput. Linguistics*, 1995, pp. 189–196.
[40] C. D. Sutton, "Classification and regression trees, bagging, and boosting," *Handbook Statist.*, vol. 24, pp. 303–329, Dec. 2005.
[41] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May 1991.
[42] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Adv. Res. Projects Agency (DARPA)*, vol. 2, no. 2, pp. 1–18, Nov. 2017.
[43] H. Hagras, "Toward human-understandable, explainable AI," *Computer*, vol. 51, no. 9, pp. 28–36, Sep. 2018.
[44] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[45] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.
[46] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Trans. Neural Netw.*, vol. 1, no. 4, pp. 296–298, Dec. 1990.
[47] M. L. De Prado, *Advances in Financial Machine Learning*. Hoboken, NJ, USA: Wiley, 2018.
[48] T. J. Brailsford and R. W. Faff, "An evaluation of volatility forecasting techniques," *J. Banking Finance*, vol. 20, no. 3, pp. 419–438, Apr. 1996.
[49] J. Korst, V. Pronk, M. Barbieri, and S. Consoli, "Introduction to classification algorithms and their performance analysis using medical examples," in *Data Science for Healthcare*, S. Consoli, D. R. Recupero, and M. Petkovic, Eds. Cham, Switzerland: Springer, 2019, pp. 39–73.
[50] T. W. Kim and B. R. Routledge, "Informational privacy, a right to explanation, and interpretable AI," in *Proc. IEEE Symp. Privacy-Aware Comput. (PAC)*, Sep. 2018, pp. 64–74.
[51] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.
[52] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 144–1135.

SALVATORE M. CARTA (Member, IEEE) received the Ph.D. degree in electronics and computer science from the University of Cagliari, Italy, in 2003. In 2005, he joined the Department of Mathematics and Computer Science, University of Cagliari, as an Assistant Professor. In 2006 and 2007, he was a Guest Researcher with the Swiss Federal Institute of Technology, as an Invited Professor, hosted by the Laboratoire des Systmes Intgrs-LSI. He is currently an Associate Professor with the Department of Mathematics and Computer Science, University of Cagliari. His research interests include clustering algorithms, social media analysis, and text mining for behavioral pattern identification and recommendation in group of users and in single users, AI algorithms for credit scoring, fraud detection and intrusion detection, AI algorithms for financial forecasting and robo-trading, E-coaching platforms, and AI algorithms for healthy lifestyles. He is author of more than 110 conference and journal papers in these research fields, with more than 1400 citations. He is a member of the ACM. He founded three Hi-Tech companies, spin off of the University of Cagliari, and is also leading one of them.



SERGIO CONSOLI received the Ph.D. degree. He is currently a Scientific Project Officer with the European Commission, Joint Research Centre (DG-JRC), Ispra, Italy, also working with the Centre for Advanced Studies on the project Big Data and Forecasting of Economic developments. Previously, he was a Senior Scientist with the Data Science Department, Philips Research, Eindhoven, The Netherlands, focusing on advancing automated analytical methods used to extract new knowledge from data for HealthTech applications. Other former experiences include the Italian Presidency of the Council of Ministers and the National Research Council of Italy. He also provided ICT consultancy services to Isab, the largest oil refinery in the Mediterranean area. His education and scientific experience fall in the areas of Data Science, Operational Research, Artificial Intelligence, Knowledge Engineering, Semantic Reasoning, and Machine Learning. He is author of several research publications in peer-reviewed international journals, granted EPO and WIPO patents, edited books, and leading conferences in the fields of his work. He is co-editor of the book *S. Consoli, D. Reforgiato Recupero, and M. Petkovic* (2019) and *Data Science for Healthcare: Methodologies and Applications* (Springer Nature).

LUCA PIRAS received the master's degree in computer science from the University of Cagliari, with a thesis titled Dynamic Industry-Specific Lexicon Generation for Financial Forecast. He is currently a Scholarship Researcher with the Department of Mathematics and Information Technology, University of Cagliari. His main research interests include machine learning and natural language processing, with a focus on applications oriented to financial forecasting and recommender systems. He has coauthored several articles about these fields, which were published in scientific journals and international conference proceedings.

ALESSANDRO SEBASTIAN PODDA received the B.Sc. and M.Sc. degrees (Hons.) in informatics from the University of Cagliari, and the Ph.D. degree in mathematics and informatics, supported by a Grant from the Autonomous Region of Sardinia, with a thesis entitled Behavioral contracts: from centralized to decentralized implementations.

In 2017, he has been Visiting Ph.D. Student with the Laboratory of Cryptography and Industrial Mathematics, University of Trento. He is currently a Postdoctoral Researcher with the Department of Mathematics and Computer Science, University of Cagliari, where he is also a member of the Artificial Intelligence and Big Data Laboratory and the Blockchain Laboratory, in which he has been a Technical Supervisor and a Collaborator of numerous research projects. His current research interests include deep learning, financial forecasting, information security, blockchains, and smart cities. To date, he has been the coauthor of several journal articles and scientific conference papers, as well as a Reviewer of different top-tier international journals.

DIEGO REFORGIATO RECUPERO received the Ph.D. degree in computer science from the University of Naples Federico II, Italy, in 2004. From 2005 to 2008, he was a Postdoctoral Researcher with the University of Maryland, College Park, MD, USA. He has been an Associate Professor with the Department of Mathematics and Computer Science, University of Cagliari, Italy, since December 2016. He won different awards in his career (such as Marie Curie International Reintegration Grant, Marie Curie Innovative Training Network, the Best Research Award from the University of Catania, the Computer World Horizon Award, and the Telecom Working Capital, Startup Weekend). He co-founded six companies within the ICT sector and is actively involved in European projects and research (with one of his companies he won more than 30 FP7 and H2020 projects). His current research interests include sentiment analysis, semantic Web, natural language processing, human–robot interaction, financial technology, and smart grid. In all of them, machine learning, deep learning, and big data are key technologies employed to effectively solve several tasks. He is author of more than 130 conference and journal papers in these research fields, with more than 1500 citations. He is co-editor of the book *S. Consoli, D. Reforgiato Recupero, and M. Petkovic* (2019) and *Data Science for Healthcare: Methodologies and Applications* (Springer Nature).