Hindawi
Complexity
Volume 2023, Article ID 5847887, 20 pages
https://doi.org/10.1155/2023/5847887

WILEY | Hindawi

Research Article

Hybrid Human and Machine Learning Algorithms to Forecast the European Stock Market

Germán G. Creamer, Yasuaki Sakamoto, Jeffrey V. Nickerson, and Yong Ren

Stevens Institute of Technology, Hoboken 07030, New Jersey, USA

Correspondence should be addressed to Germán G. Creamer; gcreamer@stevens.edu

Received 1 July 2022; Accepted 13 August 2022; Published 26 April 2023

Academic Editor: Sheng Du

This paper explores the power of news sentiment to predict financial returns, particularly the returns of a set of European stocks. Building on past decision support work going back to the Delphi method, this paper describes a text analysis expert weighting algorithm that aggregates the responses of both humans and algorithms by dynamically selecting the best answer according to previous performance. The proposed system is tested through an experiment in which ensembles of experts, crowds, and machines analyzed Thomson Reuters news stories and predicted the returns of the relevant stocks mentioned right after the stories appeared. In most cases, the expert weighting algorithm was better than or as good as the best algorithm or human. The algorithm's capacity to dynamically select the best answers from humans and machines results in an evolving collective intelligence: the final decision is an aggregation of the best automated individual answers, some of which come from machines and some from humans. Additionally, this paper shows that the groups of humans, algorithms, and expert weighting algorithms have associated with them, particularly, news topics that these groups are good at making predictions from.

1. Introduction

Decision support has at times relied on people's opinions and at other times on computers. As online crowds and communities have shown potential in various tasks, interest has increased in studying the reasons and applications of ensembles of humans, some experts, some not, in performing tasks and making forecasts. A related stream of research has looked at how crowd input can be used to improve machine learning. In the past, machine algorithms were under the control of humans. In some recent work, humans are under the control of machines. Machines request humans to perform tasks; the output of those tasks is used as input to machine algorithms.

The new phenomena present challenges to current theories of information systems. For example, much of information systems have focused on the willingness of users to adopt new technology. In some of the present work on crowds, the focus is on the extent to which computers should adopt the cognitive output of humans, whose performance is variable.

Current research trends can lead to different conceptualizations of decision support, in which humans and computers are both viewed as participants in a complex decision process. Given that humans and computers have different cognitive skills and performances according to the Turing test [1], how should tasks be allocated?

This general question can be addressed through a series of studies. We report the results of one such study here. We chose a cognitively complex task where ground truth exists. Specifically, we picked the domain of market forecasting, and within that, we asked humans and computers to predict future stock prices based on past prices and news stories. Neither humans nor computers can perform this task with a high degree of accuracy [2]: indeed, some argue that such markets are random walks [3]. Many other empirical studies [4–6] have already demonstrated that qualitative variables affect stock prices as there is a psychological element at play: prices are not based solely on the expected present value of their future cash flows. For instance, there is an association between trading volume and message activity in an Internet

chat room [7]. Also, trading volume is associated with the sound levels in the Chicago Board of Trade proprietary traders pit [8].

Additionally, many recent studies [9, 10] have shown that investors react at different speeds to the release of news information, so it is at least, in theory, possible to make predictions about how investors will respond to news shortly. This task is complex because it requires understanding news stories and how they may violate investors' expectations, leading to changes in investment strategies. Investing experience and domain knowledge of markets, companies, behavioral economics, and psychology may be helpful.

In sum, the task is one of anticipation that depends on various skills. Success on the job can be readily measured. Thus, it provides a way of understanding the relative strengths of people and machines in a data-rich domain. This understanding may, in turn, lead to new and different architectures for decision support systems related to forecasting markets. The findings may also provide a basis for further studies to determine how such architectures might work on different kinds of problems.

The complementarity between humans and machines is that humans contribute with unique data given by their expertise, knowledge, memories, intuition, association, analysis, and the capacity to learn from unrelated tasks. In contrast, the machines can optimize a function using the available data. Hence, it is also possible to design a two-stage algorithm: in the first stage, the input of humans and different machine learning algorithms are calculated, and in the second stage, an algorithm optimizes the proper combination of humans and machines based on specific performance criteria. As a result, the combination of humans and machines should enrich the decision process. The results might be at least about the same as the best individual forecast, even though no theoretical support leads to this outcome. In principle, humans may excel in tasks with limited machine-readable data, while machines may perform very well in jobs that require a large amount of data where the predictive function should be calibrated to particular dimensions such as different time horizons.

The other relevant aspect that motivates our human-machine ensemble algorithm is that any ensemble algorithm benefits from the diversity of its experts, either humans or machines. The aggregation of very different experts helps to manage the tradeoff bias-diversity avoiding a bias in a certain direction when all the experts are very similar [11–13].

As several studies [14] have shown that machine algorithms can perform at least as weak learners, their predictions are slightly better than chance; then, it is possible to minimize the prediction error when the diversity of an ensemble of weak learners is maximized as their errors mutually cancel out.

In summary, the strength of ensemble algorithms is in the diversity of prediction models and the data used. As humans build their mental models, the models of crowd members might differ according to their demographic origin.

The hypothesis that computers can classify financial news has received strong support in academic literature [15–17]. However, these studies have only used natural language processing (NLP) methods to classify news without incorporating the input of humans. To the best of our knowledge, only a few papers mix computational linguistic methods with the inputs of humans. Archak et al. [18] combines NLP techniques with the crowd inputs to develop an econometric model that evaluates the influence of textual product reviews on decisions to buy a particular type of product. Also, Ipeirotis and Gabrilovich [19] proposed a system that assesses users and acquires knowledge from them and selects the best for specific tasks. From our perspective, this is equivalent to choosing the best workers in a crowd [20, 21]. But none of these studies considers the possibility of incorporating crowdsourced sentiment classification into an intelligent system integrating human interpretative capabilities with computer processing [22] to predict financial returns. Hence, this paper's critical challenge is bridging the gap in our understanding of machine-generated and human classification, considering their diverse origins. In particular, we ask the following:

(1) Can a learning algorithm that combines the output of machines and humans improve a prediction?

(2) Considering that humans (individuals and crowds) and machines have different cognitive skills, do they focus on other characteristics when analyzing and classifying news stories?

This paper addresses these questions with the development of a new text analysis expert weighting algorithm that aggregates, according to previous performance, the forecasts of both humans and algorithms based on several automated text analysis methods that we describe in the appendix:

(1) About our first question, we compare the performance of different human groups and several machine learning methods using different text analysis methods to classify the news sentiment. NLP studies typically use labeled datasets by humans or other mechanisms as the gold standard to develop and test new algorithms. Our study uses several human groups and machine learning algorithms to classify news, back tests these classifications with historical asset returns and evaluates their performance with several performance measures.

(2) About our second question, we build a network of humans and machine learning algorithms based on the number of the most relevant and common topics among the different groups. We cluster the groups in communities and compare them according to the topics they predict best. Our experiments show that humans and machine learning algorithms naturally separate into communities that predict better when presented with topics associated with each community.

These are the additional contributions to the literature of this paper:

(1) We propose a novel method that automates several dimensions of text analysis to forecast assets returns: multiple features that include frequency of n-grams (bag-of-words), parts of speech, keywords preselected by a dictionary associated with emotions, semantic frames that interpret the meaning of short phrases, and common topics among the news. Most of the previous studies in finance or accounting follow either a bag-of-words or a dictionary approach. However, our method captures several aspects of the news using a combination of several text analysis methods.

(2) We rank the different text analysis methods and demonstrate that the combination of bag-of-words, keywords of a dictionary, and semantic frames show important improvements compared to the use of singular methods to predict asset returns.

(3) We also propose the aggregation of many features, machine learning algorithms, and human groups with a hybrid expert weighting algorithm capable of combining different forecasts according to their performance. This approach is also related to the recent work of Archak et al. [18] and Bajari et al. [23] and to extensive econometric literature that combines independent predictions into a single forecasting model [24–26]. However, our approach uses a dynamic model capable of making online decisions as required by large and heterogeneous datasets generated by the news and social media. As far as we know, previous finance studies have not simultaneously explored a variety of text analysis methods and algorithms and optimized them dynamically.

(4) Our final results show that selecting features generated by the bag-of-words, dictionary, and semantic frames, processed by many algorithms and integrated with the human classification using the expert weighting algorithm can obtain significant correlations with assets returns in most cases. Even though the machine learning algorithms can process large amounts of information adjusted to several time horizons, the combination of their outputs with the human classification simulates an artificial collective intelligence capable of producing at least similar or better results than any individual methods explored.

2. Background

2.1. Machine Readable News and Sentiment Analysis. There are many reasons why decision-makers may want to classify news automatically. This has created a market for machine-readable news; for example, Thomson Reuters provides a news feed specifically intended to be read from computers and its interpretation of the sentiment of this news. Many have attempted to parse these and other electronic news sources in academia. For example, one study showed that a proportion of negative words (media pessimism) in sections of the Wall Street Journal (WSJ) and the complete WSJ and Dow Jones News Service [9] are associated with a tendency for prices to fall. Along these lines, certain words of corporate annual reports such as "risk" are associated with low returns [27] and also finds that positive or negative words of earnings press releases are related to return on assets [28]. Positive and negative sentiments have a corresponding influence on stock price [29]. These studies are a few examples of a growing literature that evaluates the economic or financial effects of media [30–36] and social media [37–41].

Other attributes of news stories also impact the price of securities. News about monetary issues significantly affects daily stock prices while unemployment, inflation rate, and real activity news have no significant influence on prices [42]. Only the first four most recent news of each news category have a high impact on 5-year T-note prices [43]. The duration of the price impact of positive news is shorter than the impact of negative news [44]. Traders underreact to released news [45], and investors overreact to old information [10].

2.2. Crowdsourcing. Since sentiment analysis aims to predict how people will react to news stories, it is sensible to involve people in the process. Human classification processes [46–48] and NLP methods [49–51] are also central in the field of cognitive science. However, such studies aim at understanding how humans think and are less interested in replicating human behavior on complex social problems. Consequently, cognitive psychologists usually study individuals' cognitive processes while performing simplified tasks. Doing so can reduce the effect of prior knowledge on an individual's task performance. But such research is difficult to generalize to real-world tasks such as classifying news sentiment and predicting market trends.

Despite this disconnect, several studies in cognitive science relate to the work reported here. For example, Griffiths and Tenenbaum [50] found that human predictions about the duration of everyday events followed a pattern predicted by Bayesian models and concluded that humans made complex inferences using prior distributions. This conclusion was challenged by Mozer et al. [52] who demonstrated that heuristic-based reasoning depending only on a few instances (1–3) could also explain Griffiths and Tenenbaum's [50] results. Mozer et al. [52] proposed that the "wisdom of the crowd" could reconcile this seeming contradiction: the combination of many decisions of individuals with limited knowledge (as opposed to prior distributions) is consistent with a Bayesian perspective.

Taken together, the findings of Griffiths and Tenenbaum [50] and Mozer et al. [52] suggested that crowds with limited knowledge can collectively predict like an expert with greater knowledge [53, 54]. The effectiveness of crowds has also been demonstrated in more open-ended tasks of generating text ideas [55] and design sketches [56]. Then crowdsourcing may be a viable approach to classifying news sentiment. Exploring the ability of crowds to classify news sentiment, which is less open-ended than idea generation but more open-ended than predicting a correct value of a thing, will contribute to research into the usefulness and organization

of crowd work [22, 57, 58]. Moreover, the resulting data, the classification performed by crowds, may provide insights into sentiment analysis' cognitive processes.

2.3. Sentiment Analysis by Humans and Algorithms. These different studies use either human-labeled datasets or NLP methods for sentiment classification. Although a combined human-machine approach has been used to improve business decisions in prediction markets [59], this literature has not explored a hybrid human-machine approach to classify news sentiment and its predictive power on asset return as we proposed in this paper. Our method, an expert weighting algorithm, aggregates the prediction of humans with the prediction of several automated text analysis methods described in the appendix.

Our paper explores the power of news sentiment to predict financial returns, particularly the returns of a set of European stocks. We find sentiment in several different ways: first, as defined by the best performers of the crowd; second, by a group of trained evaluators; third, by a finance expert; and fourth, by several learning algorithms. We build a hybrid expert weighting algorithm that aggregates the automated responses of both human groups and/or machine learning algorithms according to their past performance. That is, we create an ensemble learning algorithm [60, 61] that uses both human and computer classification histories. We compare the effectiveness of these different groups. We then seek to understand how humans and machines differ by examining their relative success in various news stories.

3. Decision Support for Forecasting

This section introduces the main analytical methods used in this research; the appendix includes a formal description of the algorithms used.

3.1. Automated Text Analysis. The prevailing method of analyzing text is bag-of-words (BoW) due to its robustness and simplicity [62]. The implementation of BoW requires the construction of dictionaries for the particular documents explored. A different approach consists of classifying every word as positive or negative according to a given lexicon or dictionary [63]. The percentage of positive words determines the overall sentiment of a news or a document among all sentiment words. When Zhang and Skiena [64] apply this approach to forecasting stock price movements, they find a profitable trading strategy as the news sentiment scores are highly correlated with stock returns. Bollegala et al. [65] and Godbole et al. [63] showed that the selection of appropriate lexicons for different tasks may improve the performance of this method. For instance, the application of dictionaries to evaluate the positive or negative sentiment of risk disclosure forms [66–68], of the MD&A section prior SEC filings [69], and of earnings announcements [70] required the development or use of specialized dictionaries. In this regard, Wuthrich et al. [71] developed a keyword-based system to forecast major market indexes. A different application of dictionaries is to generate new features by the emotional meaning of every word and use them as inputs of a supervised learning system as we use it in this research. The dictionary of affect in the language (DAL) [72] evaluates the emotional meaning of words or text using 8,742 words that are evaluated on a scale of 1 to 3 along the dimensions of pleasantness (P.I.s), activation (Act), and imagery (Img). Agarwal et al. [73] applied the normalized average of the DAL scores of different parts of speech (POS) for sentiment analysis, and Xie et al. [17] used this sentiment for stock price prediction.

The main problem with BoW and any dictionary is that they do not identify the sentiment's target. For example, "*X* beats *Y*" is positive for *X* but is negative for *Y*. BoW cannot provide such information. Since different words in different domains may present different sentiments, it is also impossible for BoW to identify the causality among the variables studied without semantic information. Since BoW uses every word as a feature, a large dataset generates an extensive feature set and much irrelevant information. Therefore, an appropriate method to extract useful information, such as the frame semantic approach, is still critical to calculate the sentiment [62, 74, 75]. Frame semantic parsing of a document refers to the automated task of finding semantic targets, disambiguating their semantic frames that refer to particular events and identifying their frame elements. For instance, we include these two examples about France Telecom:

(1) France Telecom will charge 10 cents more for its service.

(2) The government charged France Telecom with fraud.

The core word in these examples is "charge." The main problem for a computer is discriminating the different meanings of "charge" in these two sentences. In this respect, semantic analysis can provide further information to understand the differences between these two sentences. Below are the parsed results based on frame semantics [76] where the frame elements are in parentheses:

(1) France Telecom (Seller) will charge (Commerce_collect) 10 cents more (Money) for its service (Reason).

(2) The government (Arraign_authority) charged (Notification_of_charges) France Telecom (Accused) with fraud (Charges).

The first sentence is about fee collection, while the second is about a lawsuit. Beyond the different meanings obtained from the semantic analysis, we can also identify the action initiator and action target. France Telecom is the action initiator in the first sentence and may have more income. In the second sentence, France Telecom is the target of a lawsuit, and its operations can be negatively impacted.

Frame semantics, as illustrated above, presents a sentence with a "frame" structure. The frame describes the type of sentence, entities in the sentence, and roles and interactions of the entities. The keywords in a sentence evoke a particular frame, and the roles of other words are identified. In general, the knowledge provided by frame semantics facilitates understanding one sentence.

The theory of frame semantics has motivated the development of ontologies such as the FrameNet lexicon [77] that acts as a repository of semantic frames (S.F.) and their frame elements. In January 2015, FrameNet had more than 10,000 word senses, 170,000 manually annotated sentences, and 1,200 frames. Each frame contains a list of words that evoke the frame. The words are defined as lexical units. For example, in the Commerce_buy frame, "Buy," "Purchase," and other similar words are considered lexical units. Frame elements describing the roles of words in a sentence are also included in each frame. Following the same example of the Commerce_buy frame, "Seller," "Buyer," and "Goods" are the core frame elements, while "Place" and "Purpose" are noncore frame elements. The FrameNet also describes the relationships between frames and the relationships between roles. For instance, the inheritance relation indicates which frame is the most general.

FrameNet provides the annotated information for further processing. An automatic semantic parser is required to analyze, identify targets, and label a new sentence, such as the rule-based system SEMAFOR proposed by Das et al. [78]. SEMAFOR uses a latent-variable log-linear model to identify semantic frames and a probabilistic model to determine their frame elements. In this research, we used SEMAFOR to identify semantic frames (F) with their targets (T) and frame elements (E) of every news item, considering that this system achieves a very high public performance with 0.9 precision.

After eliminating stop words, this research extracts features from news using mostly bag-of-words, the dictionary of affect in language by part of speech, and elements of semantic frames. We combine these features in large sparse matrices to predict asset return direction using the unsupervised and supervised methods described in the appendix. We introduce a text analysis expert weighting algorithm in the next section.

3.2. Prediction Markets and the Delphi Method. We propose a hybrid human-machine system based on the Delphi method and prediction markets, in which crowds, experts, and machine learning algorithms contribute their responses to forecasting. Prediction markets are based on aggregating orders to buy and sell on future events. There is an extensive list of successful examples of prediction markets, such as the Hollywood Stock Exchange to predict returns of movie box offices, the Iowa Electronic Market to forecast political events, and NewsFutures' World News Exchange to anticipate future events. These prediction markets assume that the aggregation of the information of a group or the wisdom of crowds can help predict the probability of an event successfully. The Delphi method is much more directed than prediction markets. It relies on a panel of experts who predict an event and has access to the predictions of each other and then predict again. This process can be repeated several times, and as a result, the forecasts tend to converge (Figure 1(a)). Dalkey and Helmer [79] describe its use to estimate the number of A-bombs that would target particular U.S. industrial sectors.

3.3. Text Analysis Expert Weighting Algorithm: Aggregation of Individual Forecasts. Our algorithm, presented in Figure 2, is a text analysis expert weighting algorithm that aggregates several experts' forecasts to predict asset returns. Its output should be similar to or better than the best individual predictions. As with the Delphi method, it collects the inputs of several experts; however, instead of returning the individual forecasts to all the experts, it combines all the predictions into a single prediction based on the past performance of every expert, as described in Figure 1(b).

To simplify the presentation of our online learning algorithm, we introduce the case of one asset, which can easily be extended to N assets. In this research, every expert is a human or a machine learning method calculated with the training set. We refer to the sequence of experts by $\psi = \psi_1, \psi_2, \ldots, \psi_E$ where E is the number of experts.

$y_t \in \{0, 1\}$ is the binary label to be predicted where 1 represents the expectation of a positive sentiment, return or output, and 0 otherwise.

The outcome of expert ψ_i at time t is the prediction h_t^i. The final score S_t at time t is obtained by summing up the experts' predictions weighted by their past performance using an exponential weights function as introduced by Littlestone and Warmuth [80] and Cesa–Bianchi et al. [81]. If this score is greater or equal to 0.5, we consider this a positive sentiment; otherwise, it is non-positive.

The exponential weights formula provides a different way to compute the posterior distribution from Bayesian techniques. An interesting feature of this algorithm is that it abstains from predicting certain instances; as a result, the predictions that it makes are very reliable.

Creamer and Freund [82] and Creamer [83] proposed the application of learning algorithms and especially the expert weighting algorithm to forecast price trends and discover new trading strategies. This approach is also related to the pioneering work of Bates and Granger [24] that combines independent predictions into a single forecasting model. We want to know if a learning algorithm that combines the output of machines and humans improves a prediction. To evaluate this question, we use news sentiment of crowds, experts, and machine learning algorithms to predict the direction of assets return. We expect that our expert weighting algorithm will outperform the individual experts. Still, it is always possible that a weighting algorithm may overfit past performance and thus do more poorly than the individuals.

4. Experiment in Forecasting

4.1. Data Collection. Using the following T.R. topic categories, we select a stratified sample of 1,000 news stories associated with STOXX50 companies from the T.R. News Archive for 2005: (1) Cross-Market, (2) Fixed income, F. X. and Money Market, and (3) Central Bank, Economy, and Institutions. These topics were the most relevant categories to generate a stratified sample of news. The single "Cross-Market" category, which includes a large variety of corporate news, represents 69% of the news, while the categories "Fixed income, F. X. and Money Market" and "Central Bank,

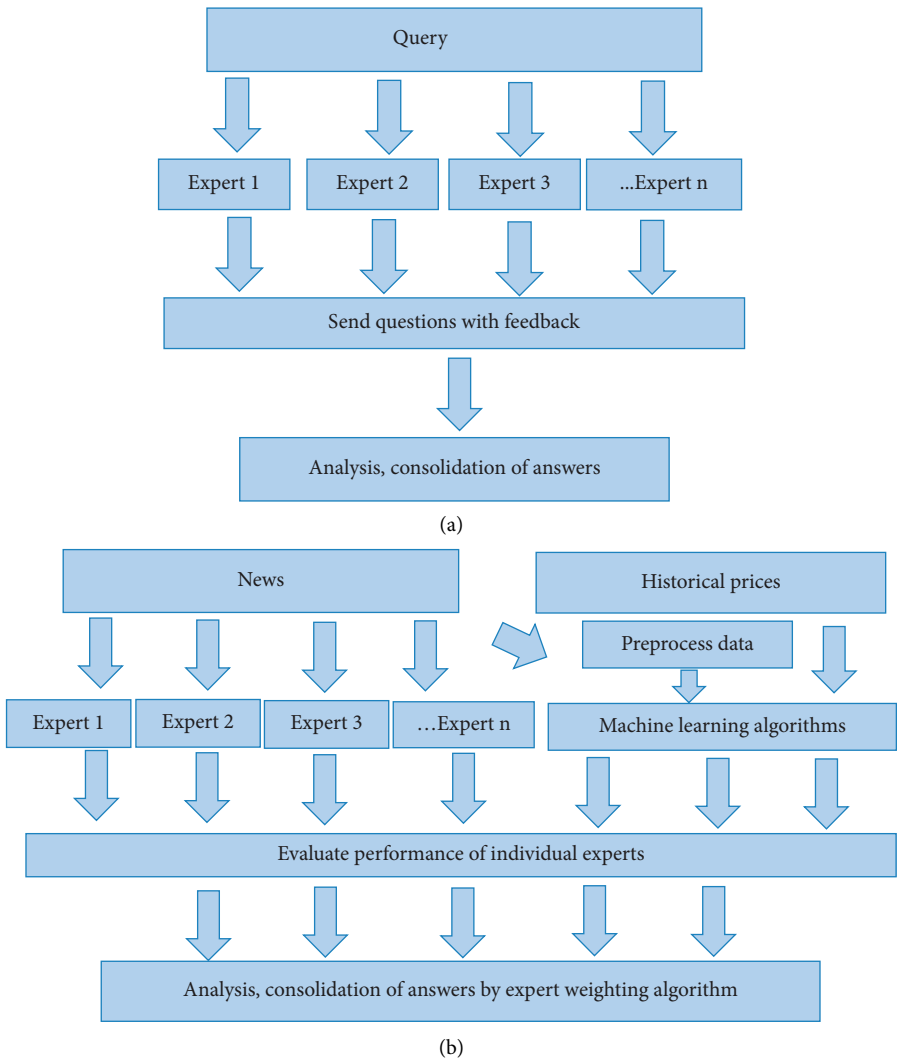

Figure 1: Delphi method (a) and expert weighting as an extension of the Delphi method (b).

Economy, and Institutions" are associated with 5% and 4% of the news, respectively. The rest of the news stories are associated with more than one category; however, the "Cross-Market" category is included in most cases. The final number of news stories that we use was 743 after we eliminate news that did not have proper prices, incomplete news, news containing market reports, and news summaries that simultaneously included information from many companies. We match the timestamp of the news with the timestamp of the most relevant associated assets' prices. The news stories are labeled as positive when the asset return is positive and 0 otherwise, after 1, 5, and 15 minutes, and 1, 2, and 3 hours. We associate every news item to a particular company using the Reuters Instrument Codes (RIC) based on the first main sentence of the news, which in most cases was the first RIC in the field Related_RICS of T.R. news archive; however, we verified that the news was about the selected RIC. Whenever a news is mentioned, first the company of an external analyst or a manager that issued an opinion or shared news about a particular company, we select the target company instead of the analyst's company. We extract the text of every news item from the T.R. news archive field Take_Text, and we substitute any missing data of this field with the corresponding header (HEADLINE_ALERT_TEXT).

STOXX50 represents the 50 most influential companies by the level of capitalization. We select the year 2005 as we want to avoid the market overreaction to any news related to the financial crisis of 2007.

We also explored the T.R. news analytics that provides news sentiment scores (T.R. sentiment) with positive, neutral, and negative values. The T.R. sentiment is associated with specific assets, based on NLP algorithms, and validated by human experts. Its timestamps align with the assets' price time series of T.R. tick history. For this review, we select the period 2003 to 2012 to introduce the overall relationship between T.R. news sentiment and financial or economic events, as demonstrated in the following section. However, we do not compare the T.R. sentiment with our learning algorithms due to T.R. sentiment's proprietary nature.

Input:
Set of news or media related to a particular asset where each observation has a unique timestamp t.
Asset returns r_t for each period t or humans that label the news.
C is an exogenous parameter for expert weighting. The default value is 1.
γ is the score threshold with a default value of 0.5.
Train with several machine learning algorithms:
1. Select a representative asset from the targeted market.
2. Collect news or social media directly associated with this asset.
3. Label (y_t) each observation with the asset return (r_t) trends of the next period where 1 represents the expectation of a positive return, and 0 otherwise.
4. Preprocess the media by eliminating stop words, generating n-grams, and selecting text indicators according to the text analysis method.
5. Include all features and appropriate labels in the training and test sets.
6. Every d periods, train a new group of learning algorithms (experts) based on the training set.
7. Every d periods, recalculate the test set, get every expert's prediction, and weigh every expert's response in the next steps. The learning algorithms use the training set to build their models, and all the experts (learning algorithms and/or humans) predict based on the test set. Call the experts as $\psi = \psi_1, \psi_2, ..., \psi_E$ where E is the maximum number of the most recent experts.
Expert weighting algorithm:
8. Calculate the weight of expert ψ_i at time t according to the cumulative return of expert i at time $t-1$ (expert weighting (EW)):
if t=1, $w_1^i \doteq \exp(C\ I_1^i\ r_1)$
else $w_t^i \doteq \exp\left(C\ \sum_{s=1}^{t-1}(I_s^i\ r_s)\right)$
where $I = \begin{cases} 1 & \text{if } h_t^i = y_t \\ 0 & \text{Otherwise} \end{cases}$
r_t is the abnormal asset return at time t.
h_t^i is the hypothesis of expert i at time t.
9. Calculate the experts' score at time t as $S_t = \frac{\sum_i w_t^i\ h_t^i}{\sum_i w_t^i}$.
Output:
Expert weighting prediction is:
$H_t \doteq \begin{cases} 1 & \text{if } S_t \geq \gamma \\ 0 & \text{Otherwise} \end{cases}$

Figure 2: Text analysis expert weighting algorithm.

4.2. Characteristics of Data. Since June of 2006, the proportion of T.R.'s positive news stories about STOXX50 companies has decreased while the STOXX50 index continued to increase (see Figure 3). Its peak was in June 2007, when the financial crisis started in the U.S. After this month, the STOXX50 only lost value until March 2009, when the U.S. economy began its recovery period. In August 2008, just before Lehman Brothers' bankruptcy, the proportion of positive news reached its minimum point. During the following period, the proportion of positive news increased, although with a high level of volatility consistent with the dynamic of the financial crisis. The correlation between one-day lagged positive news and the STOXX50 index from January 2005 until June 2006 is −0.33. During the most challenging period of the crisis (June 2007 until March 2009), this correlation is 0.49, and during the recovery period (after March 2009), the correlation is 0.43.

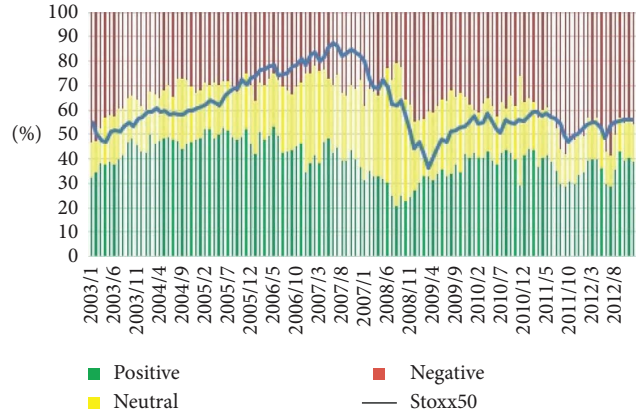

Figure 3: Proportion of T.R. news sentiment (positive, neutral, or negative) and STOXX50 index: 2003-12. The STOXX50 index is scaled to be comparable with the sentiment's distribution.

Figure 3 indicates that negative and positive sentiment distribution is almost symmetric. The correlations between the number of daily positive and negative news stories before, during, and after the financial crisis of 2008 are −0.44, −0.21, and −0.65, respectively. The correlations between the STOXX50 index and the lagged non-positive (neutral and negative) news stories during the above periods are 0.33, −0.49, and −0.43, respectively. The correlations for the one-day lagged negative news stories during the same periods are −0.18, −0.41, and −0.22, respectively. The non-negative values are larger than those of the negative news. Considering this symmetric behavior of the news sentiment and the additional information captured by the non-positive news, we evaluate only the binary classification between positive and non-positive news sentiment as a market sentiment indicator instead of assessing the three sentiment categories: positive, negative, and neutral. Additionally, the reverse of the sign of the correlation observed by the number of positive news stories and non-positive news stories before and during the financial crisis of 2008 indicates that news sentiment can also be used to anticipate major market movements.

4.3. Procedure. Our framework to forecast financial markets trends using text analysis automates and aggregates the answers of several machine learning algorithms and humans are presented in Figure 4 and the following sections.

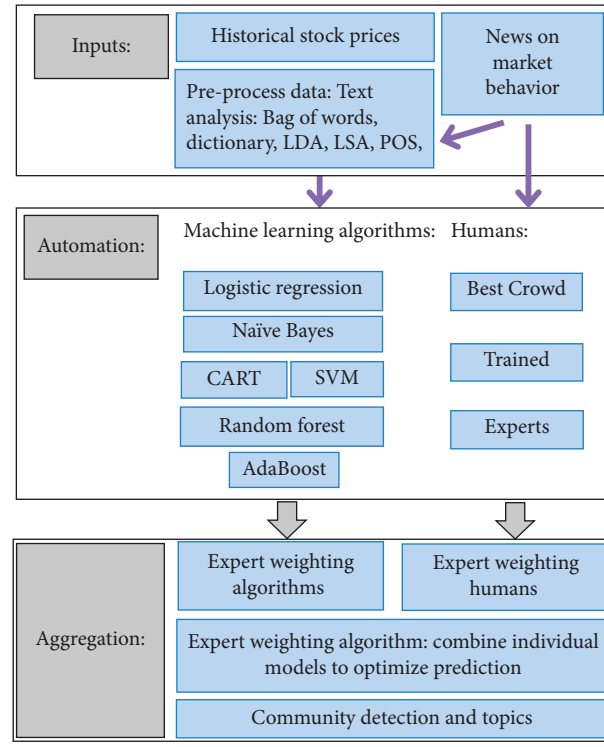

Figure 4: Expert weighting combining humans and machine learning algorithms to preprocess and classify news.

4.3.1. Automation. In this paper, automation of financial news classification includes preprocessing and classifying news according to humans and learning algorithms. In a fully automated system with a large flow of news, the human component may either disappear or become very small as it may have been converted into rules of the final computational model.

The human participants evaluated 743 news stories in batches of 50 news stories each. To reduce fatigue, the participants were obligated to wait at least one hour to examine a new batch if they read more than one batch. News stories were classified as positive and non-positive based on overall sentiment. The following groups performed the classification:

(1) Best crowd: This is a selected group of workers from the crowd that in the first 70% of the news (training data set) have an error rate below the median of all the workers when their sentiment scores were used to forecast the asset return trends. This group is recalculated for every time horizon (1, 5, and 15 minutes and 1, 2, and 3 hours) used to predict asset return. The crowd is based on the majority vote of 9 Amazon Mechanical Turk workers (workers) per news story. This information was collected in July 2012. We also requested basic information about each worker: age, gender, city, state, zip code, and if the worker considered themselves a finance expert. The workers had a minimum HIT approval rate of 95%, they were U.S. residents as all the news were in English, and they were compensated with Dollar 0.01 for the classification of every news story.

(2) Trained evaluators: classification based on the sentiment of 3 evaluators with training in finance.

(3) Expert: an evaluator with advanced training and professional experience in finance.

The final classification of each news story is based on the majority vote of each group.

As a basic sanity check, we observed that about two-thirds of the T.R. sentiment, an industry standard, is the same as the sentiments of the crowd (67%), trained evaluators (65%), and the expert (62%).

After eliminating stop words, we use the following methods introduced in Section 3.1 and a combination of them to extract quantitative features from the news:

(1) Bag-of-words (BoW): tf and tf—idf of 1-, 2-, and 3-grams for all words.

(2) Dictionary of affect language by part of speech (PDAL): P.I.s, Act, and Img scores for all words (All), verbs only (V.B.), adverbs only (R.B.), adjectives only (J.J.), and nouns only (N.N.).

(3) Latent Semantic Analysis (LSA): k dimensions selected (see Section A.1).

(4) Latent Dirichlet Analysis (LDA): topics generated (see Section A.2).

(5) Part of speech (POS): tf and tf—idf of 1-, 2-, and 3-grams of the parts of speech tags as described in PDAL.

(6) Semantic frames (SF): tf and $tf—idf$ of 1-, 2-, and 3-grams of semantic frames (F), targets (T), and frame elements (E).

We also combined these methods with and without BoW due to its simplicity and predictive capacity in text analysis studies. The default number of topics used for LSA and LDA is 100.

We forecast the asset return trend using the above features and the following learning algorithms explained in the appendix because of their different methodological approaches to classification:

(1) Logistic regression (LR): a well-known linear regression algorithm used as the baseline algorithm.

(2) Naïve Bayes (NB): a Bayesian parameter estimation problem based on some known prior distribution.

(3) Support vector machine (SVM): classifier based on a linear discriminant function.

(4) CART: decision tree or nonparametric method that uses nominal or continuous data for classification.

(5) Ensemble methods are as follows:

  (a) AdaBoost: an ensemble method that minimizes bias.
  (b) Random forests (RF): an ensemble method that minimizes variance without increasing bias.

(6) Expert weighting algorithms are as follows:

  (a) Expert weighting (EW): it combines the output of the machine learning algorithms and the relevant human groups under study. The weights of the learning algorithms and human groups depend on their past cumulative return. This algorithm follows the research proposed by Archak et al. [18] that combines features generated by text mining techniques and the crowd in an econometric model, and the perspective of Bates and Clive [24] that combines independent predictions into a major econometric model.
  (b) Expert (algorithms) weighting (EWa): EW is only based on the combination of the learning algorithms.
  (c) Expert (humans) weighting (EWh): EW is only based on the combination of the human groups.

4.3.2. Implementation. We used the first 70% of the observations (520) as the training data set and the remaining 30% (223) as the test data set to predict the asset return trend after 1, 5, and 15 minutes and 1, 2, and 3 hours. We split our test data set into ten subsets sorted by date, which were used to calculate the average of the performance indicators. Every subgroup has about 21 observations (like a month based only on trading days), as few observations were lost due to initialization issues. We used the Stanford POS tagger to tag the POS of the news, the NLTK Python package to preprocess the data, the gensim package to run the LDA and LSA models, and the Scikit-learn package to train the classification algorithms with the default values. After several tests to avoid overfitting and find the best performers, we run a linear SVM, AdaBoost with 50 iterations, CART with 200 minimum samples required to be at a leaf node and a maximum depth of 3, and NB for multivariate Bernoulli distributions with an alpha smoothing parameter of 0.01 (minimal smoothing). We applied the expert weighting algorithm (Figure 2) to combine our forecasting methods according to their performance to generate one weighted sentiment prediction.

Unlike other papers such as Gao et al. [84] that use accuracy to evaluate the forecast of stock market direction, we select the Matthews correlation coefficient (MCC) [85] as our performance measure for binary classification. MCC can manage unbalanced datasets characterized by a significant difference in the number of observations assigned to each label. In our case, the positive labels represent from 39% to 54% of the complete test sample depending on the time horizon, so using MCC seems appropriate.

We compared the difference between the MCC of each group and algorithm against logistic regression, the baseline algorithm, using the t-test of the mean difference. We also run the ANOVA procedure to compare performance differences among humans, individual learning algorithms, and EW algorithms. This comparison helped us to evaluate the first main question of this paper: Can a learning algorithm that combines the output of machines and humans improve a prediction?

The formula of MCC or of the ϕ coefficient as it is also known is $|\mathrm{MCC}| \doteq \sqrt{(\chi^2/n)}$, which is equivalent to the Pearson correlation coefficient for binary cases. MCC can also be calculated from the confusion matrix using the following formula: $\mathrm{MCC} \doteq (\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}) / \sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}$, where n is the number of observations and TP, FP, TN, and FN stands for true positive, false positive, true negative, and false negative observations, respectively. When the two classes have the same number of observations, extreme values of 1 or −1 are expected. A value of zero implies an accuracy of 50%.

4.3.3. Aggregation. As a stress test, we evaluate the capacity of the human and algorithm predictions (without retraining) to detect the top 30% returns of the test set even when the algorithms were not trained to recognize these extreme events.

Using random forests with 100 trees, without a limit on its size and with a limit of ten features to evaluate the best split, we extracted the top 5% T.R. topics (4) for all our classifiers (humans and algorithms) that were relevant to predict: (1) return direction, and (2) the top 30% returns. We selected the topics with the highest random forests' importance score, which is calculated based on the impact of each feature on the out-of-sample prediction accuracy.

With these results, we built undirected networks for each type of return. The nodes of these networks are the classifiers, and the edges are the number of common topics among each pair of classifiers weighted by the random forest's

Table 1: Weighted average error rate, average response time, and frequency by number of answers between the crowd sentiment and the 1-minute asset return trend. Average response time is in seconds.

Number of answers	1	2–5	6–10	11–20	21–50	51–250	251–750
Avg error rate (%)	61.11	47.01	52.44	50.00	50.47	54.79	49.97
Avg response time	44.44	35.54	36.90	36.09	25.44	23.00	17.63
Frequency	36	42	11	14	16	6	9

importance score. The most common topic across all the classifiers is excluded from the network calculation to eliminate unnecessary links.

We clustered the network of humans and algorithms using the community detection method based on the greedy optimization of the measure known as modularity [86]. Modularity measures the quality of a partition that evaluates if there are many links within communities and few links between communities.

We selected the most frequent topics associated with every community to evaluate the second main question of this paper: do humans (individuals and crowds) and machines concentrate on different topics when they analyze and classify news stories? This analysis has two steps: (1) to evaluate if the new communities or clusters are consistent with our main groups of analysis (humans, machines, and expert weighting algorithms), and (2) to recognize specific topics for these communities that may indicate differences in their decision-making process.

5. Results

5.1. Crowd. The crowd consisted of 134 workers from 38 U.S. states and had a mean age of 37.7 ($\sigma = 11.4$) and was 53% female.

A logistic regression analysis shows that the following workers' variables do not affect the error rate when the sentiment is used to predict the associated asset return trend after 1 minute: gender, state, average response time, and if the worker considers himself/herself an expert. However, age and the number of answers per worker are directly associated with the error rate. In the complete sample, the overall error rate weighted by the number of responses of each worker is 50.5%. This value decreases when the worker evaluates at least five news and reaches its minimum of 47% with 2–5 news stories. The error rate slightly increases to 52.44% with 6–10 news stories, and after ten news stories, the average error rate is 50%. The average response time decreases with the number of news stories processed (see Table 1).

The crowd results improve after eliminating casual workers (less than two news stories evaluated) and workers that answer too fast. Considering the effect of casual workers and that the average error rate of the crowd is close to random, we eliminate the casual workers when we select the best workers from the training set to classify the news in the test set. We refer to this group as the best crowd.

5.2. Learning and Expert Weighting Algorithms. Our tests show that BoW-SF-PDAL (bag-of-words with semantic frames and dictionary of affect language) is the top text analytics method according to SVM and EW, our best individual and hybrid algorithms, respectively. This combined method captures different aspects of the news: PDAL includes relevant words that might be associated with the market sentiment, BoW captures most of the combined effects of the terms of the news, and SF captures the basic meaning of each phrase. Additionally, the expert shows the most consistent forecast (lowest volatility). All the human groups perform better within the 2 hours forecast. SVM and all the EW algorithms offer their best performance with the most extended horizon of three hours. The differences in performance across the classifiers and the periods for the BoW-SF-PDAL case are significant according to the ANOVA p-value of Table 2. Therefore, we only present the results of our simulations with the three hours horizon and BoW-SF-PDAL as our preferred method.

SVM is the dominant algorithm to forecast the 3-hour asset return and is significantly different from logistic regression, our baseline algorithm, random forest, AdaBoost, and naïve Bayes. Although SVM is slightly better than CART (see Figure 5(a)), the standard error of CART is the largest of this group of algorithms. The expert and the best crowd have similar performance among the human groups, while the trained evaluators show a lower MCC (Figure 5(b)).

About our first question, every expert weighting algorithm is similar to or better than its components. All of them are better than logistic regression, while EW and EWa show a better performance than EWh (see Figure 5(c)).

Most of the methods explored show a deterioration of MCC when the original classification is used to forecast the direction of the top 30% return. AdaBoost, followed by naïve Bayes, are the best algorithms and the only methods that significantly improve their performance. The expert outperforms the rest of the humans, and EW and EWa are still the dominant algorithms (Figure 5).

The differences of performance across each main group (humans, individual learning algorithms, and EW algorithms) and levels of returns (return, and top 10%, 20%, 30%, 40%, and 50% return direction) are significant according to the ANOVA p-value of Figures 5(a)–5(c). Likewise, the performance differences across all the groups and all the returns are also significant.

Assuming that every algorithm is associated with a cognitive style or a learning procedure, the expert weighting algorithm (EW) that combines the forecast of the individual algorithms and humans can be considered an artificial collective intelligence that can process a complex and large amount of information. However, the forecasting capability of the algorithms is significantly affected by the quality of the inputs used. As mentioned above, the procedure BoW-SF-

Table 2: Average MCC for asset return prediction by time horizon and learning algorithms using BoW-SF-PDAL. Even though the table only includes the best performing individual algorithm, SVM, and the ANOVA p-values are based on all the algorithms and human groups.

	SVM	Trained	Expert	Best crowd	EW	EW algorithm	EW human
1 minute	0.03	0.06	0.08	0.06	0.04	−0.03	0.07
5 minutes	0.10	0.09	0.07	0.09	0.10	0.03	0.10
15 minutes	0.09	0.04	0.05	0.00	−0.03	0.03	0.04
1 hour	0.08	0.03	0.06	0.00	0.10	−0.04	0.04
2 hours	0.09	0.11	0.10	0.12	0.09	0.11	0.11
3 hours	0.15	0.05	0.08	0.08	0.15	0.16	0.11
Mean	0.09	0.06	0.07	0.06	0.08	0.04	0.08
St. deviation	0.04	0.03	0.02	0.05	0.06	0.08	0.03
ANOVA p-value:	Rows: 0.011		Columns: 0.002				

PDAL includes features associated with the emotions of every word (PDAL) and the meaning of basic phrases (S.F.) that substantially improve the learning algorithms' forecasting capability. When the EW algorithm is calculated using BoW alone, MCC is 0.14. This value increases to 0.2 when BoW-SF-PDAL is used instead. The importance of specific phrases associated with particular topics is studied in Section 5.4, where we discuss the performance of the different communities aggregated by the relevance of their common topics.

5.3. Complexity. The training time complexity of the studied machine learning algorithms in terms of the number of operations and big O notation is in Table 3. We tried to maintain simple implementations of these algorithms to facilitate the calculation of the expert weighting algorithms. We used 206,478 features generated by the NLP methods introduced in Section 4.3.1: BoW, PDAL, LSA, LDA, POS, and SF. Only BoW generated 196,636 features, and the other methods generated the rest.

Logistic regression, linear SVM, and naïve Bayes' complexity is $O(np)$ as these methods must take $O(p)$ steps to evaluate p features and iterate over n data points. CART must follow the same procedure for every node of the tree. Hence, its complexity is $O(np\ d)$ as the previous values are multiplied by the depth d of the tree. In this paper, we worked with $d = 3$, so CART's complexity is three times the complexity of the previous algorithms (see Table 3). In the worst case, the depth of a tree could be $O(n)$ while the depth of a binary balanced tree is $O(\log_2 n)$. We used the latter to calculate the random forest's complexity as this algorithm is based on an ensemble of a forest of trees, and we did not limit the depth of the trees. Hence, its complexity $O(np\ \log_2 n\, k)$ is the same as the complexity of a decision tree where $d = \log_2 n$ multiplied by k trees or about 300 times CART's complexity.

As AdaBoost uses decision stumps (decision trees with a single split or $d = 1$) as its weak learners, then AdaBoost's complexity $O(npe)$ is the complexity of a decision tree with $d = 1$ multiplied by the number of estimators or weak experts (e). We used not more than 50 estimators; then its complexity is about 16 times the complexity of CART with $d = 3$. The complexity of these algorithms may change according to the models' definition. For instance, the complexity of a nonlinear or kernel SVM could go from $O(n^2 p)$ to $O(n^3 p)$.

The complexity of the expert weighting algorithms would be equivalent to the sum of the complexity of every individual algorithm plus $O(n)$. However, once we obtain the forecast of every method, the marginal complexity of the expert weighting algorithm is $O(n)$ as for every observation, it only has to combine the different methods according to their performance. The same logic applies to the expert weighting algorithms that include humans. The overhead or time complexity of the human component of these algorithms can be approximated by the average response time of humans as presented in Table 1.

5.4. Community Detection. We can regard our groups of machine and human experts as constituting a set of nodes connected through the stories they are good at rating as a social network. We can then use community detection algorithms to cluster the nodes in this network to understand whether and how these groups differ. The algorithm separates the community and characterizes each with a mixture of topics based on the most frequent topics used as edges in the construction of the network. We use the most frequent topics for each community to approximate the most relevant aspects used by each community to classify the news and then evaluate our second question, if humans and machines concentrate on different topics.

The most common topics' categories that we found are:

(1) "Geopolitical Units": countries or states such as Germany and France,

(2) "Business Sector": broad corporate areas such as banking services or insurance, and

(3) "Events": relevant corporate actions such as "Results Forecasts/Warnings" of future corporate results.

We use the detection method proposed by Clauset et al. [86]. This method identifies three communities that forecast return and are consistent with our main groups of analysis (see Figure 6(a) and Table 4):

(1) Algorithms: nonlinear classifiers (CART, random forests and AdaBoost) and naïve Bayes,

(2) Humans (expert and trained evaluators) and linear classifiers (SVM and logistic regression), and

(3) Crowds: human crowd (best crowd) and artificial intelligent crowds (EW, EWa, and EWh).

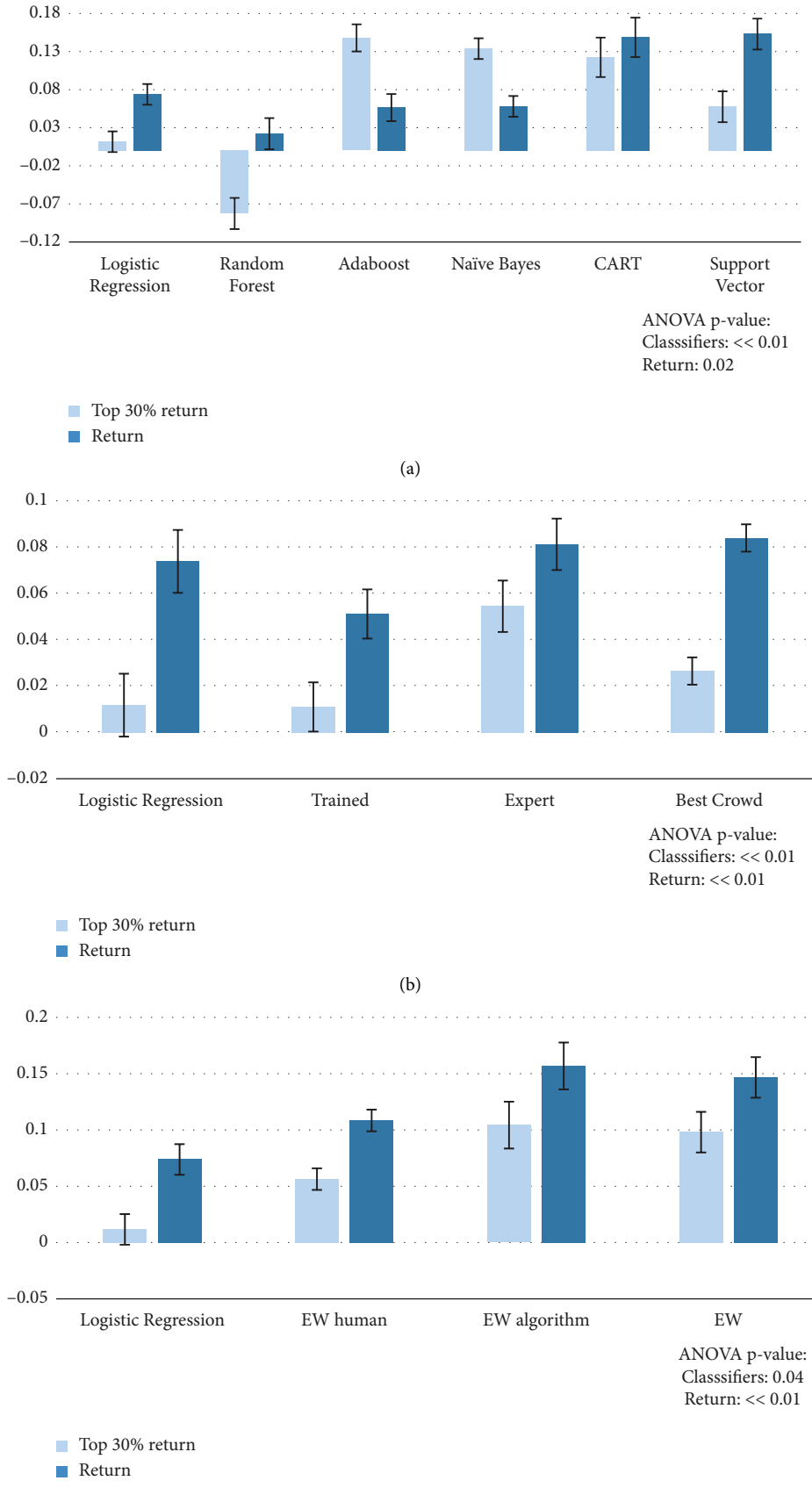

Figure 5: Average values of MCC for asset return prediction. ANOVA p-values refer to the classifiers and to the different levels of return direction: return, and top 10%, 20%, 30%, 40%, and 50% return. The graph only includes return and top 30% return. The error bars represent standard error. (a) By algorithms. (b) By humans. (c) By expert weighting algorithms.

Table 3: Training time complexity of machine learning algorithms in terms of the number of operations and big O notation. n, p, k, d, and e stand for the number of observations (479), number of features (206,478), number of trees of random forests (100), CART's depth of trees (3), and AdaBoost's number of estimators (50), respectively.

Algorithm	Formula	Time complexity
EW algorithm (total)	$O(np\log_2 nk) + O(npe) + O(npd) + 3O(np) + O(n)$	93, 600, 598, 488
Random forest	$O(np\log_2 nk)$	88, 062, 028, 784
AdaBoost	$O(npe)$	4, 945, 148, 100
CART	$O(npd)$	296, 708, 886
Support vector	$O(np)$	98, 902, 962
Logistic regression	$O(np)$	98, 902, 962
Naïve Bayes	$O(np)$	98, 902, 962
EW (marginal)	$O(n)$	479

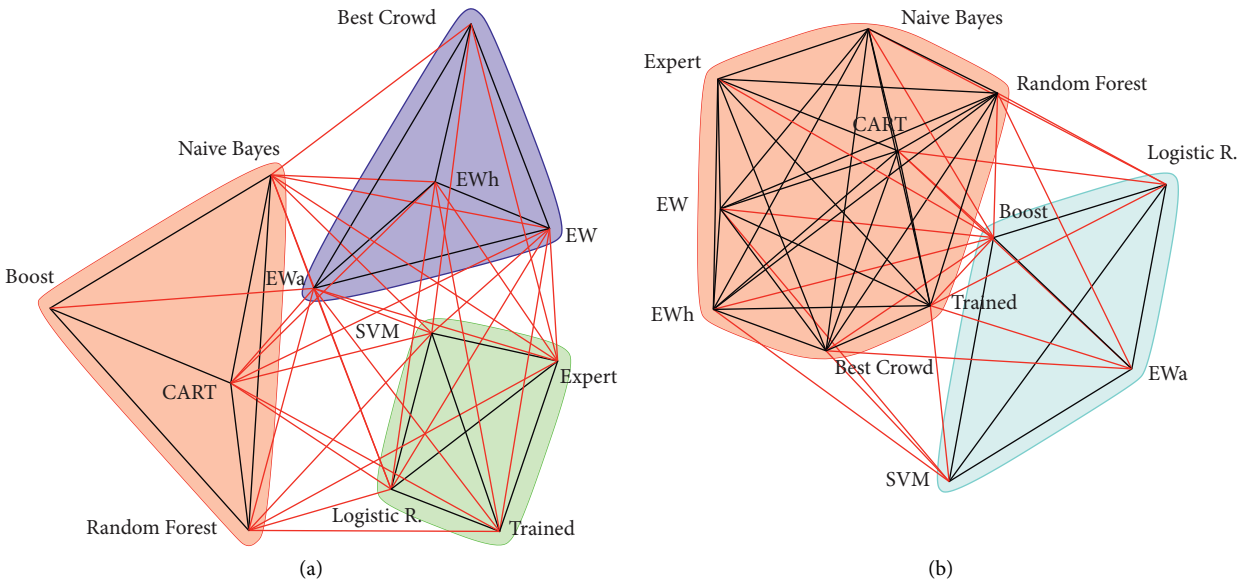

Figure 6: Network of classifiers based on the top 5% of common news topics and community structure according to Clauset et al. [86]. Boost, logistic R., SVM, and EW are AdaBoost, logistic regression, support vector machine, and expert weighting algorithm (humans h and a algorithms). (a) 3 hours return. (b) Top 30% for 3 hours return.

Table 4: Community membership of the network of classifiers based on the top 5% of common news topics for 3 hours return according to Clauset et al. [86]. NB, RF, LR, SVM, and EW are naïve Bayes, random forests, logistic regression, support vector machine, and expert weighting algorithm, respectively. MCC is the community average of the Matthews correlation coefficient. Modularity = 0.08.

I.D.	Components	MCC	Ranking of main topics by categories
1	NB, CART, RF, AdaBoost	0.07	(1) Geopolitical unit, (2) Events (results forecasts)
2	Expert, trained, LR, SVM	0.09	(1) Events (results forecasts, mergers, and acquisitions), (2) Geopolitical unit
3	EW, EWh, EWa, best crowd	0.12	(1) Geopolitical unit, (2) Business sectors

Only the second community that includes humans has "Events" as the main topics' category. This community comprises internal forecasts about a company's future results ("Results Forecasts/Warnings") and a less critical topic ("Mergers, Acquisitions & Takeovers"). For the other two communities, the main category is "Geopolitical Unit," which includes mostly news related to Germany (see Table 4). Hence, the human classification is driven by the recognition of events that may impact a future price and that are more difficult for a machine learning algorithm to recognize that they involve anticipating human reactions to human expectations. By contrast, the learning algorithms detect that most news articles that lead to a price change include a particular keyword, such as Germany. The best performer is the third community that contains all the versions of the EW algorithm and the best crowd. EW has some similarities with random forests, considering that its results are based on the combination of the output of several individual experts. However, as their past results in most cases weigh the experts, the EW algorithms show better results than random forests, so they are assigned to a different community.

Table 5: Community membership of the network of classifiers based on the top 30% 3 hours return forecast according to Clauset et al. [86]. NB, RF, LR, SVM, and EW are naïve Bayes, random forests, logistic regression, support vector machine, and expert weighting algorithm, respectively. MCC is the community average of the Matthews correlation coefficient. Modularity = 0.11.

I.D.	Components	MCC	Ranking of main topics by categories
1	NB, CART, RF, EW, EWh Best crowd, expert, trained	0.05	(1) Geopolitical unit, (2) Events (results forecasts) (2) Banking services
2	LR, SVM, AdaBoost, EWa	0.08	(1) Business sector, (2) Geopolitical unit

The classification of the top 30% return leads to the following communities of classifiers (see Figure 6(b) and Table 5):

(1) Humans, EW, EWh, and several algorithms.

(2) Best algorithms: AdaBoosting, and EWa. It also includes the best algorithm of the previous network (SVM) and logistic regression.

While the first community still selects "Geopolitical Units" (Germany and France) as the main category, the second community recognizes that "Business Sector" is the most critical category. The topics in this group are all associated with financial companies: banking services, finance, and insurance. Even though the expert has the best performance among humans, AdaBoost and EWa can detect the major return changes and converge in the second community. As common topics among its members define every community, the predictive capacity of the selected topics partially explains the performance of every community and method used. In this case, the dominance of the different versions of the EW algorithm can be explained by its capacity to select the algorithm with the best performance, and indirectly with the best combination of topics. The EW algorithm can be described as an evolving collective intelligence because of its capacity to evaluate the inputs of several experts and to take its decisions according to the experts with the lagged best performances.

6. Discussion

The importance of "Events" topics category to classify returns by humans might be due to their capacity to recognize the price impact of events using history and the information provided in the news story. Events are unique, complex, and their outcome may change according to the circumstances. They require sophisticated interpretation. For instance, an event of the type "mergers, acquisitions, and takeovers" may imply that company X takes over company Y or vice versa. However, the news story may be very ambiguous or may not offer any details: "Global steel giant Mittal Steel C. . .. confirmed on Wednesday that it was in talks on a possible deal with Hunan Valin Steel Tube & Wire Co., Ltd. . .. but would not detail the nature of the discussions" [87].

An expert with the correct information might anticipate the price impact of this deal, and even the crowd can also guess that any corporate deal may have a price impact even though it does not know the correct direction. However, an algorithm may have difficulty interpreting the stories related to a merger as being different views of the same unfolding event. Moreover, semantic frames, an advanced interpretative NLP technique used by the algorithms, can be applied only to phrases or relatively short paragraphs. But the description of an event might be extended, spread out across different news, or may not offer enough information, such as the following news alert: "TCI believes that a large majority of BOERSE shareholders would oppose BOERSE'S LSE takeover offer" [88]. This news includes the word "oppose" which the algorithm might interpret negatively. However, a human being may recognize that the show of opposition to a takeover offer might be positive for the company in the long run; therefore, it may have a positive price impact.

For the top 30% return forecasts, the first community includes humans, EWh, EW, and several algorithms. As a result of this mixture of different groups, "Results Forecasts/Warnings" (part of "Events" category) becomes the second most important topic. An additional explanation for reducing the importance of events in an extreme situation is that they may lead to either overconfidence or emotional reaction by humans [89]. In these situations, the algorithms have the advantage that they are dispassionate. The importance of each topic is based on the random forests' importance score which indicates the contribution of each topic to the out-of-sample accuracy prediction.

For learning and expert weighting algorithms, the most critical topic used to forecast return is the category "Geopolitical Units" which includes topics such as "Germany" or "France."

In most of cases, this topic is associated with the residence of a corporation instead of referring to a particular country: it is easy for an algorithm to recognize the name of a company and associate it with its past performance. In this case, if a company's performance in the test dataset is consistent with its performance in the training dataset, then the return prediction is adequate. The problem happens when the company's performance changes or the news may refer to a different company or event. For instance, the algorithm may remember that the performance of company X is positive. However, the news story may mention company X as the company affiliation of a financial analyst that follows company Y. Therefore, the news target is company Y. In these situations, semantic frames might be beneficial to disambiguate several news stories with common words but with very different targets. A similar analysis applies to "Business Sector," the most critical topics category to forecast the top 30% return by the community of algorithms.

The expert and the best crowd belong to different communities, even though they show similar performance in forecasting returns considering that crowds with limited knowledge can collectively predict like an expert with greater

$$
\begin{aligned}
&F_0(x) \equiv 0 \\
&\text{for } t = 1 \ldots T \\
&\qquad w_i^t = e^{-y_i F_{t-1}(x_i)} \\
&\qquad \text{Get } h_t \text{ from } \textit{weak learner} \\
&\qquad \alpha_t = \frac{1}{2} \ln \left(\frac{\sum_{i:h_t(x_i)=1, y_i=1} w_i^t}{\sum_{i:h_t(x_i)=1, y_i=0} w_i^t} \right) \\
&\qquad F_{t+1} = F_t + \alpha_t h_t
\end{aligned}
$$

FIGURE 7: The AdaBoost algorithm [14]. y_i is the binary label to be predicted, x_i corresponds to the features of an instance i, w_i^t is the weight of instance i at time t, h_t and $F_t(x)$ are the prediction rule and the prediction score at time t, respectively.

knowledge as the cognitive literature [50, 52] is proposed. The fact that the best crowd and all the expert weighting algorithms are in the same community and share similar topics ("Geopolitical Units" and "Business Sectors"), and that the prediction of all of them are based on the aggregation of the best individual decisions leads us to believe that their cognitive processes are similar. For this reason, we refer to the expert weighting algorithm as an artificial collective intelligence capable of predicting at least as well as any of the individual experts. The same distinction we made above between topics selected by humans and algorithms applies to the difference in topics and probably underlying cognitive processes between the expert and the community of EW algorithms and the best crowd.

To summarize, combining individual predictions through alternative versions of the EW algorithm leads to an improved aggregated prediction. Additionally, the communities of humans, algorithms, and expert weighting algorithms are associated with different topics.

7. Conclusion

As part of an effort to understand the role that crowds and experts can play in interpreting news and predicting news impact, several different kinds of crowds and algorithms were assembled to characterize the sentiment of news stories. As evaluated by the crowds, this sentiment was used to predict returns directly. Also, an expert weighting algorithm was designed and tested; this algorithm automated and aggregated both machine predictions and crowd input to make predictions. The results, tested against securities over time, showed that the expert weighting algorithm is very similar to or better than the best algorithm or human in most cases. From a cognitive perspective, the dynamic selection of the best expert or a combination of them based on previous performance in the expert weighting algorithm that follows the spirit of evolving collective intelligence.

Introducing financial and accounting variables that combine news sentiment with information about companies' stories enriches the expert weighting algorithm. Additionally, other aspects of the content—the people, places, and concepts mentioned— can also be explored as additional features that could improve the prediction of asset returns. Even though this paper has focused on asset returns, this work might possibly be extended to predicting other social, political, or market events.

Concerning research on crowd work, our findings suggest that, while crowds in aggregate may not always be good at interpretation, some individuals in the crowd or some experts are likely to be very good at it and that finding these experts is a desirable thing to do. They can make accurate predictions and generate a strong signal that machine learning algorithms can use. Humans can perform an act of interpretation that machines are not yet capable of. However, once this act of interpretation is performed, machines may learn from it. Moreover, in many situations, human input may prove necessary. For example, when the context changes suddenly, humans are arguably better at adapting than computers. As this research shows, humans and algorithms belong to different communities characterized by different prediction skills about particular news topics. Humans and machines can be integrated into a hybrid system that combines the machine's ability to process large amounts of data with the humans' ability to interpret.

The current business environment processes a large amount of heterogeneous data that arrives very fast. Using experts or the crowd to interpret it is not always practical. This research shows that the distance between the machine and human interpretation can be reduced, transforming the original text into components that simulate how humans process the information incorporating meaningful keywords that capture human emotions combined with semantic interpretations of short news stories. When an expert weighting algorithm evaluates this preprocessed information, an artificially intelligent crowd, its predictive capacity outperforms the other classification methods. The results suggest one path toward an artificial collective intelligence [90].

Appendix

A. Automated Text Analysis Methods

A.1. Unsupervised Learning Algorithms. Unsupervised learning algorithms typically cluster categories or topics of a series of documents without previous knowledge of those categories. Among the most popular methods are latent semantic analysis and latent Dirichlet allocation:

A.1.1. Latent Semantic Analysis. Latent semantic analysis (LSA) is a method to approximate a term-document matrix C using singular value decomposition. This procedure generates a low-rank approximation C_k to C, which allows representing each document with k dimensions. These k dimensions are determined by the k principal eigenvectors that represent the largest eigenvalues of CC^T and C^TC. The k dimensions of each document can be used to compute similarities between documents, terms, a query, and a document [92].

A.1.2. Latent Dirichlet Allocation. The topic model methodology [93] discovers common topics among a series of

documents. This approach assumes that documents are a mixture of topics, where a topic is based on a probability distribution over words. Latent Dirichlet allocation (LDA) evaluates a new document by selecting topics according to their distribution and keywords. We can infer the topics used to generate the documents inverting this process. LDA is an accepted topic model methodology for capturing the latent structure of a large set of documents. LDA simply supposes that the topic distribution follows a Dirichlet prior [94]. This approach helps us to cluster topics across large data sets of news.

The application of these unsupervised methods to business problems is still very limited. Aral et al. [95] used LDA to extract common topics among 2,397 stock recommendations; Creamer et al. [91] applied LDA to extract common topics in a corporate network and used it to forecast return; Bao and Datta [96], in a Management Science article, used an extended version of LDA topic model to evaluate the effect of risk disclosures in 10-K forms on the risk perception of investors; and Xie et al. [17] tested several NLP methods such as BoW, LDA, DAL, and semantic frames for stock price prediction.

A.2. Classic Supervised Learning Algorithms

This section introduces a group of classification or supervised learning algorithms that use different text analysis features to predict the asset return direction. All of them produced similar categorical output. However, we decided to explore these diverse methods as they emphasize different aspects of the classification problem.

The following methods include a vector of inputs $X = (X_1, X_2, \ldots, X_p)$ that predict the binary output Y. The training dataset of (X, Y) includes pairs (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), where x_i corresponds to the features of an instance, and $y_i \in \{0, 1\}$ is the binary label that is to be predicted.

A.2.1. Logistic Regression. The logistic regression models [60] and the posterior probabilities $\Pr(Y = l \mid X_i)$ of L classes Y using linear regression in the vector of features X as

$$\Pr\left(Y = l|X_i\right) = \left(\frac{e^{\sum_{j=1}^{n} \beta_j X_j}}{1 + e^{\sum_{j=1}^{n} \beta_j X_j}}\right). \tag{A.1}$$

The summation of these probabilities equals one. Logistic regression results are better interpreted using the odds ratio as

$$\frac{\Pr\left(Y = l \mid X_i\right)}{\Pr\left(Y = L \mid X_i\right)}. \tag{A.2}$$

We use logistic regression as our baseline algorithm.

A.2.2. Naïve Bayes. Naïve Bayes (NB) is a simple Bayesian classifier which assumes that a cause Y has an impact on several effects $(X_1, X_2, \ldots, X_p)$ that are conditionally independent given a particular class Y_j. Their full joint distribution or a naïve Bayes model is $P(Y_j, X_1, X_2, \ldots, X_p) = P(Y_j)\pi_i P(X_i \mid Y_j)$. The conditional independence assumption significantly simplifies the computation of the joint distribution. Despite its simplicity, this model has been demonstrated to be very effective even if the conditional independence assumption is not satisfied. For this reason, NB is typically used by the BoW model. This model classifies documents using the frequency of n-grams as features to train a classifier. Additionally, NB is a reasonable model to simulate cognitive judgments considering that humans use prior distributions for specific daily prediction tasks either individually [50] or collectively [52].

A.2.3. Support Vector Machine. Vapnik [97] proposed support vector machine (SVM) as a classification method based on the use of kernels to preprocess data in a higher dimension than the original space. This transformation allows an optimal hyperplane to separate the data into two categories or values.

A hyperplane is defined as

$$\left\{X\colon F(X) \doteq X^T\beta + \beta_0 = 0\right\}, \tag{A.3}$$

where $\|\beta\| = 1$.

The strong prediction rule learned by a SVM model is $\operatorname{sign}(F(X))$ (see Hastie et al. [60]).

NB and SVM are the most popular methods used in NLP for text classification. Li [98] built a NB model to classify the forward-looking statements of 10-K and 10-Q forms. Antweiler and Frank [7] used a small number of manually labeled text messages to train NB and SVM models and then calculated the sentiment of 1.5 million text messages. Schumaker and Chen [75] built an SVM model using BoW, noun phrases, and named entities of financial news as the model's features to forecast stock price changes. Noun phrases outperform the rest of the features achieving an accuracy rate of 57.1%. Mittermayer [99] designed a similar system with SVM and K-nearest neighbors using tf-idf as features of the model. SVM with a polynomial kernel showed the best performance. Hagenau et al. [62] used word combinations as features, a χ^2 method to select the best features, and SVM as the classification algorithm. Luss and Alexandre [100] used SVM for intraday price prediction.

A.2.4. Decision Trees. Classification and regression trees (CART) is a popular decision tree algorithm proposed by Breiman et al. [101]. CART builds a binary decision tree following a top-down approach where the root node is the best feature where its branches are its values and can separate the data into two parts according to a test such as information gain. CART repeats this process successively with the descendants of each node, creating two new nodes until there is no further information gained or any other stopping rule is satisfied. An alternative implementation is to grow the tree as much as possible and then prune each node that most improves accuracy. At that point, CART includes the leaf node with the most common value of the target attribute.

The information gain $(\Delta E(S, A))$ for the introduction of the target feature A and the sample of training observation S is defined as

$$\Delta E(S, A) \doteq E(S) - \sum_{v \in \text{values}(A)} \frac{S_v}{S} E(S_v), \tag{A.4}$$

and entropy impurity is

$$E(S) \doteq -\sum_{i=1}^{c} p_i \log_2 p_i, \tag{A.5}$$

where p_i is the proportion of observations S that belongs to class i and c and represents the different values of the target feature.

Oh and Sheng [102] manually labeled 7,109 blog postings and used the decision tree C4.5, which is very similar to CART to assess the remaining blog postings. Their results show that the supervised method performs slightly better than a lexicon. Yu et al. [103] showed that CART, followed by random forests, outperforms SVM and artificial neural networks to forecast credit ratings of decarbonized firms.

A.3. Supervised Learning: Ensemble Algorithms

This section introduces boosting and random forests, two well-known ensemble methods that automate the task of forecasting asset returns, aggregating the output of many individual experts using the same algorithm. In the case of boosting, every iteration increases the weight of the misclassified observations while random forests randomly select different samples and features to build several decision trees.

A.3.1. Boosting. AdaBoost is a classification learning algorithm proposed by Freund and Schapire [14] and introduced in Figure 7. AdaBoost repeatedly applies a weak or base learner (an algorithm with a performance at least slightly better than random guessing) to each instance i of a training set and produces a prediction rule h_t that maps x to $\{0, 1\}$. Mapping x to $\{0, 1\}$ instead of $\{-1, +1\}$ increases the flexibility of the weak learner. Zero can be interpreted as "no prediction." After every iteration t, a performance prediction score $F_t(x)$ is updated by adding to its small corrections given by the weak prediction functions. The weight w_i^t is assigned to each instance i using an exponential function where the misclassified instances receive a larger weight than the rest. The hypothesis that is generated by the weak learner in each iteration is combined into a single strong rule using a weighted majority vote as $\text{sign}(F(x))$.

We decided to use boosting as one of our learning algorithms because of its feature selection capability, its error bound proofs [14], its interpretability, its capacity to combine both quantitative and qualitative variables, and its previous application to financial forecasting by Creamer and Freund [104].

A.3.2. Random Forests. Random forests is a variant of bagging decision trees proposed by Breiman [105]. We chose this algorithm because it presents the best publicly available combination of decision trees and bagging.

If the training set Y consists of pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, x_i corresponds to the features of an example, and y_i is either a class label or a numerical response to be predicted. Random forests generate multiple trees (θ_i) from uniform bootstrap samples with replacement of Y and from several variables X randomly selected using any decision tree. As a result each tree generates a predictor of Y ψ (X, θ_i). When Y is a numerical response, the average of the predictors is the final prediction:

$$\psi_{RF}(X) = \frac{\sum_{i=1}^{n} \psi(X, \theta_i)}{n}. \tag{A.6}$$

If Y is a class label, $\psi_R F(X)$ is obtained by the majority vote of the individual predictors $\psi(X, \theta_i)$.

When the number of trees is vast, the generalization error for forests converges. Breiman [105] indicated that the accuracy of random forests is as good as AdaBoost or better. Random forests generates a standardized score, a z-score, that indicates the importance of each variable in the final classification.

Onan et al. [106] showed that bagging and random forests, with the most frequent measure of keyword extraction, outperform other classic machine learning algorithms such as NB, logistic regression, and SVM for text classification.

Data Availability

The set of news with the stock prices used to support the findings of this study were supplied by Thomson Reuters under license and so cannot be made freely available. Requests for access to these data should be made to https://www.reutersagency.com/en/contact-us.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Grant from the Howe School Alliance for Technology Management of the Stevens Institute of Technology.

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[2] G. G. Creamer, Y. Ren, Y. Sacamoto, and J. V. Nickerson, "News and sentiment analysis of the European market with a hybrid expert weighting algorithm," in *Proceedings of the IEEE International Conference on Social Computing*, pp. 391–396, Alexandria, VA, USA, September 2013.

[3] E. Fama, "The behavior of stock-market prices," *Journal of Business*, vol. 38, no. 1, pp. 34–105, 1965.

[4] R. J. Shiller, "Do stock prices move too much to be justified by subsequent changes in dividends?" *The American Economic Review*, vol. 71, pp. 421–436, 1981.

[5] R. W. Roll, "R-squared," *The Journal of Finance*, vol. 43, pp. 541–566, 1988.

[6] D. Cutler, J. M. Poterba, and L. H. Summers, "What moves stock prices?" *Journal of Portfolio Management*, vol. 15, pp. 4–12, 1989.

[7] W. Antweiler and M. Z. Frank, "Is all that talk just noise? the information content of internet stock message boards," *The Journal of Finance*, vol. 59, pp. 1259–1293, 2004.

[8] J. D. Coval and S. Tyler, "Is sound just noise?" *The Journal of Finance*, vol. 56, no. 5, pp. 1887–1910, 2001.

[9] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: quantifying language to measure firms' fundamentals," *The Journal of Finance*, vol. 63, no. 3, pp. 1437–1467, 2008.

[10] P. C. Tetlock, "All the news that's fit to reprint: do investors react to stale information?" *Review of Financial Studies*, pp. 1481–1512, 2011.

[11] V. Choudhary, A. Marchetti, Y. R. Shrestha, and P. Puranam, "Human-algorithm ensembles," *Tech. rep., INSEAD Working*, 2021.

[12] F. A. Csaszar and O. James, "A contingency theory of representational complexity in organizations," *Organization Science*, vol. 31, pp. 1198–1219, 2020.

[13] T. Simons, L. Hope Pelled, and K. A. Smith, "Making use of difference: diversity, debate, and decision comprehensiveness in top management teams," *Academy of Management Journal*, vol. 42, pp. 662–673, 1999.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[15] A. Devitt and K. Ahmad, "Sentiment polarity identification in financial news: a cohesion based approach," in *45th Annual Meeting of the Association of Computational Linguistics*, pp. 984–991, Association for Computational Linguistics, Prague, Czech Republic, 2007.

[16] S. A. Haider and R. Mehrotra, "Corporate news classification and valence prediction: a supervised approach," in *2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, pp. 175–181, Association for Computational Linguistics, Portland, Oregon, 2011.

[17] B. Xie, R. J. Passonneau, L. Wu, and G. Creamer, "Semantic frames to predict stock price movement," *51st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sofia, Bulgaria, 2013.

[18] N. Archak, A. Ghose, and P. G. Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews," *Management Science*, vol. 57, no. 8, pp. 1485–1509, 2011.

[19] P. G. Ipeirotis and E. Gabrilovich, "Quiz: targeted crowdsourcing with a billion (potential) users," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 143–154, WWW '14, ACM, New York, NY, USA, April 2014.

[20] H. Li, B. Zhao, and A. Fuxman, "The Wisdom of Minority: Discovering and Targeting the Right Group of Workers for Crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 165–176, WWW '14, ACM, New York, NY, USA, April 2014.

[21] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang, "Mining collective intelligence in diverse groups," in *Proceedings of the 22Nd International Conference on World Wide Web*, pp. 1041–1052, Geneva, Switzerland, May 2013.

[22] A. Kittur, J. V. Nickerson, M. Bernstein et al., "The Future of Crowd Work," in *Proceedings of the 2013 Conference on Computer supported Cooperative Work*, pp. 1301–1318, New York, NY, USA, February 2013.

[23] P. Bajari, D. Nekipelov, S. P. Ryan, and M. Yang, *Machine Learning Methods for Demand Estimation*, Working Paper, 2014.

[24] J. M. Bates and W. J. G. Clive, "The combination of forecasts," *Operational Research Quarterly*, vol. 20, pp. 451–468, 1969.

[25] R. T. Clemen, "Combining forecasts: a review and annotated bibliography," *International Journal of Forecasting*, vol. 5, pp. 559–583, 1989.

[26] K. F. Wallis, "Combining forecasts: forty years later," *Applied Financial Economics*, vol. 21, no. 1-2, pp. 33–41, 2011.

[27] F. Li, P. Dechow, I. Dichev et al., "Do stock market investors understand the risk sentiment of corporate annual reports?" *SSRN Working Paper Series*, 2006.

[28] A. K. Davis, J. M. Piger, and M. S. Lisa, "Beyond the numbers: measuring the information content of earnings press release language," *Contemporary Accounting Research*, vol. 29, no. 3, pp. 845–868, 2012.

[29] S. R. Das and M. Y. Chen, "Yahoo! for amazon: sentiment extraction from small talk on the web," *Management Science*, vol. 53, no. 9, pp. 1375–1388, 2007.

[30] D. Stromberg, "Radio's impact on public spending," *Quarterly Journal of Economics*, vol. 119, no. 1, pp. 189–221, 2004.

[31] M. Gentzkow and J. M. Shapiro, "What drives media slant? evidence from U.S. daily newspapers," *Econometrica*, vol. 78, no. 1, pp. 35–71, 2010.

[32] M. Gentzkow, "Television and voter turnout," *Quarterly Journal of Economics*, vol. 121, no. 3, pp. 931–972, 2006.

[33] A. Dyck, N. Volchkova, and L. Zingales, "The corporate governance role of the media: evidence from Russia," *The Journal of Finance*, vol. 63, no. 3, pp. 1093–1135, 2008.

[34] A. S. Gerber, K. Dean, and D. Bergan, "Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions," *American Economic Journal: Applied Economics*, vol. 1, no. 2, pp. 35–52, 2009.

[35] L. Fang and J. Peress, "Media coverage and the cross-section of stock returns," *The Journal of Finance*, vol. 64, no. 5, pp. 2023–2052, 2009.

[36] C. A. Parsons, "The causal impact of media in financial markets," *The Journal of Finance*, vol. 66, no. 1, pp. 67–97, 2011.

[37] J. Bollen and H. Mao, "Twitter mood as a stock market predictor," *Computer*, vol. 44, no. 10, pp. 91–94, 2011.

[38] X. Luo, J. Zhang, and W. Duan, "Social media and firm equity value," *Information Systems Research*, vol. 24, no. 1, pp. 146–163, 2013.

[39] X. Luo and J. Zhang, "How do consumer buzz and traffic in social media marketing predict the value of the firm?" *Journal of Management Information Systems*, vol. 30, no. 2, pp. 213–238, 2013.

[40] T. Preis and H. S. Moat, H. E. Stanley, Quantifying trading behavior in financial markets using Google trends," *Scientific Reports*, vol. 3, pp. 1–6, 2013.

[41] H. Sul, A. R. Dennis, and L. I. Yuan, "Trading on twitter: the financial information content of emotion in social media," in *Proceedings of the 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, January 2014.

[42] D. K. Pearce and V. Vance Roley, "Stock prices and economic news," *Journal of Business*, vol. 58, no. 1, pp. 49–67, 1985.

[43] M. J. Fleming and E. M. Remolona, "What moves the bond market?" *Economic Policy Review*, vol. 3, no. 4, pp. 31–50, 1997.

[44] R. F. Engle and V. K. Ng, "Measuring and testing the impact of news on volatility," *The Journal of Finance*, vol. 48, no. 5, pp. 1749–1778, 1993.

[45] W. S. Chan, "Stock price reaction to news and no-news: drift and reversal after headlines," *Journal of Financial Economics*, vol. 70, pp. 223–260, 2003.

[46] D. L. Medin and M. M. Schaffer, "Context theory of classification learning," *Psychological Review*, vol. 85, pp. 207–238, 1978.

[47] J. K. Kruschke, "ALCOVE: an exemplar-based connectionist model of category learning," *Psychological Review*, vol. 99, pp. 22–44, 1992.

[48] B. C. Love, D. L. Medin, and T. Gureckis, "SUSTAIN: a network model of human category learning," *Psychological Review*, vol. 111, pp. 309–332, 2004.

[49] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, "Topics in semantic representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[50] T. L. Griffiths and J. B. Tenenbaum, "Optimal predictions in everyday cognition," *Psychological Science*, vol. 17, no. 9, pp. 767–773, 2006.

[51] M. Steyvers and J. B. Tenenbaum, "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth," *Cognitive Science*, vol. 29, no. 1, pp. 41–78, 2005.

[52] M. C. Mozer, H. Pashler, and H. Homaei, "Optimal predictions in everyday cognition: the wisdom of individuals or crowds?" *Cognitive Science*, vol. 32, pp. 1133–1147, 2008.

[53] F. Galton, "Vox populi," *Nature*, vol. 75, pp. 450-451, 1907.

[54] H. Gurnee, "Maze learning in the collective situation," *Journal of Psychology*, vol. 3, no. 2, pp. 437–443, 1937.

[55] Y. Sakamoto and B. Jin, "Testing tournament selection in creative problem solving using crowds," *International Conference on Information Systems (ICIS) Proceedings*, Association for Information Systems, 2011.

[56] L. Yu and J. V. Nickerson, "An internet-scale idea generation system," *ACM Transactions on Interactive Intelligent Systems*, vol. 3, no. 1, pp. 1–24, 2013.

[57] Y. Sakamoto, Y. Tanaka, L. Yu, and J. V. Nickerson, "The crowdsourcing design space," in *Foundations of Augmented Cognition Directing the Future of Adaptive Systems, Lecture Notes in Computer Science*, D. D. Schmorrow and C. M. Fidopiastis, Eds., vol. 6780, pp. 346–355, Springer Berlin Heidelberg, 2011.

[58] L. Yu, J. V. Nickerson, and Y. Sakamoto, *Collective Creativity: Where We Are and where We Might Go*, Collective Intelligence 2012, Cambridge, MA USA, 2012.

[59] Y. Nagar and T. W. Malone, *Making Business Predictions by Combining Human and Machine Intelligence in Prediction Markets. 32nd International Conference on Information Systems*, Association for Information Systems, Shanghai, 2011.

[60] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York, 2nd ed edition, 2008.

[61] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[62] M. Hagenau, M. Liebmann, M. Hedwig, and D. Neumann, "Automated news reading: stock price prediction based on financial news using context-specific features," *Hawaii International Conference on System Sciences (HICSS, Big Island)*, 2012.

[63] N. Godbole, M. Srinivasaiah, and S. Skiena, *Large-scale sentiment analysis for news and blogs*, Vol. 2, 1st International Conference on Weblogs and Social Media, Boulder, CO, 2007.

[64] W. Zhang and S. Skiena, *Trading Strategies to Exploit Blog and News Sentiment*, 4th International AAAI Conference on Weblogs and Social Media, Washington, D.C., 2010.

[65] D. Bollegala, D. Weir, and J. Carroll, "Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 132–141, ACL, Oregon, January 2011.

[66] S. P. Kothari and X. Li, J. E. Short, The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis," *The Accounting Review*, vol. 84, no. 5, pp. 1639–1670, 2009.

[67] T. D. Kravet and V. Muslu, "Textual risk disclosures and investors' risk perceptions," *Review of Accounting Studies*, vol. 18, no. 4, pp. 1088–1122, 2011.

[68] J. L. Campbell and H. Chen, D. S. Dhaliwal and H. M. Lu, L. B. Steele, The information content of mandatory risk factor disclosures in corporate filings," *Review of Accounting Studies*, vol. 19, no. 1, pp. 396–455, 2014.

[69] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal, "Management's tone change, post earnings announcement drift and accruals," *Review of Accounting Studies*, vol. 15, no. 4, pp. 915–953, 2010.

[70] J. L. Rogers, A. V. Buskirk, and S. L. C. Zechman, "Disclosure tone and shareholder litigation," *The Accounting Review*, vol. 86, no. 6, pp. 2155–2183, 2011.

[71] B. Wuthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam, "Daily prediction of major stock indices from textual www data," in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp. 364–368, AAAI Press, New York, August 1998.

[72] C. M. Whissell, "The dictionary of affect in language," in *Measurement of Emotion*pp. 113–131, Academic Press, 1989.

[73] A. Agarwal, B. Xie, I. Vovsha, R. Owen, and R. Passonneau, *Sentiment analysis of twitter data. Workshop on Languages in Social Media*, Vol. 30–38, LSM '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.

[74] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[75] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: the AZFin text system," *ACM Transactions on Information Systems*, vol. 27, no. 2, pp. 1–19, 2009.

[76] C. J. Fillmore, "Frame semantics," *Cognitive linguistics: Basic readings*, vol. 34, pp. 373–400, 2006.

[77] C. F. Baker and C. J. Fillmore, J. B. Lowe, The Berkeley Framenet project," in *Proceedings of the 17th international conference on Computational linguistics*, pp. 86–90, August 1998.

[78] D. Das, N. Schneider, D. Chen, and N. A. Smith, *Probabilistic Frame-Semantic Parsing*, North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Los Angeles, CA, 2010.

[79] N. Dalkey and O. Helmer, "An experimental application of the delphi method to the use of experts," *Management Science*, vol. 9, no. 3, pp. 458–467, 1963.

[80] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, pp. 212–261, 1994.

[81] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *Journal of the ACM*, vol. 44, no. 3, pp. 427–485, 1997.

[82] G. Creamer and Y. Freund, "Automated trading with boosting and expert weighting," *Quantitative Finance*, vol. 10, no. 4, pp. 401–420, 2010.

[83] G. Creamer, "Model calibration and automated trading agent for Euro futures," *Quantitative Finance*, vol. 12, no. 4, pp. 531–545, 2012.

[84] R. Gao, X. Zhang, H. Zhang, Q. Zhao, and Y. Wang, "Forecasting the overnight return direction of stock market index combining global market indices: a multiple-branch deep learning approach," *Expert Systems with Applications*, vol. 194, Article ID 116506, 2022.

[85] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[86] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 70, Article ID 066111, 2004.

[87] M. Cheary, *Mittal Steel Confirms in Talks with Chinese Firm*, Reuters, 2005.

[88] R. Reuters, *Tci Believes Large Majority of Boerse Shareholders Would Oppose Boerse's Lse Takeover Offer*, Reuters, 2005.

[89] R. J. Shiller, *Irrational Exuberance*, Princeton University Press, Princeton, NJ, 3rd ed edition, 2015.

[90] G. G. Creamer, Y. Ren, Y. Sakamoto, and J. V. Nickerson, "Ensembles of Crowds and Computers: Experiments in Forecasting," *Stevens Institute of Technology School of Business Research*, vol. 54, 2015.

[91] G. G. Creamer, Y. Ren, and J. V. Nickerson, "Impact of dynamic corporate news networks on asset return and volatility," in *Proceedings of the IEEE International Conference on Social Computing*, pp. 809–814, Alexandria, VA, USA, September 2013.

[92] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

[93] M. Steyvers and T. Griffiths, *Probabilistic topic models, Handbook of latent semantic analysis*pp. 439–460, Psychology press, 2007.

[94] D. M. Blei, Y. N. Andrew, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[95] S. Aral, P. G. Ipeirotis, and S. Taylor, "Content and Context: Identifying the Impact of Qualitative Information on Consumer Choice," in *Proceedings of the 32nd International Conference on Information Systems*, pp. 511–525, AIS, Shanghai, March 2011.

[96] Y. Bao and A. Datta, "Simultaneously discovering and quantifying risk types from textual risk disclosures," *Management Science*, vol. 60, no. 6, pp. 1371–1391, 2014.

[97] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[98] F. Li, "The information content of forward-looking statements in corporate filings: a naive Bayesian machine learning approach," *Journal of Accounting Research*, vol. 48, pp. 1049–1102, 2010.

[99] M. A. Mittermayer, "Forecasting Intraday Stock price Trends with Text Mining Techniques," in *Proceedings of the 37th Hawaii International Conference on System Sciences*, Big Island, HI, USA, January 2004.

[100] R. Luss and d'A. Alexandre, "Predicting abnormal returns from news using text classification," *Quantitative Finance*, vol. 15, no. 6, 2011.

[101] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Wadsworth and Brooks, Belmont, 1984.

[102] C. Oh and O. Sheng, "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement," in *Proceedings of the 32nd International Conference on Information Systems*, AIS, Shanghai, December 2011.

[103] B. Yu, C. Li, N. Mirza, and M. Umar, "Forecasting credit ratings of decarbonized firms: Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models comparative assessment of machine learning models," *Technological Forecasting and Social Change*, vol. 174, Article ID 121255, 2022.

[104] G. Creamer and Y. Freund, "Using boosting for financial analysis and performance prediction: application to S&P 500 companies, Latin American ADRs and banks," *Computational Economics*, vol. 36, no. 2, pp. 133–151, 2010.

[105] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[106] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.