AIMS Mathematics, 9(1): 1683–1717.
DOI: 10.3934/math.2024083
Received: 12 September 2023
Revised: 21 November 2023
Accepted: 23 November 2023
Published: 13 December 2023

http://www.aimspress.com/journal/Math

Research article

Stock market uncertainty determination with news headlines: A digital twin approach

Pedro J. Gutiérrez-Diez[1,*] **and Jorge Alves-Antunes**[2]

[1] Department of Economic Theory, University of Valladolid, Valladolid, Spain
[2] Doctorate School, University of Valladolid, Valladolid, Spain

* **Correspondence:** Email: pedro.gutierrezdiez@uva.es; Tel: +34983423282; Fax: +34983423292.

Abstract: We present a novel digital twin model that implements advanced artificial intelligence techniques to robustly link news and stock market uncertainty. On the basis of central results in financial economics, our model efficiently identifies, quantifies, and forecasts the uncertainty encapsulated in the news by mirroring the human mind's information processing mechanisms. After obtaining full statistical descriptions of the timeline and contextual patterns of the appearances of specific words, the applied data mining techniques lead to the definition of regions of homogeneous knowledge. The absence of a clear assignment of informative elements to specific knowledge regions is regarded as uncertainty, which is then measured and quantified using Shannon Entropy. As compared with standard models, the empirical analyses demonstrate the effectiveness of this approach in anticipating stock market uncertainty, thus showcasing a meaningful integration of natural language processing, artificial intelligence, and information theory to comprehend the perception of uncertainty encapsulated in the news by market agents and its subsequent impact on stock markets.

Keywords: stock market; efficient market theory; market selection theory; uncertainty; digital twin; natural language processing; artificial intelligence; Shannon entropy
Mathematics Subject Classification: 91-08, 91-10

1. Introduction

Generally speaking, uncertainty describes a situation that cannot be accurately predicted, and is, by its nature, a subject difficult to formalize and study. Indeed, the formal analysis of uncertainty and its consequences did not appear until the 16th-17th centuries, linked to the study of games of chance [1–3], and progressively moved to other sciences, specially economics and finance. This is the logical consequence of the pervasive role that uncertainty plays in economic life, which has compelled economists to develop formal and mathematical tools able to tackle this question and its consequences.

In fact, strictly speaking, even the interest behind games of chance was an economic interest.

As a result, today, a large part of the central concepts and techniques analyzing uncertainty are contributions derived from Economics. This is the case of the basic formal construction used to describe how uncertainty is structured and perceived, the so-called state of nature [4, 5]. Roughly defined, this state of nature is any of the possible uncertain outcomes, which is a simple concept that opened the analysis of more complicated issues. For instance, depending on whether or not the actually existing states of nature can all be identified and their probabilities can be assigned, economics distinguishes between risk and uncertainty itself [6, 7]. Risk defines a situation in which the states are all identified and their randomness is expressed in terms of numerical probabilities. On the other hand, situations where not all the possible states of nature can be identified, as well as those where probabilities are not or cannot be assigned, are said to involve uncertainty in a strict sense. In addition, when probabilities are assigned, they can be either subjectively or objectively assessed, which is a question with far-reaching implications also closely related to the true source of uncertainty. When the uncertainty/risk framework is the result of non-manageable factors, it is said to be exogenous. On the contrary, when the uncertain scheme depends on the action itself of inspecting or dealing with the uncertain situation, the source of uncertainty is endogenous. Both sources can coexist and interact, as usually happens in a wide variety of natural, social or economic phenomena, especially when the involved agents present heterogeneous beliefs on the occurrence of the different states of nature and/or when formal models are used to forecast behaviors [8–11]. Obviously, the presence of uncertainty is an important negative factor in any task or action, and its prediction and quantification emerge as useful and crucial when designing decision processes, particularly in economics and finance. However, as a result of all these dimensions inherent to the nature of uncertainty in asset markets, its identification and measurement becomes a hard problem, both conceptually and practically. Indeed, to the present day, there is no completely satisfactory way to identify and quantify stock market uncertainty, and research in this area continues.

In this respect, our contribution comes from interpreting two basic results of financial economics through the lens of Ockham's Razor, thus opening up a simple and operative definition of stock market uncertainty. The two results provided by economic theory are the efficient market theory (EMT) and the market selection theory (MST) [12–17], which assert that stock prices are the result of the decisions of rational agents that correctly select and consider all the relevant information. According to these theories, the evolution and dynamics of stock market prices are simply the result of the interpretation of the available information by the human mind; consequently, times of turmoil and uncertainty in the stock market must correspond to the lack of a certain and sure interpretation of the available information by the market agents. This opens up the possibility of building a digital twin [18, 19] of the information processing carried out by the human mind, and of using this digital twin to identify situations of vague and unclear interpretation of the available information. According to the EMT and MST, this uncertain reading by the human mind of the available information must correspond to present or future real episodes of market uncertainty, and therefore, the proposed digital twin model can be used to forecast and quantify stock market uncertainty.

To illustrate the usefulness of our approach, it suffices to describe the complexities of identifying and measuring uncertainty in the simplest case, which is that associated with the concept of risk. As we have mentioned, in financial economics, risk is defined as a situation in which all the possible states are identified and their numerical probabilities are assigned. As an example, let us consider

the measurement of the risk associated to the returns of an equity. First, it is necessary to identify all the possible states of nature, that is, all the possible returns for the considered equity. Then, for each value of the return (each state), some probability must be assigned. The result is a probability distribution for the returns, from which we should quantify the associated risk. Under normality of the distribution for the returns, this measure of risk is provided by the standard deviation, as in the Markowitz model of portfolio management [20]. Nevertheless, a simple statistical analysis of the distribution of equity returns indicates that this distribution is usually not normal but skewed and with heavy tails, as well as being non-stable in time [21]. In this context, and as described by the consumption-based asset pricing model [22, 23], the quantification of risk involves several complex tasks. First, it is necessary to correctly identify the set of all possible states of nature describing the behavior of the considered return, which is complicated when the statistical distributions are not stationary as usually happens in real financial markets. Second, the assignation of probabilities to the identified states of nature is also required, wich is a statistical question implying (as the former) non-trivial economic subjective considerations. Indeed, a trade of assets is only possible if agents have different beliefs on these probabilities and different perceptions of the underlying risk. In simple words, although the distribution of the returns of any equity is the same for all the investors in the market, the quantification of the inherent risk is theoretically only possible through the subjective perception and valuation of this objective distribution by the specific considered investor. As a result, we can have subjective measures of the risk implied by the returns volatility, but not an objective one [23–25]; then, some external criterion must be applied to infer some kind of objective quantification of risk. This complexity for the quantification of risk, which is a relatively simple scheme of uncertainty, becomes much higher when the environment does not allow the set of possible values of the variables and/or the associated (subjective) measure functions to be determined, or when the different sources of uncertainty interact. This is without any doubt the most realistic situation, which, by its nature, requires alternative approaches to correctly quantify uncertainty. Indeed, given that the identification of the feasible outcomes and the assessment of the associated (subjective) probabilities are not possible, suitable methods of objective quantification must use alternative instrumental or indirect evidence of the presence of uncertainty.

In this regard, to date, most of the existing approaches seek to predict and identify market uncertainty through the statistical analysis of the historical behavior of asset prices and market indices, which is an interpretation supported by the EMT and MST. These fundamental theories of finance state that market agents act rationally in processing the available information. Therefore, asset prices must reflect all the relevant information, including that concerning future uncertainty. However, the rather limited success of these price-based models in forecasting uncertainty suggests that this approach needs to be either expanded or modified. In this light, our approach makes a significant contribution by reversing the application of the EMT and MST, thus formulating a model that directly identifies and measures stock market uncertainty by addressing its origin -specifically, the interpretation of the available information by the market agents- rather than inferring from its consequences. To this end, our model builds a digital twin of the process conducted by the human mind, thereby associating market uncertainty with a random, fuzzy, vague, or incomplete comprehension of the available textual information.

Leaving aside technical issues related to the design of the digital twin model, the first question to evaluate in greater depth is the suitability and appropriateness of our approach, especially the use of

textual variables. In this respect, the underlying guiding ideas go from the hypothesis of rationality and its transposition into the EMT and MST, to the subsequent overcoming of their shortages. As explained above, in its origins, EMT and MST reduce the forecasting analysis to a mere statistical study of the historical series of prices. However, although these market-based measures (in any of its formulations, such as probability, credibility, chance, and uncertain measure functions) have shown some ability to anticipate future uncertainty [26–31], their performance is very limited, thereby questioning the use of structured variables alone in providing information on the underlying future uncertainty [32]. The shortages of this chartist approach were contested by fundamentalists, who stressed the great importance that information on economic variables, rather than just prices,has on the market evolution. On this point, the use of unstructured data, particularly news and reports, emerged as a relevant alternative to quantify and predict uncertainty. Indeed, the empirical evidence clearly confirms that news and reports constitute market drivers, and therefore that they can be used to anticipate the future evolution of asset prices [33–38]. Upon these theoretical bases, and facilitated by the availability of data on the information exchanged in news media on the internet and social networks, there has been great growth in the formal and quantitative use of news and other textual forms to explain the future evolution of asset prices and its causes (among them the presence of current textual elements anticipating uncertainty). From the above quoted initial studies measuring the efficiency of asset markets' reactions to news releases, research interests have progressively moved to the design of models that incorporate news as either a direct or indirect predictive variable for asset prices, and/or as a variable capturing and forecasting market uncertainty and making the mitigation of the identified uncertainty possible [39–47].

Our research falls within this line, thus incorporating several new features, both theoretical and technical. First and as commented before, it focuses on the ultimate origin of the stock market uncertainty, namely the interpretation of the available information by the market agents. Second, our model considers textual news as the primary source of information, thus building a digital twin of the cognitive process undertaken by the human mind. Specifically, to identify market uncertainty, this model applies artificial intelligence (AI) techniques to associate market uncertainty with a random, fuzzy, vague, or incomplete comprehension of the available textual information. The result of these theoretical innovative perspectives is a novel and worthwhile approach integrating neuro-biological evidence into digital twin modeling, which has yet to be applied to the study of financial markets. From a technical point of view, the new features are present in practically all the dimensions of the model. Concerning the text analysis process, we take not only how frequent a particular word is in the analyzed period, but also how its use is distributed along this period into account, as well as the particular contexts of its appearances. Then, the contextual and timeline patterns identified by the implemented statistics procedures are then used as inputs in the classification algorithms, which also present new features. In particular, to gain efficiency in the objective of quantifying uncertainty, we sequentially apply a set of data mining techniques without imposing previous target variables, which is another differential attribute of our approach; this leads to the definition of regions of knowledge. Defining uncertainty as the absence of a clear assignation of an informative item to a specific region of knowledge, we propose the use of Shannon entropy to measure the market uncertainty present in the news. Finally, we evaluate the performance of our method, making use of real data and carrying out a comparison with the existing procedures to forecast stock market uncertainty.

The rest of the paper is organized as follows. Section 2 provides a general perspective of

the philosophy behind the proposed digital twin model, the implemented AI techniques, and its contributions to the objective of quantifying and forecasting market uncertainty. Section 3 explains the implemented methods in greater depth, paying special attention to their formal aspects and the clarification of their suitability. Section 4 describes the empirical analysis and explains its peculiarities. The obtained results are presented and discussed in Section 5. Lastly, Section 6 provides some concluding commentaries on the findings.

2. Digital twin model: Contributions

It is obvious that news and stock prices are closely related and influence each other in a complex way. To study these relationships, and given their ability to monitor and handle large amounts of information, AI techniques are of a great efficacy. In this respect, from neural networks to genetic programming, these techniques are being progressively implemented to explain how news, sentiments and opinions affect the stock market behavior, and how they can help to predict its evolution. Logically, the first objective of these machine learning financial models has been the prediction of market returns [48–51], with sometimes contradictory results on the effects of infrequent events; this suggests the need for additional research for these episodes with an unclear and uncertain understanding [52, 53]. Our research moves in this direction by proposing an innovative digital twin model that identifies situations of uncertainty through the replication of the human mind's interpretative process of the available information. In this sense, our model is the natural continuation of existing research in several fields. On the one hand, although recent results show that a digital twin mirroring the human mental processes is conceptually possible and feasible [54, 55], its application to study financial markets in general, and the stock market uncertainty in particular, remains unexplored. On the other hand, the reversal in the reading of the EMT and MST that we propose opens an operational interpretation of uncertainty in asset markets, which is compatible with the neurological evidence found by other authors [56–61] and also not explored to date.

From the technical point of view, our model is a rolling-window system that incorporates the characteristics of long short-term memory networks [62], which are highly suitable for our purpose of mirroring the human brain's information processing mechanisms. Broadly outlined, the replicated process is as follows. First, assuming that market agents, in their role as investors, pay special attention to the presence of a set of selected words and terms, we build an ad-hoc dictionary for the whole temporal horizon. Then, we model the agents' information interpretative process, which, according to the historical evidence and the different contexts of appearance of the different terms, is understood to be a process that learns as time passes. In our model, we achieve this by implementing AI techniques that seamlessly integrate prediction and learning through rolling windows across the entire considered period. As a result, at any period t, our digital twin continuously learns from previous data to predict the subsequent periods from $t + 1$ to $t + n$, thus reevaluating the meaning of the different words and terms according to the past evidence and the current contexts period by period.

To accomplish these tasks, our theoretical digital twin model applies three successive phases: text-to-number transformation, processing of the obtained statistics, and uncertainty measurement. All these stages present some specific characteristics that are worthy of being commented and situated within their general context to properly evaluate our contributions.

2.1. Text-to-number processing

For computers to interpret the human language, it is previously necessary to transform text expressions into something that computers can process and analyze. An immediate first option is the numerical transformation of words into vectors and the subsequent treatment and analysis of the numbers associated to each specific word. Under the general term of natural language processing (NLP), this approach, which has been present from the 1980s [63–66], has evolved to pursue the logical objectives of a higher accuracy in the numerical translation of texts and meanings, and a lower consumption of time and computer resources [67–71]. Initially, these models merely focused on the frequency of occurrence of words (bag-of-words models, [72]), later incorporating the analysis of word order to solve the arising shortages. The complexity of handling word order and not just word frequency created the necessity of using a higher computational power and more efficient techniques, thus giving rise to more complex models applying neural networks. These neural network based models can handle intricate patterns because of their ability to tackle the curse of dimensionality problem, inherited from NLP; today, they are encompassed within the heading neural network language processing (NNLP). In parallel, the development of AI, specially machine learning algorithms and the idea that statistically oriented approaches -such as those present in language acquisition in humans- can provide machine learning from text, lead to significant improvements of the models.

In our context of NNLP, whose general objectives are the information retrieval from texts and its subsequent analysis and treatment, the specificities of each model are the consequence of its particular aims. Regarding our proposal, whose objective is the identification, quantification and prediction of economic uncertainty, this implies a selection of the appropriate text sources and of the informative dimensions on uncertainty contained in these texts. Concerning the first question, the natural and logical option is to consider specialized journals in economics and finance. The determination of the informative dimensions to be captured and evaluated is more complicated, and requires a reflection on how economic uncertainty is manifesting itself in texts, and on how the previous approaches can be improved.

Our contributions regarding text-to-number processing are related to this second aspect. As recent studies show, the impact of news is directly related not only to its frequency of appearance and lifespan, but also to the speed of its spreading, its correlation with other events and news, and how it is replaced by other topics [56–61, 73]. To extract signals on the encapsulated uncertainty from the analyzed texts, we consider not only the frequency of appearance of specific words, as most models do, but also temporal and contextual elements captured by its time distribution and the presence of co-occurrences. As the economic theory shows, a very important characteristic of economic uncertainty is that it simultaneously affects several variables in different ways and with different temporal patterns [74, 75]. These relevant features can be captured through the consideration for each word of its contextual and temporal attributes, which, together with its history of frequencies, originate a high-dimensional vector characterizing the meaning of the word at each moment of its appearance, another distinctive characteristic of our method. On this point, it is worth noting again that our proposal and methods connect with recent studies on the way the human brain understands the combined meaning of words in sentences (i.e., the so called supra-word meaning, see [60, 61] and the references therein).

In addition and unlike the existing models, our text transformation method introduces an unsupervised method for the identification of contexts in a timeline without previous labeling of the data. The methods to understand or identify contexts (to perform document classification) rely

on previously classified target variables, therefore losing the capability of identifying hidden not-contemplated categories, which is an issue solved by our approach. On the other hand, when applied in NNLP, the unsupervised methods have been designed to transform words into numerical vectors without looking for the identification of spatial-temporal contexts. Our proposal does consider these contexts, which in our framework are understood as the contemporaneous (spatial) and the timeline (temporal) contexts of each word. Finally, from the computational perspective, the wide consideration of the co-occurrence of words in these two dimensions enabled by our approach results in a faster training and a higher accuracy.

Once these statistics have been obtained, an artificial neural network, a self-organizing map, and a hierarchical clustering followed by a gaussian mixture model are implemented to efficiently extract the relevant information. These contributions will be described and clarified in the following section.

2.2. Processing of the obtained statistics: Mixture model clustering

After applying our text-to-number algorithms, we count on a statistical characterization for each of the relevant words, which informs us on their frequency of appearance in different contexts and times. Given their multidimensional nature, these statistical distributions must be properly processed to extract the relevant information in a manageable way. For this objective, mixture models are the common practice of processing data points under the assumption of certain underlying probability distributions [76]. The main property of this approach is that each group has its specific probability distribution, with all its elements simultaneously belonging to all the groups but with different chances. Mixture models present the ability of identifying subpopulations alongside unobserved population heterogeneities [77]. Additionally, they are known for their flexibility in fitting data of different types, given the possibility they offer of tuning a large number of parameters. Nevertheless, with an increase of parameters, the attached complexity rises as well. The implementation of mixture models is typically made by maximum likelihood estimation via an expectation-maximization algorithm [78]. To implement this algorithm, the practitioner must be aware of unbounded likelihood functions, as well as facing complex issues related to the number of components, the initialization of the expectation-maximization procedure, and variance estimation.

To solve this and other issues and to broaden the analytical scope, self-organizing maps (SOM) have been tested as potential complementary and/or substitutes of typical mixture models [79, 80]. This approach allows for an efficient simplification of the data informative structure. Indeed, in our work, we used the SOM methodology as a topological reducer of our high-dimensional dataset (in informative terms). After this topological transformation, and with the aim of clustering text, we applied a static Gaussian mixture model of hierarchical clustering. This method seems to provide the best performance for multidimensional data [81–83], which is our case. The application of soft clustering for information retrieval is essential because, by its nature, information is not specific to a particular topic or context: An informative piece can simultaneously influence several topics, contexts and knowledge environments [84]. This is very important in our digital twin model, which is designed to interpret, predict and identify uncertainty in a manner similar to the human perspective, a perspective characterized by the existence of regions of knowledge [57–59]. Indeed, one of its abilities is the recognition of changing movements in the financial markets as signs of uncertainty, as indicated by data points located in cluster borders, and thus not clearly belonging to a specific region of homogeneous knowledge.

2.3. Uncertainty: Concept and measurement

As commented above, the existing financial literature has identified and measured uncertainty mainly by considering its effects on stock prices and its volatility [26–31, 33–38]. In this respect, the main feature distinguishing our model from previous approaches is our interpretation of market uncertainty, paying close attention to its origins and not to its effects. Specifically, on the basis of EMT and MST, our model assumes that market uncertainty is the consequence of the behavior of agents who are not completely sure about the meaning of the information they handle, and proposes to measure uncertainty from this fact.

Our digital twin of the human mind's information processing mechanisms is the result of this novel approach. Since the unclear interpretation of information by the human mind is related to the existence of randomness, fuzziness, vagueness or incomplete knowledge, we develop a model that captures these dimensions. In this respect, two notable innovations stand out. First and as explained above, the set of statistical and AI techniques applied to our text dataset allows us to define regions of homogeneous knowledge/interpretation, and to interpret uncertainty as the absence of a clear assignation of a word to a specific region of knowledge. As a second contribution, we use the Shannon entropy of the processed text series for measuring and forecasting the encapsulated uncertainty taking the degrees of allocation of each daily informative content to the different informative homogeneous regions into account, since it responds to the concept of informative uncertainty described above. On the suitability of the Shannon entropy to measure uncertainty, we refer the interested reader to [85–87].

3. Digital twin model: Methods

3.1. Text transformation

As explained in the previous sections, our text-to-number transformation process has been designed to capture not only the frequency of appearance of the relevant words, but also their contemporaneous and time line contexts. For this purpose, the process of feature extraction from the considered news headlines has the following main lines.

Let u denote each word in the collected news headlines. Let N be the total number of days, and u_n represent the number of times this word appears at day n, $n = 1, 2, \ldots, N$. Since we are interested in characterizing how the appearance of each word u is distributed over time, we compute the skewness and the kurtosis coefficients of the time distribution for each word, informing on the asymmetry and tail thickness of these distributions, respectively. For each word u, these normalized measures, denoted by kr_u and sk_u, are as follows:

$$kr_u = \frac{\sum_{n=1}^{N} \frac{(u_n - \overline{u})^4}{N}}{\sigma_u^4}, \qquad sk_u = \frac{\sum_{n=1}^{N} \frac{(u_n - \overline{u})^3}{N}}{\sigma_u^3},$$

where $\overline{u}$ is the daily mean of the occurrence of word u, and σ_u is the standard deviation of its time distribution.

For each word, the kurtosis and skewness coefficients provide shape measures of its occurrence over time. In addition, to improve our knowledge of the time distribution, we consider the specific day at which each word reaches its 25%, 50%, 75% and 95% percentiles. The selection of the 95% percentile instead of the 100% obeys the fact that, by their nature, most relevant words are going to appear on

the last day, in which case the 100% percentile loses a large part of its informative content. For each word u, these days will be denoted by q_u^i, where $i = 25\%, 50\%, 75\%, 95\%$; obviously, they determine whether the considered word is more frequent in distant or recent moments, and how this presence changes over time.

These measures characterize the time distribution of each word in isolation. In addition and as commented previously, our research seeks to incorporate contextual elements. This is done through the following two variables. For each day, the first takes the number of times that the word u appears in the previous T days into account. In this paper, we have focused on the preceding 7 and 30 days: By fixing $T = 7$ and $T = 30$, we assume that weeks and months have a special relevance in determining temporal contexts, which is a reasonable assumption given the high number of financial reports with a weekly or monthly periodicity. Let $u_{cn}(T)$ be this contextual measure for word u, at day n, and for the T previous days, which is defined as follows:

$$u_{cn}(T) = \left[\frac{(\frac{1}{T-1}) \sum_{i=n-T}^{n-1} u_i - u_n}{u_n} \right] \times u_{idf},$$

where u_{idf} is the inverse document frequency of the word, given by

$$u_{idf} = \log \frac{N}{\text{card}\{d \in N : u \in d\}},$$

where $\text{card}\{d \in N : u \in d\}$ is the number of days for which the word u is observed. The factor u_{idf} is boolean-calculated over the word, and aims to normalize the presence of frequent words, which are less informative. Therefore, for each day n, the contextual variable $u_{cn}(T)$ captures the continuity/discontinuity of the word u during the last T days, weighted by the relative informative importance of this specific word. Note that since this variable is defined for each word and each day in two different time windows (7 and 30 days), we have a total of $2(N - 30)$ contextual variables of this type for each word u.

Contexts are also defined by the coincidence of words. Then, jointly with the former variable $u_{cn}(T)$, which defines contexts from a temporal criterion for each isolated word, we build a second contextual variable to identify those contexts defined by the co-occurrence of words. In terms of information retrieval, this can be done by computing the co-occurrence matrix for the set of considered words. In addition, we also take the temporal dimension of this co-occurrence into account. Let m be the vocabulary length (i.e., the total number of considered words), and let the upper index i, $i = 1, 2, \ldots, m$, denote each specific word u^i. Let $u_n^{ij}(T)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, m$, be the number of times the words u^i and u^j co-occur during the T days previous to day n. To identify the really important pairs of words, we compute the following ratio

$$ru_n^{ij}(T) = \frac{u_n^{ij}(T)}{u_n^{ii}(T)},$$

selecting them when $ru_n^{ij}(T) \geq 0.95$. This means that, during the T days preceding day n, the word u^i appears jointly with the word u^j in at least 95% of the appearances of u^i. From these ratios, the following co-occurrence matrix $M_n(T)$ is calculated:

$$M_n(T) = \begin{bmatrix} ru_n^{11}(T) & \cdots & ru_n^{1m}(T) \\ \vdots & \ddots & \vdots \\ ru_n^{m1}(T) & \cdots & ru_n^{mm}(T) \end{bmatrix}.$$

Since there are two time windows (of 7 and 30 days), it is worth noting again that there is a total of $2(N-30)$ co-occurrence matrixes capturing these contexts defined by the simultaneity of topics.

To summarize, after our text transformation process, for each word, we count on the following: Four variables characterizing its distribution over time; $2(N-30)$ variables informing on its temporal contexts; and $2(N-30)$ variables capturing its topical contexts. The last two are common for all the considered words, the first two being specific for each word. Each of these informative variables is denoted by b_r, r, $r = 1, 2 \ldots, R$, where R is the total number of variables/dimensions.

3.2. Uncertainty quantification

As explained in the former section, to identify and measure uncertainty, we sequentially apply three data mining algorithms on the numerical variables obtained from the texts, namely a self-organizing map, followed by a hierarchical clustering and a Gaussian mixture model. The output of these algorithms is a set of statistical distributions characterizing daily information in terms of regions of homogeneous knowledge, on which we compute the Shannon entropy to quantify day-by-day uncertainty. The guidelines of these steps are explained in the following subsections.

3.2.1. Self-organizing map

In order to work with the huge amount of variables resulting from the text-to-number processing, it is mandatory to reduce the complexity of the obtained output without diminishing its informative content. The implementation of a self-organizing map (SOM) algorithm constitutes the first step in this simplification. More specifically, the aim of this SOM technique is to efficiently condense the N daily multidimensional informative units into a small number of units, the so-called neurons.

Formally speaking, our SOM algorithm is a neural network, trained using unsupervised learning techniques, which in terms of informative content, allows for the optimal topological placement of the informative units (i.e., the neurons). The main guidelines of the proposed SOM algorithm are the following. First, we consider the output generated by the initial text-to-number process. As explained in the previous section, for each word u, this output consists of R variables. For each word u, let $\mathbf{b_u}$ be the R-dimensional vector defined by these informative variables. In a second phase, we compute these vectors for all the words appearing on each day n, and calculate their average. Let $\mathbf{D}_n$, $n = 1, 2, \ldots, N$ be these daily average vectors. Since the components of $\mathbf{D}_n$ are given by the average of the corresponding variable for each word observed on day n, and all these word variables have been previously normalized in terms of their informative content, these daily average vectors $\mathbf{D}_n$ summarize the informative content of the daily production of news, tweets and headlines.

These $\mathbf{D}_n$, $n = 1, 2, \ldots, N$ vectors constitute the input of our SOM, which is run according to the procedure defined by Kohonen [88–90]. Roughly speaking, this procedure considers each component in $\mathbf{D}_n$ (i.e., each informative variable) as a dimension. In the multidimensional space thus defined, this algorithm starts by placing an a priori reduced number of neurons, denoted by $\mathbf{w}_k$, $k = 1, 2, \ldots, K$. Then, these neurons must move and find their optimal positions in the multidimensional information

space according to the position of the N daily vectors-points $\mathbf{D}_n$ and the movement rules established in the training example. In our case, given a daily $\mathbf{D}_n$ informative vector-point, each neuron approaches this point depending on their distance: closer neurons move to the $\mathbf{D}_n$ vector-point more than distant neurons. After iteratively applying this rule for all the daily $\mathbf{D}_n$ points, the process converges and leads to a final position of the neurons, that is to a final neural network. In their final places, these neurons can be envisaged as information environments or information attractors, in the sense of sharing and determining the informative content of the closer days, given by the $\mathbf{D}_n$ points. Logically, the a priori fixed number of neurons -or network size- must be properly chosen while looking for the highest informative homogeneity inside the cluster defined by each neuron, which is compatible with the clear existence of different informative environments or neurons. In the literature, this network size is determined by a simple rule of thumb. Here, we follow a more rigorous method by applying the silhouette coefficient, which informs on the number of neurons from which the network size tends to be more ineffective. Basically, the silhouette coefficient measures the homogeneity inside each informative environment-cluster, defined from the associated neuron. The idea is that as the number of neurons increases, the average silhouette coefficient score decreases. As a result of the higher granularity of the map, some daily data points become close to data points assigned to other clusters, and the assignation of the daily informative points to a neuron tends to be fuzzier. Therefore, to choose the optimal number of neurons in terms of this trade off between informative homogeneity and informative difference, we must identify the inflection point in the silhouette coefficient. Once this network size is optimally determined, these neurons $\mathbf{w}_k$, $k = 1, 2, \ldots, K$, are hierarchically clustered following Ward's method, which is justified and explained in the next subsection.

3.2.2. Hierarchical clustering

In terms of topological visualization, the resulting neural network provides a snapshot of the informative patterns in the collected news. However, we are interested in an interpretation of uncertainty similar to that arising from the human perspective, characterized by the existence of regions of knowledge [57–59]. With this aim of reducing the neural network (i.e., the number and relative position of the neurons) into interpretable informative regions, we apply an additional data mining technique over the network of neurons, more specifically, a hard clustering process. The proposed hierarchical clustering algorithm follows Ward's method, which has proven to be resilient in this type of task.

In essence, our hierarchical clustering process begins by considering that each neuron is a cluster itself. Then, the algorithm progressively joins the clusters-neurons $\mathbf{w}_k$, giving rise to merged clusters which look to minimize the variance inside the resulting merged clusters [91]. In our model, this criterion can be envisaged as pursuing the greatest informative homogeneity for the neurons inside the cluster, therefore justifying its application. At the end of this clustering phase, all the original neurons $\mathbf{w}_k$ -and thus all the original daily input data $\mathbf{D}_n$ linked to them- are assigned to a specific cluster C_z, $z = 1, 2, \ldots, Z$. As explained in Subsection 4.2.1, the number of clusters is optimally determined in terms of the trade-off between the operational simplification and the conservation of information.

3.2.3. Multivariate Gaussian distribution

As explained, the previous phases result in a hard-clustering assignment, where each daily data point $\mathbf{D}_n$ is fully attached to a single cluster. However, keeping the concept and meaning of uncertainty and its determination in mind, it is necessary to classify each data point according to the probability of belonging to any of the available cluster representatives. As commented before, each cluster can be envisaged as a homogeneous and intelligible informative environment. Therefore, by determining the probability of belonging to each cluster C_z for each daily informative unit $\mathbf{D}_n$, we are characterizing how confident we are about the right meaning and interpretation of the observed daily information $\mathbf{D}_n$.

To accomplish this task we apply a soft-clustering method, where all the daily points $\mathbf{D}_n$ belong to all the clusters C_z with different probabilities, which are calculated according to the multivariate Gaussian probability distribution. For each cluster C_z, $z = 1, 2, \ldots, Z$, let $P(C_z)$ be its probability of occurrence (in informative terms), given by the following expression:

$$P(C_z) = \frac{N_z}{N},$$

where N is the total number of days and N_z is the number of daily informative data $\mathbf{D}_n$ assigned to the cluster after the aforementioned hierarchical clustering process. Let $\overline{C}_z$ be the cluster mean, given by the average of the daily informative data $\overline{\mathbf{D}}_n$ assigned to the cluster, and let Σ_{C_z} be its covariance matrix.

Then, for any daily informative point $\mathbf{D}_n$, the probability of its occurrence conditional to its membership to cluster C_z is given by the following expression:

$$P(\mathbf{D}_n \mid C_z) = \frac{1}{\sqrt{2\pi \mid \sum_{C_z} \mid}} exp\{-\frac{1}{2}(\mathbf{D}_n - \overline{C}_z)^T \Sigma_{C_z}^{-1}(\mathbf{D}_n - \overline{C}_z)\}.$$

Therefore, the probability of being in the informative environment defined by cluster C_z once the informative daily data $\mathbf{D}_n$ has been observed, is as follows:

$$P(C_z \mid \mathbf{D}_n) = \frac{P(\mathbf{D}_n \mid C_z)P(C_z)}{\sum_{z=1}^{Z} P(\mathbf{D}_n \mid C_z)P(C_z)}.$$

Once these conditional probabilities have been calculated, we can proceed to the computation of the Shannon entropy associated to each daily informative unit $\mathbf{D}_n$, which we take as a measure of the encapsulated uncertainty.

3.2.4. Entropy

One of the novelties of our model is the consideration of the Shannon entropy as a convenient and reasonable measure of the uncertainty enclosed in the collected news. This interpretation of uncertainty, which is a consequence of the design of all the previous steps, lies in the modeled existence of homogenous informative environments (i.e., the clusters C_z) and the different degrees of membership of each daily informative unit $\mathbf{D}_n$ to each of these informative environments. In short, when a daily information unit is at 99% assigned to a specific cluster, this daily information can be very accurately identified with a unique single informative environment, and the uncertainty is almost negligible. However, when the daily news is a source of several topics with no clear assignation to a particular informative homogeneous environment -in terms of a great similarity of the probabilities of

membership of the daily informative vector to several clusters-, the uncertainty degree encapsulated in the daily news is obviously much greater. This is just the quantification provided by the Shannon entropy of the daily informative units $\mathbf{D}_n$, denoted by $H(\mathbf{D}_n)$ and defined as follows:

$$H(\mathbf{D}_n) = -\sum_{z=1}^{Z} P(C_z \mid \mathbf{D}_n) \log_2 P(C_z \mid \mathbf{D}_n).$$

In summary, our proposal is able to obtain a consistent and coherent quantification of financial uncertainty by using unstructured data (news, tweets and headlines) as input, which are conveniently processed to allow the daily Shannon entropy (our proposed measure of the encapsulated uncertainty) to be computed.

3.2.5. Measuring and forecasting uncertainty

As explained in the previous sections, our model relies on measuring and predicting market uncertainty by conveniently transforming the information encapsulated in the text news into numerical variables, thus extracting the relevant information. To do so, these numerical variables are classified and interpreted by applying AI techniques, thus opening up the possibility of quantifying the underlying uncertainty through the computation of the Shannon entropy associated to the daily information. Since uncertainty is associated with periods of turmoil, extreme volatilities and strong tendency changes in financial literature, we must design a procedure able to identify these periods from our data of daily Shannon entropy.

Concerning this point, we propose to predict future periods of uncertainty in stock markets by identifying previous and significant changes in our textual daily informative environments, as characterized through the computation of their Shannon entropy. This is made in two steps. First, to capture trend elements in the information flow, we calculate a moving average for the daily Shannon entropy. Given that 14 natural days (two weeks, 10 working days) is a reasonable length in news duration (see [73]), we consider the window of the 14 previous natural days for the moving average, denoted by $\overline{SE}_{14}$. Then, to identify changes inside these tendencies, we consider the slope gradient $\nabla(t)$ of $\overline{SE}_{14}$for each period, calculated as follows:

$$\nabla(t) = \mid \max_{t-7:t} \overline{SE}_{14} - \min_{t-7:t} \overline{SE}_{14} \mid .$$

Regarding the length of the period used to identify informative changes, related to the calculation of $\nabla(t)$, we consider the previous seven natural days (i.e., the previous 5 working days). According to our reasoning, informative uncertainty is associated with the presence of short periods of significant changes in the type of information and/or increases in information heterogeneity, both leading to extreme values for ∇ (i.e., to outliers). Following a well established rule to identify outliers, they appear when $\nabla(t) \geq (Q3 + 1.5 \times IQR)$, where IQR is the interquartile range. Assuming a proper functioning of our model, these outliers should predict the presence of uncertainty in the stock market, that is, they should advance the future occurrence of periods of turmoil, extreme volatilities and/or strong tendency changes in stock market prices. This is the model that we test with real data in the following section.

4. Experiment

4.1. Data

As explained above, our research makes use of two types of data: Textual data, processed to extract the uncertainty encapsulated in news by applying AI techniques, and numerical indices of uncertainty, which are used to validate model predictions. The success of any forecasting model strongly depends on the quality and quantity of the data sources, which are the characteristics that have guided our selection of the data sources.

In terms of news, the relevant and trustworthy sources must reach large audiences, since they lead to more significant impacts and will generate more interest in agents, practitioners and researchers. In this study, the news source is the Wall Street Journal, which is a daily publication reaching 42 million digital readers per month, and is the main reference in the world of finance [92]. More specifically, text data have been retrieved from the public historical dataset stored in the Wall Street Journal archive from January 01 2007 to December 31 2019. In this dataset, covering 4,748 days, the total amount of headlines considered as input were 739,726, from which we obtained 3,772,613 words labeled by the respective article category.

To validate our model, we consider the Chicago Board Options Exchange's CBOE Volatility Index (VIX) [93] and the Bears Power (BP) indicator [94] for the same period of time. The VIX index provides a measure of stock market volatility, and is the most used index for measuring uncertainty. For its part, the BP indicator allows changes in trends and regimes to be identified.

4.2. Text pre-processing

As referred to above, the data source is the Wall Street Journal Archive from January 01, 2007 to December 31, 2019. A total of 739,726 headlines were considered, and then screened and processed according to the methodology of our model. The specific methods and techniques constituting our contributions to the literature, as well as the logic behind them, have been explained in detail in the previous sections. Therefore, here we only discuss the operational questions arising from the particular characteristics of our dataset, which must be ad-hoc designed, and are not related to the model goal, namely the quantification of uncertainty. All the work was performed within Python® 3.8 with the supercomputer resources provided by the supercomputing center foundation of Castilla and León (SCAYLE) [95].

From the point of view of the question to be analyzed, real-world data contain features ranging from irrelevant or redundant to deeply important. The selection of the truly relevant features depends on the AI techniques to be used in the investigation [96], which determine the type of so-called pre-processing practices. In our case, our method relies on the quality of the information extracted from the news headlines and the applied AI techniques are those of NLP; therefore, the ad-hoc pre-processing tasks must have a mainly semantic nature.

Besides being a finance journal, the WSJ releases news not related to economics and finance, thus acting as a noise generator; therefore, the first step in preparing the data for our model is the elimination of these irrelevant headlines and words. As commented on above, we count on 739,726 headlines, that the WSJ archive has previously labeled by a word defining the related topic. More specifically, the collected headlines were classified into 5,691 topics, many of them with no interest for our purposes.

In order to keep only finance-related news, we used the regular expression (regex) patterns in Table 1 for screening the initial 5,961 topics.

Table 1. Headline topics for regex screening.

.*business.*	.*market.*	.*econom.*	.*money.*	.*invest.*	.*financ.*	.*bank.*	.*stock.*

After the regex screening of the original 5,961 topics and 739,726 headlines, the resulting dataset is comprised of 573 topics (9.6% in terms of topics) and 186,001 headlines (25.1% in terms of headlines).

Some of these final headlines contained alphanumeric values. In this respect, we decided to remove the numerical characters, given that the information provided by numbers is only understandable in terms of context and we already count on four contextual variables for each word. With the same aim of avoiding redundancy, lemmatization was implemented to reduce the inflectional forms -which have only contextual meaning- into a common root.

Another important aspect to consider in the process of information retrieval is the existence of groups of words with a unique assigned meaning, the so-called n-grams. For example, "United States" must be considered as a single piece of information, because isolating "United" from "States" generates both new spurious items and loss of true information. To tackle this issue, we proceeded to detect the existing n-grams. After the analysis of their informative relevance, we decided to focus only on those bi-grams appearing more than 30 times, thereby transforming them into single pieces of information. This is reasonable, since we are dealing with headlines in which n-grams with more than three words are practically non-existent. Also, when present, all their constituent words provide relevant information.

After this initial pre-processing phase of the text data and the subsequent noise reduction, we count on a proper set of vocabulary, which constitutes the input of our model.

4.2.1. Feature selection and number of clusters: Specificities

The data obtained after the screening phase were processed by the vectorization algorithms described in the methodological section. The results are the daily informative vectors (defined in Section 3.2.1) $\mathbf{D}_n$, $n = 1, 2, \ldots, 4748$, each of whose components measures an informative dimension. However, given that our objective is to identify, quantify and predict uncertainty, it is necessary to remove those informative features that can be predicted. There are two reasons behind this reduction of the information dimensionality. The first one meets the need to ensure an accurate identification and quantification of the uncertainty degree, which is the main goal of the model. In this respect, it is obvious that when an informative dimension in $\mathbf{D}_n$ can be systematically explained or forecasted by any of the others, this dimension does not entail idiosyncratic uncertainty, and must be removed to avoid a double accounting of a unique and common uncertainty element. As related to the first, the second reason has an operational nature, since the elimination of redundant dimensions entails a simplification of calculation and gains of efficiency by avoiding the curse of dimensionality.

In the related literature, dimensionality reduction is mainly carried out by applying principal component analysis. This method is essentially a technique that identifies the underlying correlations, thus proposing a change of basis on the data enabling the identification -and removal- of the less explicative dimensions. Since this method is based on the analysis of the existing correlations, here, in order to avoid additional manipulations of the data, we decided to apply Pearson's correlation and

remove those features or dimensions that could be explained by others (i.e., showing a high correlation coefficient with any of the others).

Table 2 collects the Pearson correlation coefficients between our informative dimensions. Assuming that each category of variables (i.e., time distribution, contextual, and cumulative frequency variables) share the type of informative content, the removal of dimensions/features has been carried out inside each category. Concerning the word context variables $u_{cn}(T)$ and $M_n(T)$, correlations above 0.89 enable us to drop $u_{cn}(30)$ and $M_n(7)$. Kurtosis kr_u and skewness sk_u are able to explain each other, so we kept kurtosis. Concerning the cumulative approach, we drop variables with a correlation above 0.82.

Table 2. Pearson's correlation for feature selection.

Pearson's Correlations										
	$u_{cn}(7)$	$u_{cn}(30)$	$M_n(7)$	$M_n(30)$	kr_u	sk_u	q_u^{25}	q_u^{50}	q_u^{75}	q_u^{95}
$u_{cn}(7)$	1.00	0.91	0.98	0.89	-0.22	-0.26	0.35	0.32	0.29	0.24
$u_{cn}(30)$	0.91	1.00	0.89	0.98	-0.23	-0.27	0.38	0.35	0.32	0.27
$M_n(7)$	0.98	0.89	1.00	0.92	-0.22	-0.25	0.36	0.33	0.30	0.25
$M_n(30)$	0.89	0.98	0.92	1.00	-0.23	-0.27	0.39	0.36	0.33	0.27
kr_u	-0.22	-0.23	-0.22	-0.23	1.00	0.97	-0.40	-0.49	-0.41	-0.52
sk_u	-0.26	-0.27	-0.25	-0.27	0.97	1.00	-0.37	-0.46	-0.40	-0.51
q_u^{25}	0.35	0.38	0.36	0.39	-0.40	-0.37	1.00	0.95	0.90	0.82
q_u^{50}	0.32	0.35	0.33	0.36	-0.49	-0.46	0.95	1.00	0.95	0.90
q_u^{75}	0.29	0.32	0.30	0.33	-0.41	-0.40	0.90	0.95	1.00	0.96
q_u^{95}	0.24	0.27	0.25	0.27	-0.52	-0.51	0.82	0.90	0.96	1.00

One of the most relevant steps in the word processing task is the definition of the numbers of neurons and clusters. As commented in the methods section, these numbers are optimally determined. In this respect, by computing the silhouette coefficient, we obtain that the optimal number of neurons is 100. Concerning the number of clusters, we used the Elbow graph to optimally determine the number in terms of the trade off between the operational simplification and the conservation of information.

The Elbow graph is represented in Figure 1. Although the discussion of the logic behind the Elbow graph criterion is beyond our sphere of interest, here we briefly describe its foundations and its relationship with our objective. As explained in the methodological section, clusters in our model group daily informative units with a similar informative content. The idea is that the distance of a daily informative unit to the center of a given cluster measures how similar the informative content of the daily unit is to the information represented by the cluster. The greater this distance, the more distinct their respective informative contents. In the Elbow graph, the y-axis (inertia) represents the distances of the daily informative units to their closest cluster center, which can be interpreted as the informational error in its best assignation. Then, the greater the y-axis value, the greater the informative discordance between the daily units and their appropriate clusters/informative regions. Obviously, inertia is at its maximum when there is only one cluster. Moreover, as the number of clusters increases, the inertia decreases since it is feasible to better allocate each daily informative unit, with the value of the inertia being zero when each daily unit becomes its own cluster. Our approach requires not only the preservation of the informative content, but also the construction of informative homogeneous regions/clusters. Then, from the clustering with total informative content (i.e., zero

inertia and each daily unit being its own cluster), we must increase the number of clusters looking for acceptable increases in the informational errors of the arising assignation. Then, it is clear that the inflection point in the Elbow graph provides a useful and reasonable criterion to determine the optimal number of clusters. As Figure 1 shows, the inflection point with our data occurs between three and five clusters. Therefore, we choose four as the optimal number of clusters to run our hierarchical clustering algorithm. After these transformations, we proceed to compute the Shannon entropy as explained in the methods section.

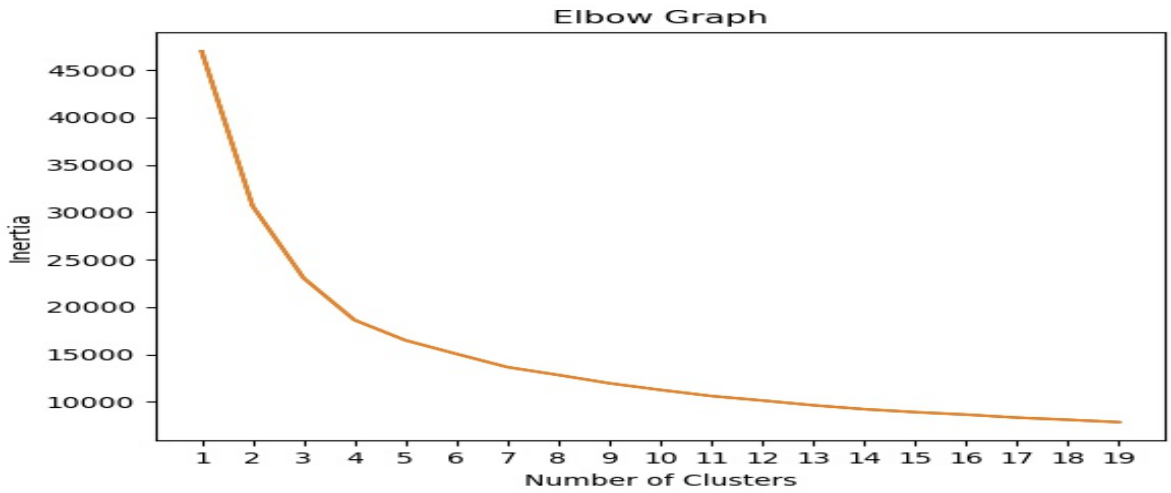

Figure 1. Elbow graph: Determination of the optimal number of clusters.

4.3. Model comparison

In order to test whether or not our proposal constitutes a methodological improvement in identifying and predicting stock market uncertainty, we compare our approach against the standard algorithms. Logically, on the basis of the EMT and MSH, the performance of both our model and the alternative algorithms are evaluated, taking the stock market prices as reference. In this respect, to test the robustness of the different models, the real presence of uncertainty, as indicated by periods of turmoil and extreme volatilities and/or strong changes in prices, is identified by applying the usual market indices, as well as a new one based on the Shannon entropy.

4.3.1. Alternative algorithms

As explained in Section 3, our model predicts stock market uncertainty by transforming the information encapsulated in the text news into numerical variables. These numerical variables allow the relevant information to be measured on a daily basis through the calculation of the daily Shannon entropy. Upon this basis, and given that in the financial literature uncertainty is associated to periods of turmoil, extreme volatilities and strong changes, we have designed a procedure which is able to identify and forecast these periods from our data of daily Shannon entropy.

To consistently compare our method, this must also be the procedure to identify and predict uncertainty for the alternative algorithms. In this respect, the algorithms to be tested are those

commonly used in the literature: k-means, self-organizing maps, and expectation-maximization (EM). As explained above, the metric we apply in our model is the daily Shannon entropy, and therefore these three alternative methods need to be conveniently adapted to ensure a coherent comparative analysis in terms of this metric. Since K-Means and SOM assign each daily data object to only one of the considered clusters and the computation of the daily Shannon entropy is no longer possible, we have decided to compute the Shannon entropy on a monthly basis according to the formula:

$$H(m) = -\sum_{z=1}^{Z} \frac{M_z}{M} \times \log_2 \frac{M_z}{M},$$

where m is each specific month, M_z is the number of times the corresponding cluster C_z is assigned during the month (i.e., the number of daily data in the month assigned to the cluster), and M is the total number of month days.

On the contrary, for the EM algorithm, the evaluation of the daily Shannon entropy is perfectly feasible through the application of the procedures, as is explained in the methodological Section 3.2.

Once the Shannon entropies have been calculated for the four models (i.e., our proposal, k-means, SOM and EM), the periods with the highest uncertainty predicted by each model are determined by applying the common procedure previously detailed.

To implement the k-means algorithm we used the library scikit-learn (v0.21.3) with randomized initial seeds. The implementation of SOM used the sompy (v1.1) library, with a 2×2 map size, random initialization, iterations equal to 1,000, and hexa lattice. The EM algorithm was implemented via scikit-learn (v0.21.3) with the module mixture.GaussianMixture, and a maximum number of 1,000 iterations. The interested reader can find exhaustive information on these algorithms in [97–100]. After applying the Elbow graph, in line with the results in our model, all the algorithms were designed to produce four final clusters.

4.3.2. Market indices

Based on the EMT and the MSH, we validate our model directly against the market prices. In the literature, the occurrence of uncertainty in the stock markets is typically identified by strong inflections in prices, volatilities, crashes, and bullish and bearish behaviours [101, 102]. To evaluate the performance of our model and of the alternative algorithms against these market measures of uncertainty, we selected the Chicago Board Options Exchange's CBOE Volatility Index (VIX) [93] and the Bear Power (BP) index [94].

As is well known, the VIX index measures market volatility, and is therefore useful to identify periods of real uncertainty characterized by the presence of high volatility in stock prices. For the VIX index, with the aim of avoiding data manipulations, we consider that the outliers identifying extreme volatility appear when

$$VIX \geq (Q3 + 1.5 \times IQR).$$

Alternatively, since the BP index is formulated as the difference in a moving average, this index is appropriate to identify uncertainty caused by changes in cycles and trends. Again, to avoid data manipulation, we directly consider its value. Taking the fact that bearish markets are related to a lack of trust and uncertain situations into account, we assume that uncertainty environments are those showing the lower values for the BP index. Then, for the BP index, relevant outliers are defined as

follows:

$$BP \leq (Q1 - 1.5 \times IQR).$$

Finally, to test robustness, we compare our model against a new index capturing both the evolution and volatility of market prices, formulated by applying a methodology similar to that considered in our model. This proposal, denoted as the price entropy index, implies a broad definition of uncertainty, and makes its quantification possible through the application of the Shannon entropy concept. The considered stock market prices p are those in the Dow Jones Industrial Average (DJI) [103], from which we compute the index as follows:

(1) Calculation of the daily closing price variation $\triangle(t) = \frac{p_t - p_{t-1}}{p_{t-1}}$, of its whole range along the N days, and of the following purposely specific quintiles Q: $Q_1 = (0, \infty)$, $Q_2 = (-0.025, 0]$, $Q_3 = (-0.05, -0.025]$, $Q_4 = (-0.075, -0.05]$, $Q_5 = (-\infty, -0.075]$.

(2) For each day t, the determination of the quintile distribution of $\triangle(t)$ along the window from $(t - 7)$ to $(t + 7)$, denoted by $Q(t)$.

(3) Calculation for each day t of the Shannon entropy associated to $Q(t)$.

Since optimism and an absence of uncertainty is associated to increases in stock prices, we have considered a single quintile for all the positive values of $\triangle(t)$. The consideration of a window of 14 natural days (i.e., 7 natural days or 5 working days) before and after each day t, $(t - 7 : t + 7)$, representing 2 natural weeks, is a reasonable period to identify changes in tendencies or extreme volatilities. The entropy calculation in step 3 was made with the scipy (v1.3.1) library. Once this Shannon entropy of market prices has been calculated, we apply the same procedure implemented for our model to detect moments of a higher uncertainty. In the next section, we present and discuss the results obtained from this comparison experiment.

5. Results

Once the methods explained in the previous sections have been applied, we count on the outliers generated by our model, the three alternative models (K-Means, SOM and EM-GMM), and the three market indexes (VIX, BP, and Price Entropy). Since we have to establish how well each of the four considered models predicts the real uncertainty as it stands in the three market indices, these outliers constitute the main output to be analyzed. Logically, it is previously necessary to define the applicable time window of forecasting for the four models, that is, the number of days previous to a real presence of uncertainty, along which it is considered that a model predicts that actual uncertainty. On this point, there is no consensus in the literature. Indeed, in studies making use of daily data such as ours, windows ranging from the 250th to the 30th previous days can be found, with no a clear criterion of decision [104–107]. Therefore, we have decided to establish a window of 60 days, which is the same as in Schema Design and Google Trends [73], mainly because that study covers news with a similar nature to ours, and analyzes their lifespan in the country where the WSJ -our text data source- is based.

5.1. Comparison of predictive accuracies

The comparative analysis of the performance of the models is carried out by applying the usual concepts of sensitivity and specificity [108]. These measures of the ability of a predictive method provide the ex-post probability of the considered model to correctly predict the future presence (sensitivity) or absence (specificity) of a real uncertainty period. As explained above, our models identify and predict uncertainty periods through the calculation of outliers. Then, the concept of sensitivity aims to penalize the inability of the model to identify true outliers within the time range, while specificity is designed to penalize models showing an over-response and the appearance of false positives/outliers. Additionally, the predictive capability of the different models is assessed by computing the area under the receiver operating characteristic (ROC) curve, abbreviated as AUC. Roughly speaking, for each model, the ROC curve displays the sensitivity (the rate of true identified outliers) as a function of the rate of falsely identified outliers for all possible threshold values determining true positives. Thus, the AUC is a robust overall measure of the performance of the considered model in predicting the presence and absence of outliers/uncertainty, because its calculation relies on the complete ROC curve and considers all possible prediction thresholds. The AUC value is within the range $[0, 1]$, where the minimum value represents a totally inaccurate prediction, 0.5 is associated to a random forecast, and the maximum value corresponds to a perfect prediction. In general, an AUC value between $[0.7, 0.8]$ is considered acceptable, in the interval $[0.8, 0.9]$ is considered excellent, and is outstanding above 0.9. The interested reader on AUC can consult [109–111].

Table 3 displays these three measures for the four considered models and the four different forecasting scenarios. These scenarios are defined as follows: real uncertainty arising from the VIX, the BP, and our price entropy indices; and a global fourth scenario for which real uncertainty arises from at least one of these former three indexes. As this table shows, our proposed model achieves much better results compared to its counterparts, particularly concerning sensitivity and AUC. Regarding specificity, although k-means and SOM present higher values and correctly forecast all the absences of real uncertainty, they are not representative of a good predictive capability. On the contrary, given their zero or close to zero sensitivity values, their forecast strategy can be envisaged as the continuous and systematic prediction of uncertainty absence. Indeed, the percentages of total right predictions (i.e., those for real presence and real absence of uncertainty, the so called accuracy value) for k-means and SOM are the lowest among the four considered models, while our model presents the highest accuracy.

Table 3. Performance of the considered models: Sensitivity, Specificity, Accuracy, and AUC.

Scenario	**Sensitivity**				**Specificity**			
Market Index	*K-Means*	*SOM*	*EM-GMM*	***Model***	*K-Means*	*SOM*	*EM-GMM*	***Model***
Price Entropy	0.06108	0	0.31414	**0.65969**	1	1	0.98195	**0.98684**
VIX	0.01562	0	0.19271	**0.92187**	1	1	0.97566	**0.98158**
BP	0	0	0.18033	**0.46721**	0.99968	1	0.97813	**0.98842**
Global	0.056	0	0.304	**0.6496**	1	1	0.98235	**0.99156**

Scenario	**Accuracy**				**AUC**			
Market Index	*K-Means*	*SOM*	*EM-GMM*	***Model***	*K-Means*	*SOM*	*EM-GMM*	***Model***
Price Entropy	0.82147	0.81033	0.86603	**0.94183**	0.53054	0.5	0.64944	**0.82708**
VIX	0.92945	0.9245	0.95421	**0.99752**	0.50781	0.5	0.58428	**0.95583**
BP	0.94647	0.94616	0.95297	**0.98205**	0.49984	0.5	0.57833	**0.728435**
Global	0.80538	0.79424	0.85303	**0.93441**	0.528	0.5	0.64432	**0.82281**

As shown in Table 3, our model successfully predicts real market uncertainty. For the main measures of predictive quality -namely sensitivity, accuracy and AUC- and for all the considered scenarios, our proposed model obtains the best results by far. In particular, our model presents a very high predictive capacity of the real uncertainty periods identified by the VIX market index, and does so without penalizing its specificity value (i.e., without wrongly predicting the absence of real uncertainty). Given the popularity and importance of this index, this is an evident advantage of our model. It is worth noting that our model provides an outstanding AUC value of 0.95583 for this VIX index, and the AUC values are excellent for the *price entropy* and *global* market indexes. The AUC value of our model for the BP index is 0.72843, which is much higher than those attained by the K-Means, SOM and EM-GMM models, and is within an acceptable range.

Figures 2–5 show the performance of the considered models for the global scenario, which is the more general one. In addition, the interactive Figure 6, which is provided in the supplementary section and is publicly available at GitHub (https://github.com/BeforePublication/Stock-market-uncertainty-determination-with-news-headlines), allows the distinct model performances for every scenario to be inspected in more detail. In these figures, for each of the considered models, green denotes a correct anticipation of a real occurrence of uncertainty (according to the visualized index), and red represents an anticipated period of uncertainty that does not actually happen. Analogously, for the different market index graphs, green indicates that the period of (real) uncertainty has been correctly predicted by the model, whilst red denotes the existence of real market uncertainty not predicted by the considered model. In this respect and as these figures depict, k-means and SOM closely follow the market evolution, thus providing valuable insights into market trends. However, due to their conservative hard clustering approach, they do not provide accurate predictions of real future uncertainty episodes in terms of sensitivity, and this happens for all of the considered market indexes. Regarding the EM-GMM model, although it predicts future real market uncertainty better than k-means and SOM, its sensitivity value is no greater than 0.31414, which is clearly unsuccessful. In comparison with the k-means and SOM models, this improvement is a consequence of the EM-GMM's higher ability to

mimic the market volatility; in any case, this improvement does not allow for a satisfactory prediction of important uncertainty periods, such as those characterizing the subprime crisis.

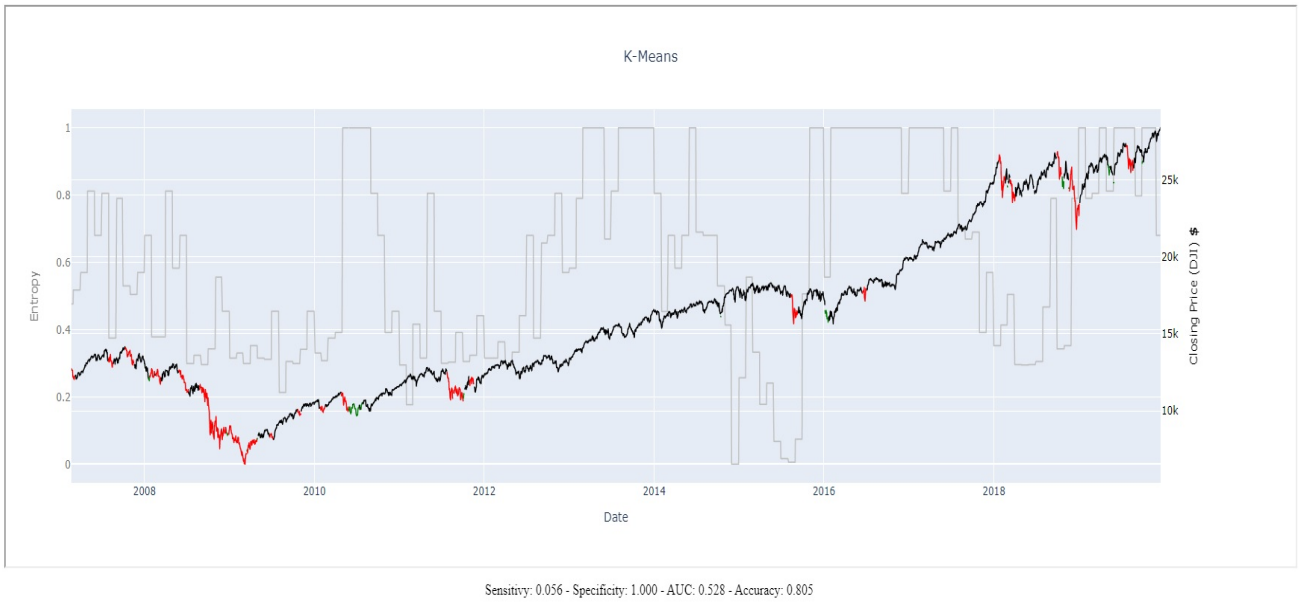

Figure 2. Global scenario: Performance of K-Means model.

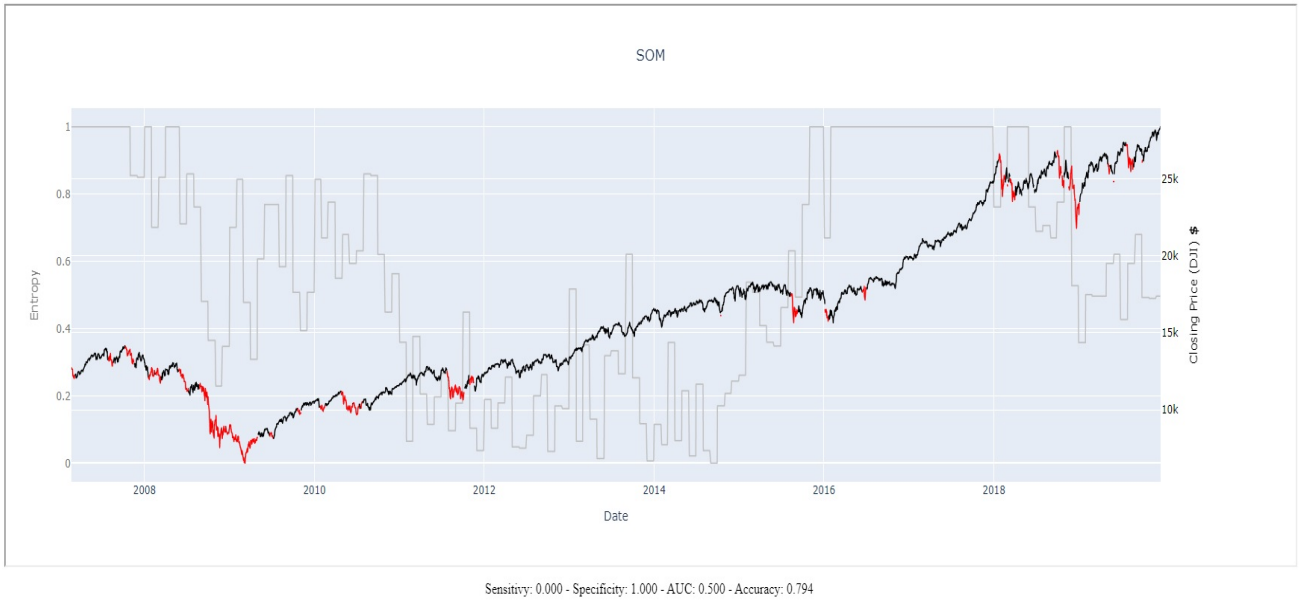

Figure 3. Global scenario: Performance of SOM model.

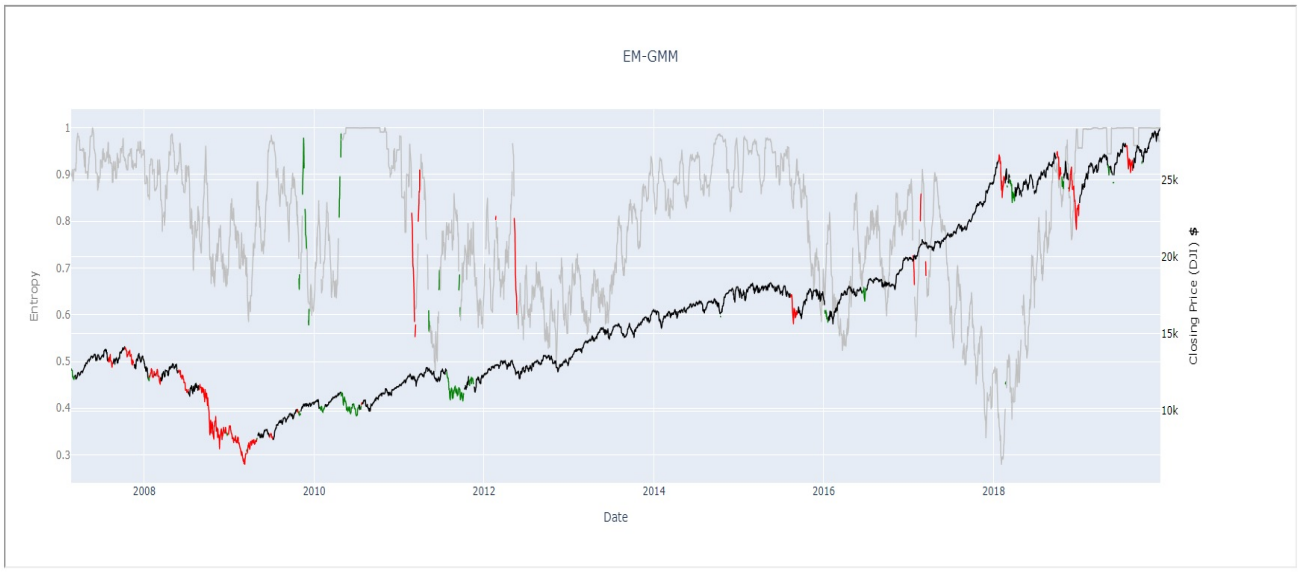

Figure 4. Global scenario: Performance of EM-GMM model.

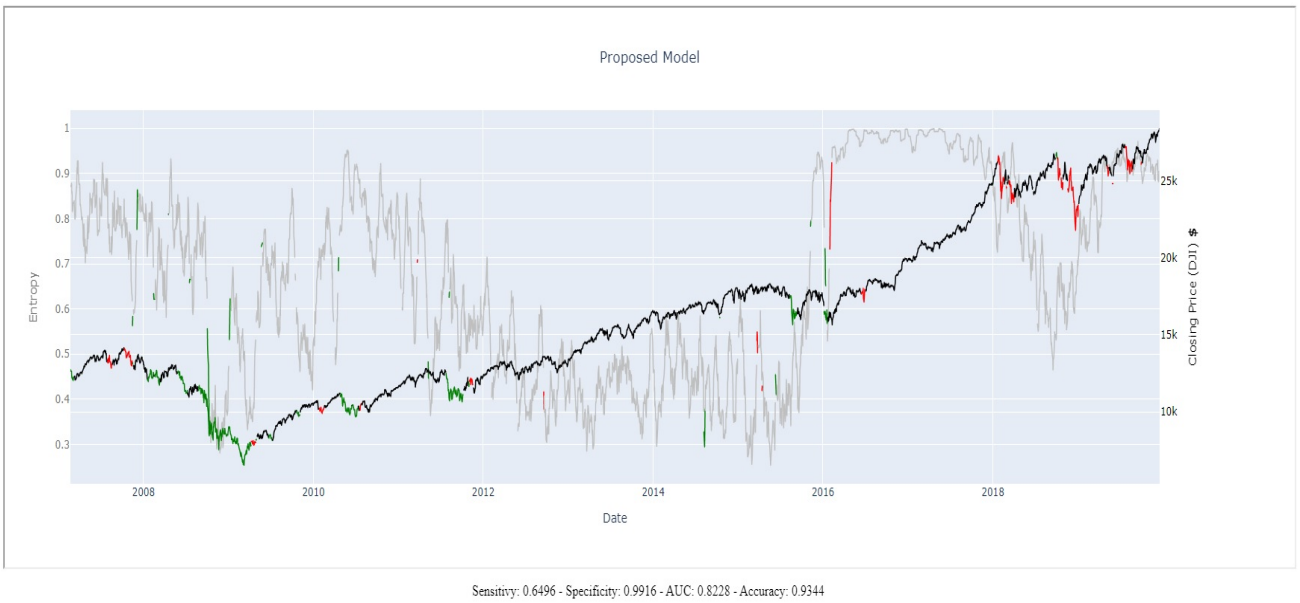

Figure 5. Global scenario: Performance of the proposed model.

As evident from its sensitivity and specificity values, the superior performance of our proposed model is also graphically manifested. In particular, Figure 2 shows (in green) how the future real occurrence of market uncertainty is correctly predicted by our model, especially in the 2008-2018 subprime crisis. Additionally, it is worth noting that most of the errors in predicting uncertainty (in red) occur in the final periods of the sample, when the Global Financial Crisis was at its end and the uncertainty episodes were less intense. Indeed, according to the VIX index, there are no market

uncertainty periods after 2016, which is the reason behind the excellent performance of our model for this market index. This fact can be interpreted from at least two perspectives. The first suggests that our methodology is particularly suitable for the prediction of episodes of high uncertainty when uncertainty is understood as volatility. The second points to the need for further analyses of the proper definition of real market uncertainty, since this definition determines the accuracy of the distinct models in forecasting real uncertainty. As commented above, our proposed model is highly suitable when real uncertainty is related to volatility, which happens for the VIX index. When market uncertainty is understood as the existence of changes in cycle and trend (as detected by the BP index), our model again provides the best predictions, although with a lower sensitivity than for the VIX index.

As explained in Section 4.3.2, to tackle these shortcomings arising from the alternative definitions of real uncertainty, and in order to check the robustness of our model, we have elaborated a more general market index, the price entropy index. Given its methodology, this index should capture a real uncertainty arising both from extreme volatility and changes in the cycle and trend. In this respect, the high similitude for all the considered measures -sensitivity, specificity, accuracy and AUC- and for all the models between the values obtained for the price entropy and global scenarios, suggests that our price entropy index correctly identifies periods of real uncertainty characterized by both price volatility and changes in cycle and tendency. For this market index/scenario, our model again provides the best forecast results, and presents the highest values for sensitivity, accuracy and AUC. This implies that our model successfully predicts episodes of real uncertainty defined as periods of high volatility in stock prices, showing change of cycles and trends in prices, or presenting the co-occurrence of these features.

5.2. Statistical significance

As is well known, the usual measures of goodness of fit used in regression models can be misleading in qualitative response models [112–114]. As a result, in the context of binary response models such as the ones considered here, sensitivity, specificity, accuracy, and AUC are considered as the most suitable metrics to determine how well a model predicts a binary condition (e.g., the presence or absence of market uncertainty in our study) [108–111]. According to these measures, and as discussed in the previous subsection, it is evident that our digital twin provides much better predictions than the alternative standard models. However, all these measures provide descriptive ex-post probabilities, but do not offer information on the statistical reliability of this better performance. In this respect, not only due to the reasons mentioned above but also to its rolling window nature, our model does not allow for an evaluation of its statistical significance through the calculation of out-of-sample R^2 statistics; therefore, alternative methods must be considered. In fact, one of the main characteristics of our approach is its evolutive nature, which involves a continuous reevaluation of the underlying parameters that renders out-of-sample prediction nonsensical. At any given time t, the model does not assume predefined parameter values enabling out-of-sample prediction; rather, these parameter values are defined through the prediction itself. This lack of a closed and concrete time-invariant specification for our digital twin model also prevents the direct application of the standard statistical hypothesis tests (for instance the Clark and West test or the Giacomini and White test) to compare the predictive ability of the models considered here, since these tests rely on assumptions that are not fulfilled in our digital twin model.

The distinctive characteristics of our approach compel us to use alternative methods to establish

the statistical significance of our results regarding sensitivity, specificity, accuracy, and AUC. In this regard, due to the immediate interpretation of the AUC and its equivalence to the Mann-Whitney U test and the Wilcoxon test of ranks [108, 115], the AUC is considered the most robust metric for evaluating and comparing predictive accuracy. After selecting this metric, the statistical significance of our results can be obtained in two different ways: By using the predicted and the observed series, or by considering the ROC curves. In the first option, we compute the Handley and McNeil test [116]. The second alternative requires the implementation of a semiparametric bootstrapping procedure [117]. As depicted in Table 4, in both cases, the p-values are below 0.05 for all the comparisons. Therefore, we can conclude that the superior performance of our digital twin is statistically significant.

Table 4. Performance of the considered models: Statistical significance of AUC differences.

Scenario	**Predicted-observed series Handley-McNeil test**			**ROC curve Semiparametric bootstrapping**		
Market Index	*Model vs K-Means*	*Model vs SOM*	*Model vs EM-GMM*	*Model vs K-Means*	*Model vs SOM*	*Model vs EM-GMM*
Price Entropy	p-value=7.584e-05 [0.04807, 0.14242]	p-value=5.595e-05 [0.04940, 0.14299]	p-value=0.01859 [0.00979, 0.10733]	p-value¡2.2e-16 [0.30037, ∞]	p-value¡2.2e-16 [0.32393, ∞]	p-value¡2.2e-16 [0.17453, ∞]
VIX	p-value=8.911e-13 [0.23089, 0.40541]	p-value=5.34e-13 [0.23366, 0.40791]	p-value=1.236e-10 [0.20265, 0.38017]	p-value¡2.2e-16 [0.44969,∞]	p-value¡2.2e-16 [0.45578,∞]	p-value¡2.2e-16 [0.37151,∞]
BP	p-value=5.328e-08 [0.13369, 0.28429]	p-value=5.197e-08 [0.13386, 0.28445]	p-value=0.00048 [0.06001, 0.21343]	p-value¡2.2e-16 [0.22858,∞]	p-value¡2.2e-16 [0.22842,∞]	p-value¡2.2e-16 [0.15955,∞]
Global	p-value=2.508e-06 [0.06252, 0.15172]	p-value=1.743e-06 [0.06372, 0.15225]	p-value=0.00211 [0.02620, 0.11837]	p-value¡2.2e-16 [0.29832,∞]	p-value¡2.2e-16 [0.31991,∞]	p-value¡2.2e-16 [0.17908,∞]
Alternative hypothesis, Handley-McNeil test: AUCs are not equal. Alternative hypothesis, semiparametric bootstrapping method: Model AUC is higher. In brackets: 95% confidence interval for the difference in AUCs						

6. Conclusions

The analysis of uncertainty, understood as a situation that cannot be accurately predicted, is a complex question due to the multiple and intricate dimensions involved in its definition and mathematical consideration. Despite this difficulty, researchers have not renounced to its formal analysis, mainly because of the pervasive role that uncertainty and its correct anticipation plays in economic life. As a result, economics has devoted important efforts to design mathematical models able to identify, quantify and predict uncertainty and its economic consequences. Most of these economic models rely on obtaining a mathematical characterization of the uncertain behavior of the considered numerical variables, usually market prices, which are then structured according to the time, space or categorical criteria. These structured numerical data give rise to alternative measure functions (i.e., probability, credibility, chance, and uncertain measure functions), which are used to identify, quantify and forecast uncertainty. In this respect, although these measures built with structured numerical market data have shown some ability to anticipate future uncertainty, the evidence shows that their performance is very limited, thus questioning the use of only these market variables to predict future uncertainty.

To overcome this shortage, the use of unstructured data, particularly news and reports, has emerged as an alternative to quantify and predict stock market uncertainty. Indeed, models considering news and other textual forms are used to confirm the verification of the EMT and MST and to elaborate

predictions for asset prices, offering results that are promising although not totally satisfactory.

Our research falls within this line. More specifically, we propose a digital twin model that mirrors the information processing carried out by the market agents, who are, according to the EMT and MST, ultimately responsible for the behavior of the stock market prices, thus establishing a correspondence between the absence of a clear and certain interpretation of the available information by the market agents and the episodes of stock market uncertainty. To this end, we use novel AI approaches and techniques to link the news and the market uncertainty, and to identify, quantify and forecast stock market uncertainty. These new features are present at all the stages of the model, and look for an identification of the uncertainty encapsulated in news similar to that of our mind. To do so, we consider the presence of words of a particular interest, taking into account their frequency of appearance, the specific moments at which each world appears, its relationships with the other words, and, most importantly, whether or not all these textual informative aspects can be clearly interpreted. These tasks are accomplished by implementing AI techniques at different levels. First, for each word, the applied algorithms allow a statistical description of the abovementioned timeline and contextual patterns of its appearances to be obtained, thus constituting an important contribution of our proposal. Then, at a second level, a set of data mining techniques -namely SOM, hierarchical clustering, and gaussian mixture clustering- are sequentially applied to define and characterize regions of homogeneous knowledge/information, which is another interesting novelty. Finally, we propose the use of Shannon entropy to measure the uncertainty present in the news, in the sense of an absence of clear assignation of the considered informative elements to a specific environment of homogeneous knowledge.

The evaluation of our model in predicting stock market uncertainty is clearly successful. To show the robustness of our proposal, the model predictions are compared against those of the standard algorithms -SOM, k-means and EM- in four different scenarios, each representing a different conception of real uncertainty. The comparative analysis of these performances is carried out by applying the usual concepts of sensitivity, specificity, accuracy and AUC. For these measures of predictive quality and for all the considered scenarios, our proposed model obtains the best results by far, specially -but not only- when market uncertainty is understood as price volatility. Indeed, when the presence of real uncertainty is associated to both price volatility and changes in cycle and tendency, our model successfully predicts the observed episodes of market uncertainty, predictions which are much better than the standard models and show excellent values for specificity, accuracy and AUC.

In light of these results, our research contributes to the literature in several aspects. First, concerning financial economics, our model constitutes a benchmark to assess the fulfillment of the EMT and MST from a different perspective. These two central theories in economics have mainly been studied by testing the existence of arbitrage opportunities with series of stock prices (see the recent studies [118, 119]), but not by determining whether or not stock prices reflect the agents' rational processing of the available information. Since our digital twin model replicates the way textual information is interpreted by the human brain, its ability for predicting real episodes of market uncertainty can be envisaged as an alternative test of the EMT and MST, as well as a novel approach to continue exploring the bounded rationality research avenue [120].

As a second contribution, our research proves the capability and flexibility of the NLP approach and AI techniques to numerically characterize and interpret textual data. In this respect, even for such a complex question as market uncertainty, the joint and adequate combination of NLP, AI and

IT techniques and concepts has provided a satisfactory measure and prediction of this magnitude. From this perspective, our methods and techniques can also contribute to the field of computational neuroscience devoted to the design of language models and to the comprehension of how the human brain elaborates supra-word meanings, which is the subject of recent interesting studies (see [60, 61, 121]).

Moreover, our research shows that models built under the umbrella of NLP can be extended to determine the processes guiding the perception of uncertainty by the agents and its transmission to the stock markets, thus providing new alternatives to study and predict the appearance of contagions, bubbles and financial panics. As our study shows, these advances must come from the combination of quantitative and qualitative improvements in the text data sources, the rigourous discussion of the processes to be implemented, and the design of the proper testing experiments. For the particular field of financial uncertainty forecasting, and in light of the empirical results where our model constitutes a valuable proposal to identify, quantify and predict market uncertainty, it is important to consider alternative and/or additional information sources, as well as new stock market indexes. In this respect, given the increasing presence and prominence of *fake news* and misinformation campaigns, special attention must be paid to the consideration and valuation of these disturbing elements, which is a question already pointed out by recent studies (see for instance [122, 123]). Since these new input data can reproduce different and not contemplated features of uncertainty, the use of different statistics in the text vectorization phase should also be addressed. Concerning our measure of the underlying uncertainty through the Shannon entropy concept, the performance of the model can be further investigated by using different levels of entropy normalization.

Last but by no means least, from a practical standpoint and concerning the investment activity in the markets and the design of optimal portfolios, our research also makes an interesting contribution. Anticipating periods of financial uncertainty is crucial for fund managers, thereby helping them to design hedging tactics to mitigate losses or maximize gains. In this context, our digital twin could be integrated into algorithmic platforms that automatically manage financial portfolios, known as robo-advisors, thus providing them with reliable real-time predictions about periods of uncertainty and turbulence and therefore enabling the possibility of higher profits.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This article was largely completed while PJGD was visiting the Department of Economics of the European University Institute (EUI) from January 2023 to May 2023. The authors express their gratitude to the EUI for providing research facilities and access to internal scientific meetings. PJGD acknowledges financial support from: Spanish Agencia Estatal de Investigación-Ministerio de Ciencia e Innovación (research project PID2020-113554GB-I00/AEI/10.13039/501100011033); University of Valladolid Research Group on Dynamic Optimization, Mathematical Finance and Recursive Utility; and University of Valladolid-Bank of Santander 2023 Mobility Aids for Researchers. This research has made use, free of cost, of high performance computing resources provided by SCAYLE [95].

Professional assistance for manuscript preparation from Alan F. Hydns, B.A. Dip. TEFL, is highly appreciated.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. G. Cardano, *Liber de ludo aleae*, In: C. Sponius (ed.), Hieronymi Cardani Mediolanensis Opera Omnia, Lyons, 1663, 1564.
2. B. Pascal, P. Fermat, *Letters*, In: Pascal Fermat Correspondence, 1654. Available from: `http://www.york.ac.uk/depts/maths/histstat/pascal.pdf`.
3. J. Bernoulli, *The art of conjecturing, together with letter to a friend on sets in court tennis*, English translation by Edith Sylla, Baltimore: Johns Hopkins Univ Press, 2005, 1713. https://doi.org/10.1111/j.1600-0498.2008.00117.x
4. F. P. Ramsey, *Truth and probability*, In: The Foundations of Mathematics and other Logical Essays, ed. R. B. Braithwaite, London: Routledge & Kegan Paul Ltd, 1926. https://doi.org/10.1007/978-3-319-20451-2_3
5. L. J. Savage, *The foundations of statistics*, New York: John Wiley & Sons, 1954. https://doi.org/10.1002/nav.3800010316
6. J. M. Keynes, *A treatise on probability*, Macmillan & Co., 1921. https://doi.org/10.2307/2178916
7. F. H. Knight, *Risk, uncertainty and profit*, Chicago University Press, **31** (1921). https://doi.org/10.1017/CBO9780511817410.005
8. M. Kurz, M. Motolese, Endogenous uncertainty and market volatility, *Econ. Theory*, **17** (2001), 497–544. http://dx.doi.org/10.2139/ssrn.159608
9. M. B. Beck, Water quality modeling: A review of the analysis of uncertainty, *Water Resour. Res.*, **23** (1987), 1393–1442. https://doi.org/10.1029/WR023i008p01393
10. S. O. Funtowicz, J. R. Ravetz, *Uncertainty and quality in science for policy*, Springer Science & Business Media, 1990. http://dx.doi.org/10.1007/978-94-009-0621-1
11. M. B. A. van Asselt, J. Rotmans, Uncertainty in integrated assessment modelling, *Climatic Change*, **54** (2002), 75–105. https://doi.org/10.1023/A:1015783803445
12. E. F. Fama, The behavior of stock-market prices, *J. Bus.*, **38** (1965), 34–105. http://dx.doi.org/10.1086/294743
13. A. Alchian, Uncertainty, evolution and economic theory, *J. Polit. Econ.*, **58** (1950), 211–221. http://dx.doi.org/10.1086/256940
14. A. Sandroni, Do Markets favor agents able to make accurate predictions? *Econometrica*, **68** (2000), 1303–1341. http://dx.doi.org/10.1111/1468-0262.00163
15. A. Sandroni, Efficient markets and Bayes' rule, *Econ. Theory*, **26** (2005) 741–764. http://dx.doi.org/10.1007/s00199-004-0567-4

16. L. Blume, D. Easley, Evolution and market behavior, *J. Econ. Theory*, **58** (1992), 9–40. http://dx.doi.org/10.1016/0022-0531(92)90099-4

17. L. Blume, D. Easley, If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets, *Econometrica*, **74** (2006), 929–966. http://dx.doi.org/10.1111/j.1468-0262.2006.00691.x

18. O. San, The digital twin revolution, *Nat. Comput. Sci.*, **1** (2021), 307–308. https://doi.org/10.1038/s43588-021-00077-0

19. G. Caldarelli, E. Arcaute, M. Barthelemy, M. Batty, C. Gershenson, D. Helbing, et al., The role of complexity for digital twins of cities, *Nat. Comput. Sci.*, **3** (2023), 374–381. https://doi.org/10.1038/s43588-023-00431-4

20. H. M. Markowitz, Portfolio selection, *J. Financ.*, **7** (1952) 77–91. http://dx.doi.org/10.2307/2975974

21. Z. Y. Guo, Heavy-tailed distributions and risk management of equity market tail events, *J. Risk Control*, **4** (2017), 31–41. http://dx.doi.org/10.2139/ssrn.3013749

22. R. E. Lucas, Asset prices in an exchange economy, *Econometrica,* **46** (1978), 1429–1445. https://doi.org/10.2307/1913837

23. J. H. Cochrane, *Asset pricing*, Princeton University Press, 2005. https://doi.org/10.1016/j.jebo.2005.08.001

24. D. Ellsberg, Risk, ambiguity, and the savage axioms, *Quart. J. Econ.*, **75** (1961), 643–669. http://dx.doi.org/10.2307/1884324

25. H. R. Varian, *Differences of opinion in financial markets*, In: C. C. Stone, (eds) Financial Risk: Theory, Evidence and Implications, Springer, Dordrecht., 1989. https://doi.org/10.1007/978-94-009-2665-3_1

26. B. Liu, *Uncertainty theory*, In: Uncertainty Theory, Studies in Fuzziness and Soft Computing, Berlin: Springer, **154** (2007). https://doi.org/10.1007/978-3-540-73165-8_5

27. B. Liu, Fuzzy process, hybrid process and uncertain process, *J. Uncertain Syst.*, **2** (2008), 3–16.

28. B. Liu, Toward uncertain finance theory, *J. Uncertain. Anal. Appl.*, **1** (2013), 1–15. http://dx.doi.org/10.1186/2195-5468-1-1

29. M. Segoviano, C. A. Goodhart, *Banking stability measures*, International Monetary Fund, 2009. https://doi.org/10.5089/9781451871517.001

30. L. Liu, T. Zhang, Economic policy uncertainty and stock market volatility, *Financ. Res. Lett.*, **15** (2015), 99–105. https://doi.org/10.1016/j.frl.2015.08.009

31. H. Asgharian, C. Christiansen, A. J. Hou, The effect of uncertainty on stock market volatility and correlation, *J. Bank. Financ.*, **154** (2023), 106929. https://doi.org/10.1016/j.jbankfin.2023.106929

32. T. Simin, The poor predictive performance of asset pricing models, *J. Financ. Quant. Anal.*, **43** (2008), 355–380. http://dx.doi.org/10.1017/S0022109000003550

33. J. H. Boyd, J. Hu, R. Jagannathan, The stock market's reaction to unemployment news: Why bad news is usually good for stocks, *J. Financ.*, **60** (2005), 649–672. http://dx.doi.org/10.1111/j.1540-6261.2005.00742.x

34. R. P. Schumaker, H. Chen, *Textual analysis of stock market prediction using breaking financial news: The AZFin text system*, *ACM Trans. Inform. Syst.*, **27** (2009), 1–19. https://doi.org/10.1145/1462198.1462204

35. M. T. Suleman, Stock market reaction to good and bad political news, *Asian J. Financ. Account.*, **4** (2012), 299–312. https://doi.org/10.5296/ajfa.v4i1.1705

36. C. O. Cepoi, Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil, *Financ. Res. Lett.*, **36** (2020), 101658. https://doi.org/10.1016/j.frl.2020.101658

37. A. Caruso, Macroeconomic news and market reaction: Surprise indexes meet nowcasting, *Int. J. Forecasting*, **35** (2019), 1725–1734. https://doi.org/10.1016/j.ijforecast.2018.12.005

38. E. F. Fama, Efficient capital markets: II, *J. Financ.*, **46** (1991), 1575–1617. https://doi.org/10.2307/2328565

39. J. D. Thomas, K. Sycara, *Integrating genetic algorithms and text learning for financial prediction*, In: Proceedings of GECCO '00 Workshop on Data Mining with Evolutionary Algorithms, 2000, 72–75.

40. P. C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, *J. Financ. Forthcoming*, **62** (2007), 1139–1168. https://dx.doi.org/10.2139/ssrn.685145

41. S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, N. A. Smith, *Predicting risk from financial reports with regression*, In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09, 2009, 272–280. http://dx.doi.org/10.3115/1620754.1620794

42. J. L. Rogers, D. J. Skinner, A. Van Buskirk, Earnings guidance and market uncertainty, *J. Account. Econ.*, **48** (2009), 90–109. https://doi.org/10.1016/j.jacceco.2009.07.001

43. J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, X. Deng, *Exploiting topic based twitter sentiment for stock prediction*, In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, **2** (2013), 24–29.

44. X. Ding, Y. Zhang, T. Liu, J. Duan, *Using structured events to predict stock price movement: An empirical investigation*, In: Proceedings of the 2014 conference on empirical methods in natural language processing, 2014, 1415–1425. http://dx.doi.org/10.3115/v1/D14-1148

45. W. Y. Wang, Z. Hua, *A semiparametric gaussian copula regression model for predicting financial risks from earnings calls*, In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, **1** (2014), 1155–1165. http://dx.doi.org/10.3115/v1/P14-1109

46. R. Luss, A. d'Aspremont, Predicting abnormal returns from news using text classification, *Quant. Financ.*, **15** (2015), 999–1012. https://doi.org/10.1080/14697688.2012.672762

47. P. K. Narayan, D. Bannigidadmath, Does financial news predict stock returns? New evidence from Islamic and non-Islamic stocks, *Pac.-Basin Financ. J.*, **42** (2017) 24–45. https://doi.org/10.1016/j.pacfin.2015.12.009

48. F. Larkin, C. Ryan, *Good news: Using news feeds with genetic programming to predict stock prices*, In: European Conference on Genetic Programming, **4971** (2008), 49–60. https://doi.org/10.1007/978-3-540-78671-9_5

49. Y. Kim, S. R. Jeong, I. Ghani, Text opinion mining to analyze news for stock market prediction, *Int. J. Adv. Soft Comput. Appl.*, **6** (2014), 2074–8523.
50. A. E. Khedr, S. E. Salama, N. Yaseen, Predicting stock market behavior using data mining technique and news sentiment analysis, *Int. J. Adv. Soft Comput. Appl.*, **9** (2017), 22–30. https://doi.org/10.5815/ijisa.2017.07.03
51. X. Zhou, H. Zhou, H. Long, Forecasting the equity premium: Do deep neural network models work? *Mod. Financ.*, **1** (2023), 1–11. https://doi.org/10.61351/mf.v1i1.2
52. X. Dong, Y. Li, D. E. Rapach, G. Zhou, Anomalies and the expected market return, *J. Financ.*, **77** (2022), 639–681. https://doi.org/10.1111/jofi.13099
53. N. Cakici, C. Fieberg, D. Metko, A. Zaremba, Do anomalies really predict market returns? New data and new evidence, *Rev. Financ.*, 2023, rfad025. https://doi.org/10.1093/rof/rfad025
54. W. Shengli, Is human digital twin possible? *Comput. Method. Prog. Biomed. Update*, **1** (2021), 100014. https://doi.org/10.1016/j.cmpbup.2021.100014
55. M. Singh, E. Fuenmayor, E. P. Hinchy, Y. Qiao, N. Murray, D. Devine, Digital twin: Origin to future, *Appl. Syst. Inno.*, **4** (2021), 36. https://doi.org/10.3390/asi4020036
56. H. D. Critchley, C. J. Mathias, R. J. Dolan, Neural activity in the human brain relating to uncertainty and arousal during anticipation, *Neuron*, **29** (2001), 537–545. https://doi.org/10.1016/s0896-6273(01)00225-2
57. H. A. Simon, Rational decision-making in business organizations, *Am. Econ. Rev.*, **69** (1979), 493–513.
58. R. M. Hogarth, N. Karelaia, Regions of rationality: Maps for bounded agents, *Decis. Anal.*, **3** (2006), 124–144. http://dx.doi.org/10.1287/deca.1060.0063
59. Y. Wang, N. Zhang, Uncertainty analysis of knowledge reductions in rough sets, *The Scientific World J.*, **2014** (2014), 576409. https://doi.org/10.1155/2014/576409
60. K. Erk, Understanding the combined meaning of words, *Nat. Comput. Sci.*, **2** (2022), 701–702. https://doi.org/10.1038/s43588-022-00338-6
61. M. Toneva, T. M. Mitchell, L. Wehbe, Combining computational controls with natural text reveals aspects of meaning composition, *Nat. Comput. Sci.*, **2** (2022), 745–757. https://doi.org/10.1038/s43588-022-00354-6
62. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
63. G. E. Hinton, *Distributed representations*, Carnegie Mellon University, 1984.
64. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, *Nature*, **323** (1986), 533–536. http://dx.doi.org/10.1038/323533a0
65. J. L. Elman, Finding structure in time, *Cognitive Sci.*, **14** (1990), 179–211. http://dx.doi.org/10.1207/s15516709cog1402_1
66. Y. Bengio, H. Schwenk, F. Morin, J. L. Gauvain, *Neural probabilistic language models*, In: Innovations in Machine Learning: Theory and Applications, 2006, 137–186. https://doi.org/10.1007/3-540-33486-6_6

67. R. Collobert, J. Weston, *A unified architecture for natural language processing: Deep neural networks with multitask learning*, In: Proceedings of the 25th international conference on Machine learning - ICML '08, 2008, 160–167. https://doi.org/10.1145/1390156.1390177
68. A. Mnih, G. E. Hinton, A scalable hierarchical distributed language model, *Adv. Neural Inform. Process. Syst.*, **21** (2008), 1081–1088. https://dl.acm.org/doi/10.5555/2981780.2981915
69. T. Mikolov, J. Kopecky, L. Burget, O. Glembek, J. Cernocky, *Neural network based language models for highly inflective languages*, In: 2009 IEEE international conference on acoustics, speech and signal processing, 2009, 4725–4728. https://doi.org/10.1109/ICASSP.2009.4960686
70. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, *J. Mach. Learn. Res.*, **12** (2011), 2493–2537. https://dl.acm.org/doi/10.5555/1953048.2078186
71. E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng, *Improving word representations via global context and multiple word prototypes*, In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, **1** (2012), 873–882.
72. Y. Zhang, R. Jin, Z. H. Zhou, Understanding bag-of-words model: a statistical framework, *Int. J. Mach. Learn. Cyb.*, **1** (2010), 43–52. http://dx.doi.org/10.1007/s13042-010-0001-0
73. *The lifespan of news stories, How the news enters (and exits) the public consciousness*, Schema Design and Google Trends, 2019. Available from: `https://newslifespan.com/`.
74. N. Bloom, Fluctuations in uncertainty, *J. Econ. Perspect.*, **28** (2014), 153–176. http://dx.doi.org/10.1257/jep.28.2.153
75. S. R. Baker, S. J. Davis, J. A. Levy, State-level economic policy uncertainty, *J. Monetary Econ.*, **132** (2022), 81–99. http://dx.doi.org/10.1016/j.jmoneco.2022.08.004
76. S. Newcomb, A generalized theory of the combination of observations so as to obtain the best result, *Am. J. Math.*, 1886, 343–366. http://dx.doi.org/10.2307/2369392
77. D. Böhning, E. Dietz, P. Schlattmann, Recent developments in computer-assisted analysis of mixtures, *Biometrics*, **54** (1998), 525–536. http://dx.doi.org/10.2307/3109760
78. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. Ser. B*, **39** (1977), 1–22. http://dx.doi.org/10.1111/j.2517-6161.1977.tb01600.x
79. T. Heskes, Self-organizing maps, vector quantization, and mixture modeling, *IEEE T. Neur. Net.*, **12** (2001), 1299–1305. http://dx.doi.org/10.1109/72.963766
80. A. Gepperth, B. Pfülb, *A rigorous link between self-organizing maps and gaussian mixture models*, In: Artificial Neural Networks and Machine Learning-ICANN 2020, Springer, Cham, 2020, 863–872. http://dx.doi.org/10.1007/978-3-030-61616-8_69
81. D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, et al., The subspace Gaussian mixture model—A structured model for speech recognition, *Comput. Speech Lang.*, **25** (2011), 404–439. http://dx.doi.org/10.1016/j.csl.2010.06.003
82. J. Yin, J. Wang, *A dirichlet multinomial mixture model-based approach for short text clustering*, In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, 2014, 233–242. http://dx.doi.org/10.1145/2623330.2623715

83. F. Najar, S. Bourouis, N. Bouguila, S. Belghith, *A comparison between different Gaussian-based mixture models*, In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), 2017, 704–708. http://dx.doi.org/10.1109/AICCSA.2017.108
84. G. Bordogna, G. Pasi, Soft clustering for information retrieval applications, *WIRES Data Min. Knowl.*, **1** (2011), 138–146. http://dx.doi.org/10.1002/widm.3
85. N. F. G. Martin, J. W. England, R. Baierlein, Mathematical theory of entropy, *Phys. Today*, **36** (1983), 66–67. http://dx.doi.org/10.1063/1.2915804
86. S. R. Bentes, R. Menezes, Entropy: A new measure of stock market volatility? *J. Phys. Conf. Ser.*, **394** (2012), 012033. http://dx.doi.org/10.1088/1742-6596/394/1/012033
87. K. Ahn, D. Lee, S. Sohn, B. Yang, Stock market uncertainty and economic fundamentals: An entropy-based approach, *Quant. Financ.*, **19** (2019), 1151–1163. http://dx.doi.org/10.1080/14697688.2019.1579922
88. T. Kohonen, The self-organizing map, *Neurocomputing*, **21** (1998), 1–6. http://dx.doi.org/10.1016/S0925-2312(98)00030-7
89. M. Y. Kiang, Extending the Kohonen self-organizing map networks for clustering analysis, *Comput. Stat. Data Anal.*, **38** (2001), 161–180. http://dx.doi.org/10.1016/S0167-9473(01)00040-8
90. T. Kohonen, *Self-organization and associative memory*, Springer Science & Business Media, **8** (2012). https://doi.org/10.1007/978-3-642-88163-3
91. B. Szmrecsanyi, *Grammatical variation in British English dialects: A study in corpus-based dialectometry*, Cambridge University Press, 2012. http://dx.doi.org/10.1017/CBO9780511763380
92. *Dow Jones & CO WSJ.COM audience profile*, comScore Media Metrix Q1, 2021. Available from: `https://images.dowjones.com/wp-content/uploads/sites/183/2018/05/09164150/WSJ.com-Audience-Profile.pdf`.
93. *VIX volatility suite*, Cboe Global Markets, Inc., 2021. Available from: `https://www.cboe.com/tradable_products/vix/`.
94. A. Elder, *Trading for a living: Psychology, trading tactics, money management*, John Wiley & Sons, **31** (1993).
95. *SCAYLE Supercomputación Castilla y León*, 2021. Available from: `https://www.scayle.es`.
96. O. A. M. Salem, F. Liu, A. S. Sherif, W. Zhang, X. Chen, Feature selection based on fuzzy joint mutual information maximization, *Math. Biosci. Eng.*, **18** (2020), 305–327. http://dx.doi.org/10.3934/mbe.2021016
97. P. V. Balakrishnan, M. C. Cooper, V. S. Jacob, P. A. Lewis, A study of the classification capabilities of neural networks using unsupervised learning: A comparison with K-means clustering, *Psychometrika*, **59** (1994), 509–525. https://doi.org/10.1007/BF02294390
98. A. Flexer, Limitations of self-organizing maps for vector quantization and multidimensional scaling, *Adv. Neur. Inform. Process. Syst.*, **9** (1996), 445–451.
99. U. A. Kumar, Y. Dhamija, *Comparative analysis of SOM neural network with K-means clustering algorithm*, In: 2010 IEEE International Conference on Management of Innovation & Technology, 2010, 55–59. http://dx.doi.org/10.1109/ICMIT.2010.5492838

100.J. Han, M. Kamber, J. Pei, *Data mining: Concepts and techniques*, 3 Eds., Morgan Kauffman, 2012. https://doi.org/10.1016/C2009-0-61819-5

101.H. M. Hodges, Arbitrage bounds of the implied volatility strike and term structures of European-style options, *J. Deriv.*, **3** (1996), 23–35. http://dx.doi.org/10.3905/jod.1996.407950

102.A. M. Malz, A simple and reliable way to compute option-based risk-neutral distributions, *FRB New York Staff Rep.*, **677** (2014). http://dx.doi.org/10.2139/ssrn.2449692

103.B. Judge, *26 May 1896: Charles Dow launches the Dow Jones industrial average*, 2015. Available from: `https://moneyweek.com/392888/26-may-1896-charles-dow-launches-the-dow-jones-industrial-average/`.

104.A. C. MacKinlay, Event studies in economics and finance, *J. Econ. Lit.*, **35** (1997), 13–39.

105.Z. Önder, C. Şimga-Mugan, How do political and economic news affect emerging markets? Evidence from Argentina and Turkey, *Emerg. Mark. Financ. Tr.*, **42** (2006), 50–77. http://dx.doi.org/10.2753/REE1540-496X420403

106.N. Aktas, E. de Bodt, J. G. Cousin, Event studies with a contaminated estimation period, *J. Corp. Financ.*, **13** (2007), 129–145. http://dx.doi.org/10.1016/j.jcorpfin.2006.09.001

107.O. Arslan, W. Xing, F. A. Inan, H. Du, Understanding topic duration in Twitter learning communities using data mining, *J. Comput. Assist. Learn.*, **38** (2022), 513–525. http://dx.doi.org/10.1111/jcal.12633

108.T. Fawcett, An introduction to ROC analysis, *Pattern Recogn. Lett.*, **27** (2006), 861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010

109.D. W. Hosmer, S. Lemeshow, *Applied logistic regression*, 2 Eds., New York: John Wiley and Sons, 2000, 160–164. http://dx.doi.org/10.1002/0471722146

110.T. Fawcett, ROC graphs: Notes and practical considerations for researchers, *Mach. Learn.*, **31** (2004), 1–38.

111.F. Melo, Area under the ROC curve, *Encyclopedia Syst. Biol.*, **2013** (2013). http://dx.doi.org/10.1007/978-1-4419-9863-7_209

112.J. Cragg, R. Uhler, The demand for automobiles, *Can. J. Econ.*, **3** (1970), 386–406. http://dx.doi.org/10.2307/133656

113.G. Maddala, *Limited dependent and qualitative variables in econometrics*, New York: Cambridge University Press, 1983. http://dx.doi.org/10.1017/CBO9780511810176

114.D. R. Cox, N. Wermuth, A comment on the coefficient of determination for binary responses, *Am. Stat.*, **46** (1992), 1–4. http://dx.doi.org/10.2307/2684400

115.P. Flach, J. Hernández-Orallo, C. Ferri, *A coherent interpretation of AUC as a measure of aggregated classification performance*, In: Proceedings of the 28th International Conference on Machine Learning, 2011.

116.J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143** (1982), 29–36. http://dx.doi.org/10.1148/radiology.143.1.7063747

117.B. Efron, *The jackknife, the bootstrap, and other resampling plans*, In: Society of Industrial and Applied Mathematics CBMS-NSF Monographs 38, 1982. http://dx.doi.org/10.1137/1.9781611970319

118.I. A. Boboc, M. C. Dinică, An Algorithm for testing the efficient market hypothesis, *PloS One*, **8** (2013), e78177. https://doi.org/10.1371/journal.pone.0078177

119.M. A. Sánchez-Granero, K. A. Balladares, J. P. Ramos-Requena, J. E. Trinidad-Segovia, Testing the efficient market hypothesis in Latin American stock markets, *Physica A*, **540** (2020), 123082. https://doi.org/10.1016/j.physa.2019.123082

120.E. M. Sent, Rationality and bounded rationality: You can't have one without the other, *Eur. J.Hist. Econ. Thou.*, **25** (2018), 1370–1386. http://dx.doi.org/10.1080/09672567.2018.1523206

121.M. Hahn, R. Futrell, R. Levy, E. Gibson, A resource-rational model of human processing of recursive linguistic structure, *P. Natl. Acad. Sci.*, **119** (2022), e2122602119. http://dx.doi.org/10.1073/pnas.2122602119

122.M. Szczepański, M. Pawlicki, R. Kozik, M. Choraś, New explainability method for BERT-based model in fake news detection, *Sci. Rep.*, **11** (2021), 23705. http://dx.doi.org/10.1038/s41598-021-03100-6

123.G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, D. G. Rand, Shifting attention to accuracy can reduce misinformation online, *Nature*, **592** (2021) 590–595. http://dx.doi.org/10.1038/s41586-021-03344-2

AIMS Press