DOI: 10.1111/eufm.12326

ORIGINAL ARTICLE

EUROPEAN FINANCIAL MANAGEMENT WILEY

Machine learning in finance: A topic modeling approach

Saqib Aziz[1] | **Michael Dowling**[2] | **Helmi Hammami**[1] | **Anke Piepenbrink**[3]

[1]Department of Finance & Accounting, Rennes School of Business, Rennes, France

[2]Financial and Operational Performance Group, DCU Business School, Dublin City University, Dublin, Ireland

[3]Department of Strategy & Innovation, Rennes School of Business, Rennes, France

Correspondence
Saqib Aziz, Rennes School of Business. 2 rue Robert d'Arbrissel, 35065 Rennes, France.
Email: saqib.aziz@rennes-sb.com

Abstract

We identify the core topics of research applying machine learning to finance. We use a probabilistic topic modeling approach to make sense of this diverse body of research spanning across multiple disciplines. Through a latent Dirichlet allocation topic modeling technique, we extract 15 coherent research topics that are the focus of 5942 academic studies from 1990 to 2020. We find that these topics can be grouped into four categories: Price-forecasting techniques, financial markets analysis, risk forecasting and financial perspectives. We first describe and structure these topics and then further show how the topic focus has evolved over the last three decades. A notable trend we find is the emergence of text-based machine learning, for example, for sentiment analysis, in recent years. Our study thus provides a structured topography for finance researchers seeking to integrate machine learning research approaches in their exploration of finance phenomena. We also showcase the benefits to finance researchers of the method of probabilistic

We thank John A. Doukas, the editor, and an anonymous referee of European Financial Management as the study has enormously benefited from their comments. We also thank Muhammad Farooq Ahmad and participants of IFABS 2019, Angers France for their valuable comments. Saqib Aziz and Michael Dowling acknowledge financial assistance from the B<>COM project: Prospect 2030. The views expressed in this article are those of the authors and all errors are our own.

modeling of topics for deep comprehension of a body of literature.

KEYWORDS
finance, latent dirichlet allocation, machine learning, textual analysis, topic modeling

JEL CLASSIFICATION
G00, C45

> How will Big Data and Machine Learning change the investment landscape? We think the change will be profound. As more investors adopt alternative data sets, the market will start reacting faster and will increasingly anticipate traditional or 'old' data sources (e.g., quarterly corporate earnings, low-frequency macroeconomic data, etc.). This gives an edge to quant managers and those willing to adopt and learn about new data sets and methods. Eventually, 'old' data sets will lose most predictive value and new data sets that capture 'Big Data' will increasingly become standardized.—JP Morgan, 2017.

1 | INTRODUCTION

The techniques of machine learning (ML) and artificial intelligence (AI) offer significant benefits to financial decision makers in terms of new approaches for modeling and forecasting from financial data. This has been recognized by the finance industry, with over three-quarters of finance firms expecting AI technologies to be of core strategic importance for their operations by 2022.[1] Major current finance AI applications are in algorithmic trading, risk management and process automation. Research on these topics among finance researchers is also growing rapidly. However, research in the area crosses many domains (computer science, finance, economics, decision making, engineering, to name a few) and so, understanding the scope and prior work of this sprawling area is difficult. In this paper, we, therefore, comprehensively structure this body of research for the benefit of researchers seeking to understand the techniques and areas of interest of ML and AI applied to finance. To do this we apply an analytic technique for structuring the research that can handle the spread across multiple research domains.

For common understanding, we first briefly give our working definitions of AI techniques. AI is an umbrella term for a range of techniques involving intelligence demonstrated by machines (Cockburn et al., 2018), with this intelligence normally focused on prediction. In a finance context, due to its numerical focus, the most relevant AI methods have been ML: Predictive algorithms and models involving statistical learning from data, and more recently, *deep learning* (DL): An approach that allows more abstracted learning from unknown

[1]World Economic Forum 'Transforming' Paradigms: A Global AI in Financial Services Survey' 2020.

relationships within the input data (Hastie et al., 2009). DL is an approach that has evolved from an earlier focus on artificial neural networks, which attempt to mimic the learning models of biological neural networks, such as the human brain. Of additional relevance in the toolkit of AI is *natural language processing* (NLP), centred on the understanding and analysis of textual data. NLP offers the potential to integrate the large body of textual data in learning and prediction and overlaps with ML to the extent that ML techniques can be applied to NLP data. The technique applied in this study, topic modeling, is an ML application of NLP, thus adding an element of symbiosis to our study with the use of an ML technique to understand research on ML in finance. We illustrate the basic relationship between AI, ML, DL and NLP in Figure 1.

The potential for ML in financial decision-making was, to our best understanding, first explored from a research perspective in Hawley et al. (1990) with a focus on neural networks as an aide to financial decision-making. Echoing its future benefit to banking, a number of early studies also appeared in the *Journal of Banking & Finance* in the 1990s which explored the potential for ML to improve lending decisions and credit risk management. Altman et al. (1994) applied neural networks to classify Italian firms based on likelihood of financial distress, whereas Varetto (1998) built on this study by applying genetic learning algorithms to the same topic. More recent research in finance journals has kept the focus on prediction but moved towards DL techniques and other advanced ML techniques. These recent applications include: The understanding determinants of firm risks (Cheng & Cirillo, 2018; Kamiya et al., 2019; Sigrist & Hirnschall, 2019), learning optimal option hedging rates (Nian et al., 2018), modeling investor sentiment (Renault, 2017; Sprenger et al., 2014), prediction of stock returns (F. Wang et al., 2020) and the detection of stock price evolution based on order books (Sirignano, 2019).

As finance journals have taken steps to acknowledge the potential of the new techniques of ML in finance, other disciplines have also been making strident efforts to apply ML approaches to financial data. In part, this is due to the attractiveness of the comprehensive, structured and easily accessible, data available in finance. Indeed, research on ML and finance outside of finance journals exceeds by a considerable multiple ML and finance research published in finance journals. Some recent examples of this outside-finance corpus of ML in finance

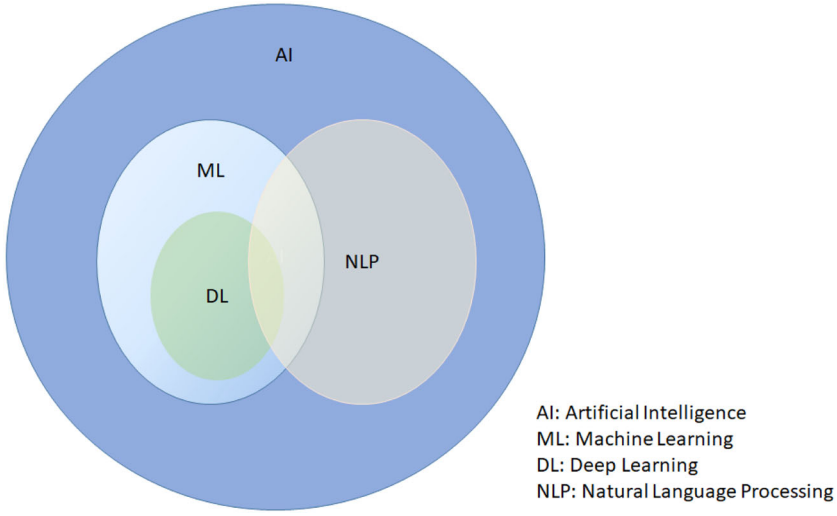

FIGURE 1 Interaction of various learning techniques of artificial intelligence

research include the application of image recognition techniques for the prediction of stock technical patterns (Sezer & Ozbayoglu, 2018), forecasting stock prices based on ensemble models of online search and sentiment data (Weng et al., 2018) and bank risk modeling (Cerchiello et al., 2017). These examples show both overlap but also distinct research approaches to mainstream financial research approaches to the modeling and application of ML in finance.

This multidisciplinary development of ML and finance research presents challenges to researchers seeking to understand the corpus of research in the area. The literature spread also risks duplication of research and incomplete inputs into idea development. Further, as the body of research is substantial—we identify 5942 studies (primarily from the last decade) in ML and finance—it is beyond reasonable human capabilities to comprehensively understand the directions and topics of this study. A standard literature review, or systematic literature review, requires a stage of human reading of individual documents to understand the core ideas in a document (Denyer & Tranfield, 2009). This process just cannot realistically be adopted for a review involving nearly 6000 individual studies. To address this problem, we adopt a data science approach of topic modeling. Topic modeling enables probabilistic machine-based learning of the core topics of ML and finance across all its research domains. The topic modeling approach also allows following the evolution of these topics over time. This provides both a comprehensive and trend-based understanding of the structure of research in this area.

We apply a latent Dirichlet allocation (LDA) approach (Blei et al., 2003; Blei, 2012) to topic modeling. LDA works off the assumption that each document is a collection of initially unknown topics and the words chosen by authors reflect and, therefore, can be used to learn the topics of a document. The probabilistic inference approach of LDA can be used to identify word clusters and, therefore, the topics from the collection of words in a document and at the same time to identify the most relevant shared topics across a corpus of documents. A core advantage of the technique, compared to other approaches to surveying a literature, is that no prior knowledge of topics is needed. The technique is also noted for its ability to arrive at quite distinct and precise topics (Chang et al., 2009).

Topic modeling in other business disciplines has recently appeared in the literature, including on how emerging markets are researched (Piepenbrink & Nurmammadov, 2015) and on the structure of approaches to organization research methods (Piepenbrink & Gaur, 2017). It has also been applied in practical business contexts such as determining patent topics (Kaplan & Vakili, 2015) and understanding online brand discussions (Tirunillai & Tellis, 2014). There has been a limited amount of topic modeling applications of closer relevance to finance. These include Moro et al. (2015) who grouped a small number of articles into topics on the application of business intelligence to the banking sector and Dyer et al. (2017) who determine 150 topics in US 10-K annual reports. Most relevant is Huang et al. (2018) who use topic modeling to draw connections between the contents of company calls with analysts and the subsequent content of analyst reports.

Our study makes a number of contributions. It is the first comprehensive structuring of the ML and finance research literature. There have been some more limited prior reviews of this literature, including Heaton et al. (2016) who structure the DL and finance prior research and Wong and Selvi (1998) who provide an early review of neural network applications in finance. There have also been recent reviews of ML applied to economics (Athey, 2018; Mullainathan & Spiess, 2017). However, perhaps because of the scale and spread of the literature, the full literature on ML in finance as researched across disciplines has not been previously structured.

Our study is also the first application of topic modeling to structuring the finance research literature. This is important as the technique offers particular benefits to financial research

comprehension. Finance research tends to be spread not just among finance journals but also, because of the transferable quantitative training of finance researchers and the attractiveness of ordered financial data as a data source, across a range of other disciplines. This makes it difficult to arrive at comprehensive understanding of the full body of research in an area of finance. We suggest that topic modeling is an important means of addressing this spread problem.

To finance researchers we also contribute in a broader sense as the application of ML to finance is of clear practical importance, but is currently (as our research review shows) primarily the research domain of computer scientists. This suggest that pertinent financial insights to the research might be being missed. By providing a structure to the topics of this study field and showing their development, we allow finance researchers to become familiar with the areas' structure as well as the approaches to research within each topic.

The paper proceeds in the following structure. A description of the topic modeling technique is provided in Section 2. The corpus of research papers selected and the preprocessing conducted to allow topic determination is covered in Section 3. Section 4 structures and analyses the identified topics and Section 5 concludes.

2 | TOPIC MODELING THROUGH LDA

Topic modeling is a process for discovering the 'hidden' or latent topics in a corpus of documents. By 'hidden' we mean that the topics are unknown at the outset and must be inferred from how words cluster in a document and are shared across the corpus. Neither the content of the topics, nor the number of topics that well-describe a corpus are known, so both must be deduced. The former through probabilistic modeling followed by expert labeling of topics and the latter through testing models with different numbers of topics and comparing the output.

There are a range of feasible approaches for implementing topic modeling, including Latent Semantic Indexing, Probabilistic Latent Semantic Analysis and LDA. The most widely accepted current approach is that of LDA developed in Blei et al. (2003),[2] due to the advantages of the incorporated Dirichlet distribution in assigning documents and terms to topics compared to older approaches. The practical advantage of LDA is that it allows individual documents to be a mixture of topics and terms are allowed to belong to more than one topic. This echoes the reality of how corpora of documents are structured, as well as recognizing that topics within an area can partially overlap.

A summary of the LDA approach to topic modeling is that it assumes a fixed number of hidden topics in a corpus of documents, uses word cooccurrence to determine what these topics are and then makes a probabilistic determination of the presence of these topics in a document. The perspective is that writers will approach writing a document with a collection of topics in mind and the words chosen by the writer will reflect this topic mixture. Thus, in an echo of an example given in the introduction, an article applying neural network approaches to credit risk modeling in banks might be written with a topic mixture of 50% credit risk modeling, 30% neural network methodologies and 20% banking context. The choice of words in a document

[2]ALthough we use the core method in this study, it is worth noting that there have been many useful variants of LDA proposed, including hierarchical topic modeling (Blei et al., 2004) and structural topic modeling (Roberts et al., 2014). These newer variants address various issues with core LDA such as identifying linkages between topics and finding global maxima. In our case, we choose the core LDA approach given the novelty to finance of topic modeling and leave the more advanced methods for future researchers.

will then reflect this mixture of topics and their relative emphasis. The key task for the topic modeling researcher is, therefore, to work backwards from the observed words to uncover the latent topics.

We now describe the LDA process in detail, starting with preprocessing of the data, moving to the implementation of the LDA and interpretation of the output. We summarize the steps of this process in Table 1.

A necessary first step in topic modeling is processing the corpus of documents. We describe this approach now, before proceeding to present the process for topic generation. Documents are initially processed by tokenizing each document into a collection of their individual words where order is unimportant (known as a 'bag of words' approach). We remove string items that are not words when tokenizing, such as special characters, grammatical characters and numbers. Common 'stop words' that offer no topic context (words such as 'and', 'of', 'the') are removed. Remaining words in a document are stemmed to reduce the unique word count further and accurately gauge unique term usage. That is, suffixes are removed to create

TABLE 1 Summary of the latent Dirichlet allocation (LDA) process

This table reports the salient steps of LDA process. Terminologies are defined in Section 2.

Step 1: Preprocess documents
a: Remove punctuation marks, special characters, numbers.
b: Convert each document into a list of words ('bag-of-words').
c: Remove stop words (common usage words that add minimal context).
d: Stem or lemmatize words to reduce to a common core.
e: Create a document-term matrix (DTM) counting term occurrence per document.
f: Make choice on removing low-frequency and high-frequency terms across documents (TF-IDF may be used for this step).
Step 2: Run LDA tests
a: Use an LDA package such as R *topicmodels* or Python *gensim*.
b: Specify number of topics across the documents (k). A range of topic numbers might be specified if unsure of the appropriate number of topics (common).
c: Decide whether to adjust α and β default Dirichlet priors.
d: Specify random starting state if wish to have a repeatable topic model. A range of random starting states might be specified to avoid local maxima.
Step 3: Interpret LDA output
a: If a range of possible topic models are run (e.g. testing a different number of topics) then choose between the topic models based on inspection of terms and matched documents to topics and by comparison of Krippendorf's α across the models as a measure of topic coherency.
b: At least two subject expert raters should independently label the topics within the selected topic model, with discussion of topic labels where raters disagree.
c: Elaborate on topic definition, beyond the label, by reference to top terms and top matched documents to each topic.

common stem terms, for example, both *finance* and *finances* might be reduced to the common *financ* stem.

A TF-IDF (Term Frequency—Inverse Document Frequency) assessment is now made of the relative importance of the remaining words in the corpus (Salton & Buckley, 1988). This process involves first calculating the percentage of occurrences of a term in a document compared to all terms in that document. This is then multiplied by the log of all documents in the corpus divided by the number of documents in a corpus that contain the term. Higher TF-IDF terms are relatively more important in the corpus and lower TF-IDF terms are less important. Very low TF-IDF terms tend to now be removed due to being too uncommonly present in the corpus to be able to describe a topic. Similarly, but for the opposite reason, very high TF-IDF terms tend to be removed due to being too widely used to be able to describe individual topics. Thus, for example, the term *equity* might occur in most investment papers and, thus, would not be a useful term for understanding subtopics within that area. In a research context this will also normally involve removing research structural words such as *results*, *hypothesis* and *analysis*, which, though not common in general word usage, are very common in a body of academic research studies. There is quite an element of researcher choice in determining high and low TF-IDF terms and no strict rule as each collection of documents can follow different word distribution structures. A common approach, hard-coded into LDA packages available for R or Python, is to remove terms that don't appear in at least 5, 10, 20 documents and, based on visual inspection, to remove the top 5%, 10% and 15% most frequent terms.

The final data set after the TF-IDF stage, is the construction of a document-term matrix (DTM). The DTM is a matrix structured as rows representing each document and columns representing each term, with values being frequency of occurrence of a term in a document. Thus, a corpus of 5000 documents with 10,000 unique terms across all documents remaining after cleaning will result in a $5000 \times 10{,}000$ dimension DTM. Topic modeling involves reducing the dimensions of this matrix to end up with the same number of rows/documents but a restricted number of columns which now represent the topics. A 20-topic LDA on the above example will, thus, result in a 5000×20 matrix with values being the probabilistic weighting of each topic within each of the 5000 documents. In turn the 20 topics will be a probabilistic weighting of each of the 10,000 original terms within each topic. This weighting allows two rankings to emerge. The first is the ranking of terms of importance to a topic. We commonly use the 10 most important terms to describe a topic, although all terms have some weighting in each topic no matter how small. The second ranking is that we can also rank documents in the DTM to determine the documents that most purely match a topic. This is useful for identifying exemplar documents for a topic and for structuring documents within a corpus.

The process for LDA is well described in Blei (2012) and in more technical detail in Blei et al. (2003) and Boyd-Graber et al. (2017), so rather than repeating the technical detail, we describe the intuition behind the process here using the plate diagram in Figure 2, as is common convention in describing relationships within algorithms in data science. We start our analysis with just the shaded circle, w, available which are specific words from each document in the corpus. All other variables (in circles) must be constructed or uncovered. N is the collection of all words w in a corpus. There are D documents d in the corpus and T topics t. A variable being in the N, D, or T plate indicates what form of the data that variable refers to.

α and β are the Dirichlet priors and are the key external input into the model to determine the generative process of the LDA. α is the parameter for per-document topic distributions and β is the parameter for per-topic word distribution. A Dirichlet distribution is a distribution that

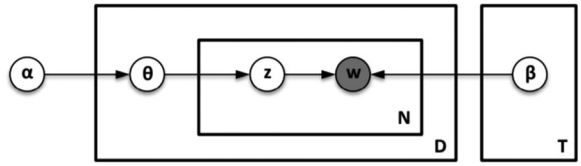

FIGURE 2 Latent Dirichlet allocation plate diagram. *Source*: Kaplan and Vakili (2015)

can be used to produce probability vectors[3] that allow in the LDA an assumption to be made about how topics are distributed across documents and words. A high α assumption indicates that each document in the corpus is likely to contain a distribution of most of the topics in the corpus and a low α indicates each document will contain only a few of the topics in the corpus. Similarly, a high β indicates that each topic is likely to contain a distribution of most of the words, whereas a low β indicates that each topic contains just a few of the words. The Dirichlet distribution, thus, gives the prior distributions across which we can approach latent topic discovery. The subsequent step in the model involve Bayesian updating to these priors based on actual word distribution across topics and documents.

Referring to Figure 2 again, α informs the initial θ which is the proportion of topics per document. The initial proportions are semirandomly allocated according to the α Dirichlet prior. Following the plate diagram arrows, θ then informs z, which is the actual topic assignment of words in a document. Separately β feeds into topic distribution of the words, which are also semirandomly allocated initially according to the Dirichlet prior. Inferring from the model, which in its initial form is intractable, is by means of a choice of inference algorithms. The two popular choices are variational expectation–maximization inference as proposed in the initial Blei et al. (2003) paper and Gibbs sampling proposed by Griffiths and Steyvers (2004). These inference techniques allow updating of the model from its initial semirandom allocation of topics to words and documents to a converged determination of probabilistic topics per document and across the corpus.

A last consideration, before moving to the next section, is to consider the specific suitability of the LDA method for reviewing academic research documents. The method is currently used widely for reviewing the academic literature and detailed support for this approach in various disciplines is provided in Asmussen and Moller (2019), Maier et al. (2018) and Quinn et al. (2010). Asmussen and Moller (2019) start their study by noting: "Manual exploratory literature reviews are soon to be outdated. It is a time-consuming process, with limited processing power, resulting in a low number of papers analysed" (p. 1). They note that LDA is particularly suited compared to systematic review for analysing a large literature, where there are practical limits on how many studies researchers can physically review. LDA is less suited for highly accurate distillations of knowledge when a smaller number of studies are being reviewed. Moro et al. (2015) also note this size-related benefit of LDA in their LDA-based review of the research literature on the returns from investing in banking information technology. As our review is focused on reviewing a large and disperse literature, our use of LDA, therefore, fits well with prior recommended use.

[3]Dirichlet distributions are well described in Boyd-Graber et al. (2017) in the context of LDA

We now proceed to describe our corpus of ML and finance articles, our preprocessing choices undertaken to generate the DTM and the LDA modeling options applied to the topic extraction.

3 | CORPUS AND PROCESSING

We use the Elsevier *Scopus* database to identify relevant prior research on ML in finance based on term searching within title, abstract and keywords. The research corpus selection was a process of experimentation in terms of identifying suitable search terms and one where we had to iterate towards the best search terms. Our search criteria required a match of both a ML and a finance term in order for an article to be selected, but in initial searches, we were retrieving significant 'intruder' articles that were being captured by the search terms but clearly not related to ML in finance. For example, the combined ML term of *neural network* and finance term of *bank*, resulted in numerous articles about the application of *neural networks* within image *banks* for improved image recognition. An advantage of the topic modeling approach is that the presence of intruder articles is clear from the topics generated, as these articles cluster in topics that are clearly not related to ML in finance.

Our final search criteria, which minimized, but didn't fully eliminate,[4] intruder articles is contained in Table 2. ML search terms were selected initially based on searching for key topics in de Prado (2018), a recent well-regarded industry-academic textbook on ML applied to finance. From this the selection of ML search terms was expanded out based on author knowledge. For finance search terms, there was a difficulty in using general finance terms such as 'finance', 'equity' and 'investment'. Often in the retrieved articles with these terms they were either used in an alternative sense to the standard finance meaning (particularly 'equity' and 'investment') or were present in many abstracts but not the core focus of the study (especially 'finance'). We, therefore, developed a list of more specific finance search terms based on author knowledge and the structure of finance research. As there is a potentially endless list of terms that could be included as search terms we experimented with alternative search terms for both ML and finance and did not find significant article count variation. This suggests our search strategy resulted in a reasonable convergence on a suitably identified body of research.

The date range is 1990–2020. We included conference papers published in conference proceedings in our sample as this is the primary mode of research communication in computer science (Bar-Ilan, 2010), in contrast to finance practice, due to the desire for speedy communication. Lastly, we limited the subjects to the *Scopus* subjects of computer science, decision making and business and economics to avoid topic sprawl. Our final data set from this search process is 5942 articles which is the starting point for our analysis. The topic modeling analysis is conducted on the abstracts of these articles. Figure 3 shows the distribution of these articles by year. This figure tells us that our articles are primarily since 2006, with 40% of articles coming from just the last 3 years. Among the top journals which are popular for publishing ML and finance research are *Expert Systems with Applications*, *Neurocomputing*, *European Journal of Operational Research*, *Quantitative Finance*, *Journal of Forecasting* and *Journal of Banking & Finance*.

[4]The cost of fully eliminating irrelevant articles would have meant also eliminating significant numbers of relevant articles.

TABLE 2 Machine learning and finance literature search strategy

This table reports the search strategy used in conducting the literature review through Elsevier *Scopus*. Search strategy required a match of at least one machine learning search term and one finance search term. Match was conducted on title, keywords and abstract. For the final data set all relevant article features were retrieved on matched articles and the topic modeling analysis was based on the article abstract.

Machine learning search terms
"machine learning"; "artificial intelligence"; "support vector"; "deep learning"; "neural network"; "random forests"; "gradient boosting"; "ensemble learning"; "reinforcement learning"; "natural language processing"; "Latent Dirichlet"; "Latent semantic"; "unsupervised learning"; "supervised learning"; "statistical learning"
Finance search terms
"financial markets"; "equity investment"; "mergers and acquisitions"; "stock market"; "stockmarket"; "investor"; "equity securities"; "commercial bank"; "investment bank"; "corporate finance"; "credit institution"; "fixed income securities"; "corporate debt"; "debt markets"; "financial derivatives"; "financial crisis"; "financial risk management"; "financial instruments"; "volatility risk"; "liquidity risk"; "credit risk"; "algorithmic trading"; "foreign exchange"; "commodity markets"; "debt capital"; "equity capital"; "cost of capital"; "dividend policy"; "treasury management"; "interest rate risk"; "capital structure"; "discounted cash flow"; "real options"; "contingent claim"; "financial model"; "Value at Risk"; "financial engineering"; "initial public offering"; "investment management"; "hedge fund"; "mutual fund"; "investor sentiment"; "financial agent"; "project finance"; "idiosyncratic risk"; "fintech"; "financial contagion"; "financial stability"; "cryptocurrency"
Other search criteria
Date range: 1990–2020
Search type: Title, abstract, keywords
Publication type: Articles; conference papers
Source type: Journals; conference proceedings
Language: English language only
Subjects: Computer science; business, management and accounting; decision sciences; economics, econometrics and finance

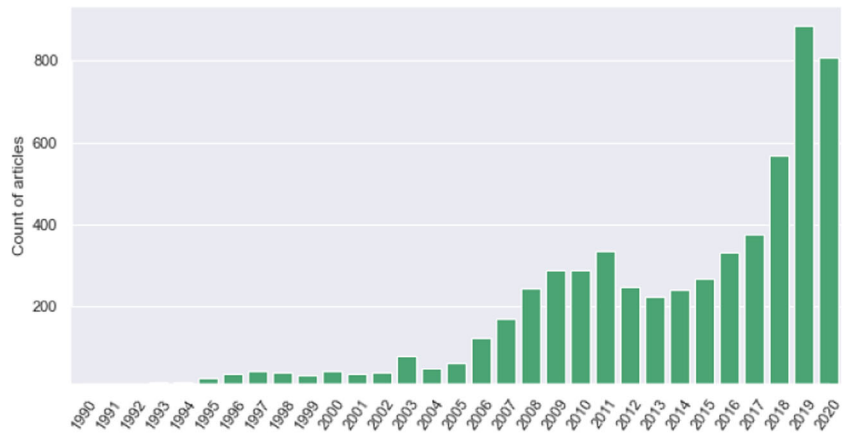

FIGURE 3 Article distribution by year

For preprocessing we used the *R* packages *tm* and *quanteda*. We performed the following steps: First, all text in each abstract is converted to lowercase letters and digits and punctuation as well as stop words (common words such as: and, or, not, I, you) are removed. Second, we applied Porter's stemming algorithm that reduces each word to its stem (Porter, 1980). Third, we removed the top 10% of terms based on the TF-IDF score. The 10% choice was based on visual inspection of terms remaining and experimenting with between 5% and 15% top TF-IDF term removal. We also removed terms that occur less than five times in the corpus. Lastly, as the processing of the vocabulary leads to substantial shortening of the abstracts, we only keep abstracts with at least five remaining terms. These preprocessing steps lead to a reduction of the vocabulary in our DTM from 17,191 initial unique terms to 3139 final terms. Due to the length restriction on abstracts we keep 5801 out of 5942 abstracts. The final DTM, therefore, has dimensions of 5801 and 3139. The values in the DTM provide the number of occurrences of each term of the vocabulary in each document, with most of the entries being zero.

For the topic modeling itself we used the R package *topicmodels* (Hornik & Grün, 2011), with inference from the variational expectation–maximization algorithm as per the original Blei et al. (2003) process. The Dirichlet prior of α is set at $50/k$, where k is the number of topics. β is estimated within the model. These are the default parameters to estimate LDA models in the *topicmodels* package and there was no notable rationale for adjusting them. The number of topics k is the major input choice parameter for our topic model. We estimate separate models for 10–50 topics in increments of five. As the resulting models do not necessarily find global extrema, for each k we run 10 models with randomly chosen (but fixed across the various k) starting points. This results in 90 sets of topic models (9 k parameters $\times$ 10 random starting seeds).

For each set of topic models a Krippendorff's α (Krippendorff, 1970) is calculated where lower values of this α identify topic models with sharper topic distributions. The assessment of the quality of topic models can be guided by this α value, however, it also requires human judgment, in particular, regarding cohesion of topics and absence of intrusive words (Chang et al., 2009; Hannigan et al., 2019). We, therefore, further screened our grid of topic models regarding our criteria of cohesive, sense-making topics as well as an overall parsimonous model. For this, we manually inspect among the lower α topic models to choose which seems to be the most coherent in terms of individual topic composition. We inspected the top 10 terms for each topic (the terms with the highest probability of attachment to a topic) and assessed them regarding their coherence as well as the top abstracts with the highest fit to the topic (this fit being referred to as θ). The topic model with the most coherent topics and best fit of top articles was chosen, which in our case is the 20 topic model and with an α of 0.1085.

4 | TOPICS OF ML AND FINANCE

Twenty topics were initially extracted in the preferred LDA model and of these we classify 15 as strong ML and finance topics. The remaining five unreported topics either have low coherency (e.g., consisting of a mix of diverse smaller topics) or are best classified as pure ML topics. This presence of low coherency topics is a common issue in topic modeling due to the method forcing all documents to match to one or more topics (Chang et al., 2009).

To facilitate reporting of the results we assign a label to each topic that reflects the essence of the topic and based on our field knowledge we arrange the topics into four categories: (1) Asset pricing forecasting techniques, (2) financial markets analysis, (3) financial risk

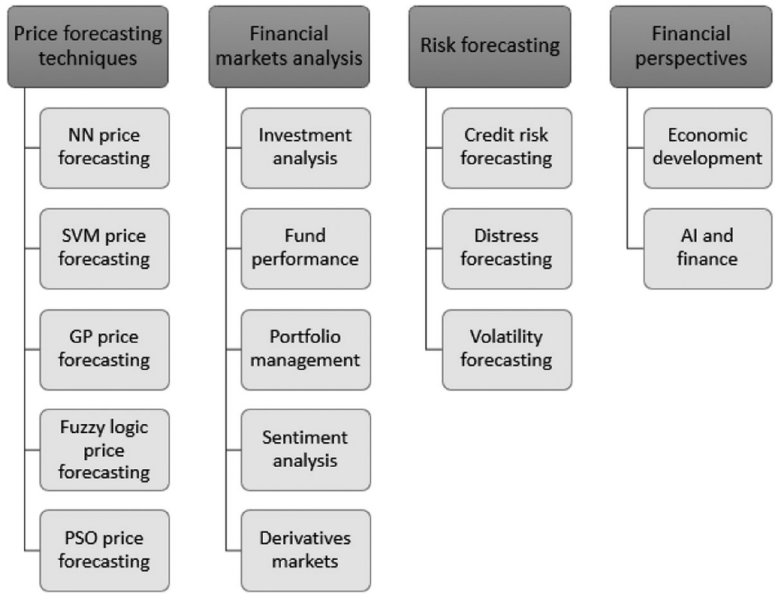

FIGURE 4 Extracted topics and categories

forecasting and (4) other topics. See Figure 4 for the topics assigned to each category. We can see from these topic and category labels that there is a strong focus on forecasting, a core strength of ML techniques. The topics are formally presented in Table 3, which includes the (stemmed) keywords that best describe the article and the articles with high matches to topics. A core strength of the LDA approach over a conventional survey is that LDA provides, as output, a sorted list of documents (articles) that are the highest match to each topic. It is, to a much greater extent, a choice of the author what articles are selected in a conventional survey and this can introduce a bias of focus to the review.[5] Our analysis of topics, therefore, focuses on these high-match articles.[6] To aide the analysis, we also provide Table 4, which shows the overall popularity of each topic in terms of the frequency of matched articles from the data set to each topic. We now proceed to discuss the topics within each category.

4.1 | Price-forecasting techniques

The category of asset price-forecasting techniques focuses on the ability of the techniques of ML to forecast financial prices. Five techniques are prominent in this category: Artificial neural networks (NN), support vector machines (SVM), genetic programming (GP), fuzzy logic and particle swarm optimisation (PSO). We discuss each technique along with the prior application. As we do not have the space to delve deeply into the techniques our focus is on the most

[5]Another potential bias of a conventional survey which LDA can overcome is a bias towards selecting older articles, as these have the highest verifiable impact through citations. LDA does not prioritise older articles.

[6]As this paper is necessarily an overview of each topic, we also have provided, through Mendeley Data, open access to the full data set which provides the topic composition of all articles (not linked to this draft due to author identification).

TABLE 3 Topics of machine learning in finance

This table presents the topics of machine learning (ML) in finance, based on the latent Dirichlet allocation model on 5942 articles that matched ML and finance search terms in *Sco—pus*. We report the 15 topics that show strong topic consistency with ML applied to finance. The topic label is assigned by the authors, as are the topic categories. Top 10 words are the stemmed words that probabilistically best describe the topic. Sample articles are articles with a minimum of 70% topic match and that provide a useful illustration of the topic. The θ % is the actual topic match of the specific article. See Section 4 for further description on the methodology and explanation of choices made in the topic extraction. The shaded rows are used to distinguish the different categories and serve no other purpose.

No	Category	Topic label	Top topic terms	Sample articles and θ %
1	Price-forecasting techniques	NN price forecasting	layer, recognit, phase, rnn, sequenc, tempor, interact, neuron, visual, graph	S.-S. Kim (1998)—96 R. Zhang et al. (2018)—92 Long et al. (2020)—73
2	Price-forecasting techniques	SVM price forecasting	svm, cluster, kernel, filter, dimension, reduct, regular, binari, matrix, probabilist	L. Wang and Zhu (2010)—84 McCluskey and Liu (2017)—83 M.-C. Lee (2009)—75
3	Price-forecasting techniques	GP price forecasting	detect, trader, frequenc, evolutionari, fraud, mlp, intraday, highfrequ, anomali, chart	Chen et al. (2009)—89 Li et al. (2010)—89 Cui et al. (2010)—88
4	Price-forecasting techniques	Fuzzy logic price forecasting	fuzzi, logic, prefer, neurofuzzi, anfi, qualit, membership, student, compet, person	Soto et al. (2019)—85 Naranjo et al. (2015)—83 Rahamneh et al. (2010)—83
5	Price-forecasting techniques	PSO price forecasting	ensembl, bp, ga, predictor, earli, pso, bitcoin, rough, warn, pca	Chen et al. (2007)—83 Z. Zhang et al. (2017)—83 Guo et al. (2006)—78
6	Financial markets analysis	Investment analysis	project, countri, flow, ma, revers, valuat, cash, interv, cryptocurr, estat	Liu (2008)—80 Liao and Ho (2010)—79 Lei and Qiu (2020)—76
7	Financial markets analysis	Fund performance	ann, fund, net, window, arima, electr, mutual, week, rbf, indian	K. Wang and Huang (2010)—85 Pan et al. (2017)—84 Rout et al. (2020)—82

TABLE 3 (Continued)

No	Category	Topic label	Top topic terms	Sample articles and θ %
8	Financial markets analysis	Portfolio management	portfolio, stage, alloc, liquid, constraint, sharp, dss, multiobject, equiti, optimis	Pai (2018)—94 Singer (1997)—91 S. Kim (2018)—83
9	Financial markets analysis	Sentiment analysis	sentiment, news, text, opinion, twitter, tweet, emot, content, document, post	Ito et al. (2017)—93 Deng et al. (2017)—93 Sun et al. (2019)—85
10	Financial markets analysis	Derivatives markets	foreign, currenc, oil, china, commod, mode, forex, pair, decomposit, dollar	He et al. (2018)—91 Yao and Tan (2000)—91 Misra and Goswami (2015)—87
11	Risk forecasting	Credit risk forecasting	credit, bank, score, loan, custom, default, commerci, lend, borrow, card	Shivanna and Agrawal (2020)—94 Turkson et al. (2016)—92 Y. S. Kim and Sohn (2004)—91
12	Risk forecasting	Distress forecasting	corpor, firm, bankruptci, distress, properti, failur, rank, dividend, sale, debt	Yang et al. (2009)—84 Le et al. (2019)—80 Bae (2012)—79
13	Risk forecasting	Volatility forecasting	nois, svr, wavelet, stochast, var, nn, garch, bpnn, bound, chaotic	Xu et al. (2011)—84 Takeda and Kanamori (2014)—83 Taylor (2000)—82
14	Financial perspectives	Economic development	crisi, lstm, enterpris, macroeconom, china, energi, suppli, extern, bayesian, chines	Z.-J. Wang and Zhao (2021)—88 Kolupaieva et al. (2019)—76 Senoguchi and Kurahashi (2013)—72
15	Financial perspectives	AI and finance	ai, servic, big, digit, regul, earn, recommend, firm, ml, ipo	Bataev (2018)—92 J. Lee (2020)—91 Lui and Lamb (2018)—87

TABLE 4 Topic frequency

This table reports, for the 15 topics in Table 3, from the 5942 machine learning in finance articles used in the analysis, the frequency of topic match for articles. We report the frequency of articles that have at least a 50% match to a particular topic and the frequency of articles where the topic is the largest topic in an article. Percentages in the 50%+ column are the percent of articles where a single topic is 50%+ of article. Percentages in the largest topic column are percent of all 5801 retained articles (5.942 before preprocessing) in the data set. Note that percentages in the last column do not sum to 100% as we originally extracted 20 topics in our modeling, of which only 15 are retained as being strong matches to machine learning in finance topics, so a proportion of articles have a largest topic match to the excluded topics. The shaded rows are used to distinguish the different categories (see Table 3) under which we have sorted the topics and serve no other purpose.
Abbreviations: AI, artificial intelligence; GP, genetic programming; NN, neural networks; PSO, particle swarm optimisation; SVM, support vector machines.

No	Topic label	Frequency 50%+ articles	Frequency largest topic
1	NN price forecasting	55 (2.63%)	251 (4.33%)
2	SVM price forecasting	79 (3.78%)	295 (5.01%)
3	GP price forecasting	68 (3.26%)	240 (4.14%)
4	Fuzzy logic price forecasting	82 (3.93%)	241 (4.15%)
5	PSO price forecasting	52 (2.49%)	224 (3.86%)
6	Investment analysis	62 (2.97%)	196 (3.38%)
7	Fund performance	93 (4.45%)	295 (5.09%)
8	Portfolio management	113 (5.41%)	292 (5.03%)
9	Sentiment analysis	233 (11.16%)	446 (7.69%)
10	Derivatives markets	124 (5.94%)	334 (5.76%)
11	Credit risk forecasting	262 (12.55%)	512 (8.83%)
12	Distress forecasting	97 (4.65%)	306 (5.27%)
13	Volatility forecasting	136 (6.51%)	376 (6.48%)
14	Economic development	57 (2.73%)	278 (4.79%)
15	AI and finance	79 (3.78%)	249 (4.29%)

relevant current applications of these techniques in finance. Overall these topics, as seen in Table 4, account for being the largest topic in nearly 30% of all articles in the data set.

The first topic in this category, asset price forecasting using NN, applies core models of DL to finance. The keywords from this topic include component parts of NN modeling (layer, neuron) as well as indicators of specific techniques, including 'RNN' for recurrent neural networks RNNs are a form of NN that incorporate memory in the model learning process as data is modeled as a sequence (Goodfellow et al., 2016, ch. 10) and this has proven particularly useful in financial time series forecasting where memory should be a factor (Lo, 1989).

The matched articles for this topic (S.-S. Kim, 1998; Long et al., 2020; R. Zhang et al., 2018) all deal with RNNs to some extent. S.-S. Kim (1998) compares a range of RNNs in terms of Korean stock market prediction ability. R. Zhang et al. (2018) shows how an integration of RNNs and another popular form of NN—convolutional NN, can improve stock market pricing

forecasts. Long et al. (2020) shows how deep NN can blend pricing data and trader behaviour information to improve forecasts of pricing.

SVMs are another popular technique of ML and covered in Topic 2. SVMs are, by origin, classification techniques based on maximising distinct classes of data on a hyperplane with a number of dimensions equal to the number of model features (Hastie et al., 2009; ch. 12). The technique can be applied to regression in addition to classification. The keywords highlight 'SVM' as the top term, but also some terms from related techniques. In the matched research, M.-C. Lee (2009) applies SVM to price prediction of the Nasdaq. The model uses feature extraction from nine commodity prices, 11 currencies and nine equity indices to predict 15 days ahead of Nasdaq performance. They are able to show that this outperforms a standard NN approach. L. Wang and Zhu (2010) proposes an improvement on standard SVM approaches, whereas the more recent McCluskey and Liu (2017) compares SVM for price forecasting to another popular ML method—gradient boosting.

In Topic 3, GP price forecasting, we see a cluster of genetic algorithm-related research applied to price forecasting. In genetic algorithms a population of possible solutions to a problem are taken as a starting point and, for a given fitness criteria and allowing for solution mutation, a preferred solution is arrived at in a manner that mimics natural evolution (Holland, 1992). GP is useful in narrowing down a range of potential solutions, such as a selection of trading strategies, to the most likely candidates for success. GP itself is an optimisation technique rather than a form of ML, but can be integrated in ML (Mabu et al., 2007). We can see that some GP-related stemmed terms appear in the topic keywords, such as 'evolutionari', but in general, the terms are not strong descriptions of the studies clustered in this topic. Chen et al. (2009) introduce the potential for Genetic Network Programming, an ML-capable version of GP, in finance. Their forecasting, based on GP applied to a range of technical stock market indicators in the Japanese market, is shown to outperform more common alternative approaches. Li et al. (2010) extend this evolutionary algorithm approach with the introduction of subroutines in the programming that act as small scale models within the overall model (an approach somewhat similar in intention to bootstrapping in regular finance econometrics). Cui et al. (2010) shows how grammatical evolution, a subtechnique of GE, can be used to improve trader ordering efficiency by lowering cost.

Fuzzy logic price forecasting is Topic 4 of the price-forecasting techniques category. In this topic, fuzzy logic is applied as an alternative to probabilistic thinking to account for uncertainty in decision making (Zadeh, 1988). We see a clear match between both the keywords and the matched research articles for the topic. Rahamneh et al. (2010) makes the general case, in comparison to NN, for how fuzzy logic can be used for stock market forecasting. Naranjo et al. (2015) shows how fuzzy logic can be used to filter and determine suitable trading signals. Soto et al. (2019) shows the power of some of the newer fuzzy logic models, including modern variations of the ANFIS (Adaptive Network-based Fuzzy Inference System) NN approach originally developed by Jang (1993). In common with the other technique category topics, these studies indicate the strong growth of complexity in methods over time.

The last topic (Topic 5) on ML price-forecasting techniques is on particle swarm optimisation (PSO). PSO is a type of evolutionary algorithm inspired by how birds flock and fish school together when searching (Kennedy & Eberhart, 1995). The technique allows multiple generations of an algorithm that learn from prior generations. Of our matched studies, Chen et al. (2007) speaks about PSO in general, whereas Guo et al. (2006) and Z. Zhang et al. (2017) show how to integrate PSO within NN to improve prediction.

In this section we have shown how five techniques of ML are particularly prominent for price forecasting in finance. This includes NN and SVM price forecasting, which would be many researchers' initial perceptions of what ML is, but also includes topics around GP, fuzzy logic and PSO. A common feature of the research on each set of techniques is that, over time, each topic has developed specialised focus on finance-specific versions of ML techniques to best work with finance data. We now move from techniques to applied topics of ML and finance, starting with financial markets analysis.

4.2 | Financial markets analysis

The second category, financial markets analysis, reflects the range and diversity of research on financial markets in general. We identify five topics in this category: Investment analysis, fund performance, portfolio management, sentiment analysis and derivatives markets. These topic offer an interesting insight into potential future research directions across financial ;markets.

The first topic covers a range of investment analysis studies that have some form of ML approach built into the investment analysis. The range is quite wide, for example, of the sample studies, Liu (2008) proposes using NN to improve discounted cash flow methods, Lei and Qiu (2020) uses an ML approach to understand the investment climate in countries on the 'Belt and Road' economic expansion initiative and Liao and Ho (2010) examines how to value investments based on a combination of real options and fuzzy logic. The interesting takeaway here, is that all subfields of investment analysis are potentially up for grabs.

Fund performance (Topic 7) is the second topic in this category, with the articles mainly concentrating on assessment and comparison of mutual fund investment performance. K. Wang and Huang (2010) develops a NN classifier to rapidly detect changes in mutual fund performance. Their fast adaptive neural network classifier (FANNC) is shown to be superior to a traditional NN for both mutual fund classification in strata of performance and for predicting future fund performance. A key feature of the FANNC is to allow new information to be constantly fed into the model and for the model to subsequently efficiently update. Pan et al. (2017) mixes NN with Data Envelopment Analysis (an operations research version of finances' efficiency frontiers) to show have this can be used to better predict investment performance. In Rout et al. (2020), they use a range of NN models to predict net fund values. A key finding they make is how sensitive predictions of fund values are to NN starting assumptions.

Topic 8, portfolio management, has some overlaps with fund performance, given that fund performance is, in large part, determined by the performance of portfolios. This category is primarily concerned with algorithms of rebalancing portfolios. Singer (1997) speaks about this from a broad algorithmic perspective. More recently in our sample studies, Pai (2018) shows how active rebalancing can be achieved for multiple (risk and return) objectives using metaheuristics and S. Kim (2018) shows how to create low volatility portfolios based on SVM.

Topic 9 is sentiment analysis. Sentiment analysis has a long history in finance, including popular early studies by Baker and Wurgler (2007) and Brown and Cliff (2004). More recently, Bollen et al. (2011) showed how social media sentiment can be a measure of investor sentiment. It is in this vein, because of the easy access to social media data and familiarity with its use, that ML researchers have examined sentiment in financial markets. The keywords for the topic highlight this social media angle intermingled with

sentiment keywords. Natural language processing is a core approach to research in this area, normally combined with SVM or Naive Bayes modeling (a classifier alternative to SVM), to distinguish between positive, neutral and negative sentiment. Sun et al. (2019) uses both the social media content as well as meta-characteristics of the posting patterns to improve predictions of Chinese stock market trends. They find an inverse U-shaped curve relationship between stock return and the attention of both news media and investors with a positive moderating effect of news media attention and social interactions on the stock return. Among our other sample studies, by Deng et al. (2017) and Ito et al. (2017), we see further analysis of social media posts for sentiment extraction.

An interesting additional matched paper to this topic that we identify is Smailović et al. (2014), which provides some indication of how this area is developing and, therefore, is interesting in this regard. They use a mixture of SVM as a classifier for grouping tweets into positive, neutral, negative, but with a testing approach based on stream-based active learning. Therefore, it moves beyond traditional ML where data is divided into training and testing groups (and similar variations thereof) and classification takes place in the training group and these classifications are then applied in the testing group. Instead, their approach takes advantage of the constant flows of tweets to allow their model to choose to update the classifier if it is experiencing particularly high uncertainty about new tweets. In this case the machine will request that these tweets are human-labelled as positive–neutral–negative so uncertainty can be reduced. Thus, the model is not just blindly applying the initial classification, but is trying to identify occasions when it needs to be improved.

Derivatives markets are the main interest of Topic 10, the last of the financial markets analysis category topics. Forex markets have long been attractive for algorithmic trading as they have historically been the largest and most liquid financial markets. From an ML researchers perspective the attractiveness is also the availability of long and continuous time series data. In the matched articles, we see that all derivatives markets are covered, not just forex as one of the largest derivatives categories. Yao and Tan (2000) presents a basic NN model to forecast the exchange rates between American Dollar and five other major currencies. Their paper highlights several methodological issues on the frequency of sampling, choice of network architecture, forecasting periods and measures for evaluating the model's predictive power. Similar styles of studies for other derivatives markets are also included in our sample studies, including on the oil market (He et al., 2018) and sugar futures (Misra & Goswami, 2015).

4.3 | Risk forecasting

The third category of the topics relates to the area of risk forecasting, for which we have three of the 15 topics extracted by our model. The topics are: Credit risk forecasting, distress forecasting and volatility forecasting. As noted in the introduction, risk forecasting was one of the original focuses of ML in finance, due to the strong methods in ML for forecasting and classifying various types of risk. Forecasting risk emanating from different and unexpected sources is a particularly pertinent task in the overall financial risk management process and something for which ML is well-suited.

The first two risk forecasting topics are credit risk forecasting (Topic 11) and financial distress forecasting (Topic 12). These two interconnected topics were a core focus following the global financial crisis, amid the seeming clear failure of traditional methods and techniques to forecast credit risk and predict financial distress. Moreover, forecasting credit risk and

predicting financial distress are the two topics that demonstrate the most practical application of ML techniques.

We started witnessing the use of ML techniques in credit risk forecasting and financial distress prediction from the early 1990s with the comparison of neural network based distress and bankruptcy prediction models with traditional statistical methods. Taking the classical linear, logit and probit regressions models proposed by Altman (1968) that dominated industry practice for decades, there was a clear and significant improvement observed in predicting financial distress and defaults when traditional models were combined with the neural network based models of credit risk starting with (Altman et al., 1994).

Building on the start offered by Altman et al. (1994), Y. S. Kim and Sohn (2004) tests neural network models to identify the misclassification of customers (good and bad borrowers) observed in classical credit scoring models. Two recent studies including Shivanna and Agrawal (2020) and Turkson et al. (2016) highlight some recent trends in ML techniques. Shivanna and Agrawal (2020) shows the comparative power of a range of ML approaches to credit risk forecasting and Turkson et al. (2016) perform a comparative analysis using bank credit data across 16 ML classification techniques and find the best method fits credit yield data with over 80% predictive accuracy about a customer's default probability. A major issue in the strength of credit risk modeling is the availability of customer characteristic data, which is noted consistently in research in this area. A working assumption is that more customer data will generative even better predictive accuracy.

Distress forecasting (Topic 12) is similar to credit risk forecasting in that both have a common feature of modeling an end result of low or zero loan repayment. Distress forecasting, however, is focused primarily on firm loan repayments, whereas credit risk forecasting is currently focused (but not by definition restricted to this) on personal loan repayment. Therefore, the ML methods that work for credit risk forecasting should also work for distress prediction. This topic is particularly concerned with the many drivers of corporate default; from firm-specific factors, to macroeconomic drivers, to various other potential predictors. As ML is particularly effective at sorting large volumes of predictive variables, it has been shown to be very effective at firm default prediction. Yang et al. (2009) applies NN for this purpose, whereas Bae (2012) applies SVM. Le et al. (2019) incorporates more recent computing power in their analysis and propose a gradient boosting technique that they argue has improved prediction ability.

Volatility forecasting (Topic 13) is the third topic in the category of financial risk forecasting. The topic is primarily focused on showing the power of ML as an alternative to econometric techniques for estimating volatility. We see this in Xu et al. (2011) which contrasts the explanatory power of SVM with GARCH modeling of volatility. They demonstrate that SVM has superior forecasting power to GARCH for Chinese stock data. It is evident from the Topic 13 keywords, in addition in the matched research, that not just volatility is a focus of the topic, but also a particular focus on volatility as part of Value-at-Risk models. This focus is seen in both Taylor (2000) and Takeda and Kanamori (2014).

4.4 | Financial perspectives

The fourth and final category of the topics covers financial perspectives, of which we identify two topics: Economic development and AI and finance. We term these topics as 'perspectives', as both categories are quite forward-looking. The first category on economic development

clusters a range of economic ideas, with a common underlying theme of forecasting based on macroeconomic variables. For example, Z.-J. Wang and Zhao (2021) look at predictability of EU ETS carbon trading schemes using a Bayesian network understanding of macroeconomic variables. Kolupaieva et al. (2019) examine the impact of economic variables on financial distress in Ukrainian metallurgical firms, while Senoguchi and Kurahashi (2013) try to predict financial crises based on macroeconomic values. In all cases, ML models are used in the forecasting.

Topic 15, our last topic, could also have been our first topic, as it is about the impact of AI on finance. To some extent this is the topic of our overall study, as the discipline has started to focus on this broader dimension of what AI means for finance. Though this could be a study in itself, in our context we have some sample studies which cover a range of technological as well as legal interactions with finance focusing on AI outcomes. Bataev (2018) provides an overview of how finance has been, and will be, impacted by the arrival of big data methodologies. They chart this to the rise of fintech as a new industry and some of the problems faced by traditional finance. J. Lee (2020) analyses some of the legal issues of implementing AI in finance with a particular focus on the rising area of 'RegTech'—using AI to meet compliance requirements. The last of our sample studies by Lui and Lamb (2018) discusses data protection in the era of AI in finance. We, therefore, see quite a range of topics related to the impact of AI in finance.

4.5 | Topic evolution over time

As a last analysis, we track the evolution of our 15 topics over time. This can be seen in Figure 5, which shows an alluvial diagram of topic distribution over the data set time period of 1990–2020. The chart is sorted to show the most popular topics at the beginning of the time period from the top of the left *y*-axis and similarly the most popular topics at the end of the time period from the top of the right *y*-axis.

On a broad analysis of the chart we see the most popular topics at the beginning of the time period were focused on what might be considered classic ML and finance—investment performance. The recent most popular topics are much more data-based and broader in scope, including market sentiment and credit risk forecasting. This probably reflects the rise in data availability in recent years as well as new implementations of techniques based on the practical success they have demonstrated.

The chart, however, needs to be viewed in conjunction with Figure 3 which showed very few ML and finance papers before about 2007. For that reason it is probably best to compared the changes from the period marked 2005–2009 on the chart to the present day. Looking at the chart with this perspective we see a very large growth in market sentiment studies over this time period: From the least popular to the most popular topic. This might be due to the current focus in ML on the promise of textual statistical learning, as text is a key feature of ML market sentiment analysis. Credit risk forecasting has remained consistently strong, as has volatility forecasting. Fuzzy logic price forecasting has declined significantly in interest. There has also been a very large rise in the general 'AI and finance' topic, from near the bottom to near the top. Given what we saw of this topic in the last section, this suggests a rise in interest in fundamental questions of the impact AI will have on the practice of finance.

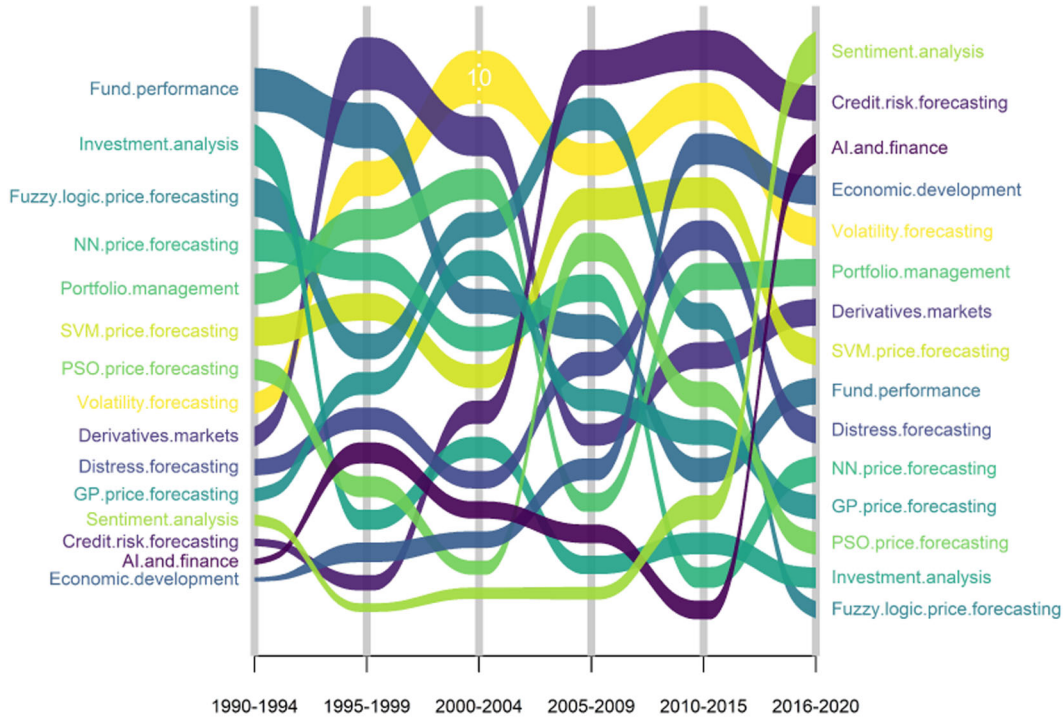

FIGURE 5 Evolution of topics over time

5 | DISCUSSION AND CONCLUSIONS

In this study, we have provided a comprehensive application of topic modeling to understanding bodies of finance research. Our demonstration with a diverse literature on the application of ML in finance shows the strength of this technique for holistically identifying and grouping relevant research on a topic. The 5942 articles we identify are spread across the fields of computer sciences, decision sciences, economics, econometrics and finance and other business disciplines; yet we show how the LDA topic modeling technique can cleanly structure this diverse corpus of research into coherent topics.

Apart from the introduction of topic modeling to financial research structuring, the core contribution of this paper is the actual mapping of the ML in finance literature. Only a small fraction of the application of ML to financial problems is published in finance journals, despite the drive within financial research to continuously improve forecasting and modeling techniques. To some extent this could become a critical issue for finance research, as industry practice in finance moves beyond the techniques discussed in finance research. Our mapping shows the structure of this literature and will be of benefit to researchers in these topics seeking to augment their current research with the techniques of ML.

We identify four overall categories of topics: Price-forecasting techniques, financial markets analysis, risk forecasting and financial perspectives, as well as 15 topics spread across these topics. These categories, while sharing some techniques, are progressing at different speeds. Risk management, thanks to an early research lead, is well advanced in ML application of the modeling of risk. This is seen in the practical field of credit risk modeling to support bank and other financial institution lending decisions. Asset modeling and forecasting is also advanced

due to the early understanding of the suitability of neural networks for modeling financial times series. Also of help in this category has been the use of the clean data sets of financial time series as sample applications of ML techniques by nonfinance ML researchers. Less advanced, but possibly more promising, is the category of financial markets analysis. In this category lies the promise of new data generation, such as on investor sentiment and investment performance analysis. Our analysis of topic research over time, shows investor sentiment to be one of the main growth categories. This category opens the potential for the future of finance with applications such as robo-advising and other financial advisory services. It also relies on textual statistical learning for which methods of analysis are currently rapidly advanced.

Of additional interest are the topics of finance that are not yet addressed by ML, particularly around corporate finance and the roles carried out by investment bankers such as mergers and acquisitions and firm financing decisions. This does not mean that research is not being conducted within these areas, but just that the research is not of a sufficient quantity to be categorized as a self-standing topic. One major issue here is likely to be availability of data to study the issues. Finance research in general has quite limited access to data within firms and within investment banks, instead being limited to externally observed data about the activities of these organizations. Research that can access this privately owned data within firms is needed to open this area. Also missing as a topic is financial network analysis, a large focus of the broader AI research movement. This probably reflects that there are quite limited data sets available for understanding the impact of networks on financial behavior. There is also scope for textual analysis to expand outside of just market sentiment. Financial activities are heavily documented activities and as the ML techniques of textual analysis improve we should see greater focus on extracting knowledge from these data reservoirs. We also don't see much mention, yet, of currently popular topics such as cryptocurrencies or the impact of COVID-19 in our topics (although cryptocurrencies is one of the topic terms in Topic 6 [investment analysis], it is not the topic itself). This is partially due to the wide time window we tested where it is difficult for brand new areas of interest to assert themselves as standalone topics.

We also note some limitations to the LDA approach to research summary. Earlier in the study we noted that LDA is suitable for summary of large bodies of research, but misses out on some element of accuracy that would be seen from more personalised research reviews, such as systematic reviews. A researcher tends to engage in a systematic research review process with an eye to what is missing and understudied, as this is of particular interest to future researchers. LDA, by contrast, is not designed to do this. We can, however, note what topics do not emerge from the LDA analysis as a signal as to what are the under-researched areas and that is what we do above. However, the level of detail is not at the same level as in a systematic review. It might be worth considering this study, rather, as a first step in a multistep process of reviewing the literature. Where our study has provided the broad structuring of the literature and follow-on studies can now take some of the individual topics for more detailed analysis. This might take the form of a systematic review or even a further level of topic modeling within each of the broader topics of the area.

Other future research avenues, apart from the deeper exploration of individual topics, include considering some of the more advanced techniques that have emerged from LDA in recent years. This includes structural topic modeling (Roberts et al., 2014), which offers some advantages in terms of topic coherence, but was considered beyond the scope of this particular study. We also suggest that there would be benefits to a deeper comparison of topics at the discipline level—to explore how different disciplines have explored the same topics with differing approaches.

We conclude on a note of optimism about the potential to expand the financial testing repertoire to incorporate the techniques outlined by this paper. The field of financial research has always been driven by the need to be practically relevant to financial practice. Our research, thus, has offered the finance industry in the past implementable approaches to investment selection and trading, applied corporate financing approaches and developed new financial products and services. ML, as demonstrated in this study, can clearly strengthen the finance researcher toolkit, due to its strong emphasis on practical understanding. We propose that the structure provided in this study offers some guidance as to how to incorporate these techniques suitably in future research.

REFERENCES

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, *18*(3), 505–529.

Asmussen, C. B., & Moller, C. (2019). Smart literature review: A practical topic modeling approach to exploratory literature review. *Journal of Big Data*, *6*(1), 1–18.

Athey, S. (2018). The impact of machine learning on economics. In A. Agrawal, A. Goldfarb, & J. Gans (Eds.), *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Bae, J. K. (2012). Predicting financial distress of the South Korean manufacturing industries. *Expert Systems with Applications*, *39*(10), 9159–9165.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–152.

Bar-Ilan, J. (2010). Web of science with the conference proceedings citation indexes: The case of computer science. *Scientometrics*, *83*(3), 809–824.

Bataev, A. V. (2018). Analysis of the application of big data technologies in the financial sphere. In *2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS)* (pp. 568–572). IEEE.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, *16*(16), 17–24.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, *11*(2–3), 143–296.

Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, *11*(1), 1–27.

Cerchiello, P., Giudici, P., & Nicola, G. (2017). Twitter data models for bank risk contagion. *Neurocomputing*, *264*, 50–56.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models, *Advances in Neural Information Processing Systems 21 (NIPS)* (pp. 288–296). Curran Associates, Inc.

Chen, Y., Mabu, S., Hirasawa, K., & Hu, J. (2007). Genetic network programming with sarsa learning and its application to creating stock trading rules. In *2007 IEEE Congress on Evolutionary Computation* (pp. 220–227). IEEE.

Chen, Y., Ohkawa, E., Mabu, S., Shimada, K., & Hirasawa, K. (2009). A portfolio optimization model using genetic network programming with control nodes. *Expert Systems with Applications*, *36*(7), 10735–10745.

Cheng, D., & Cirillo, P. (2018). A reinforced urn process modeling of recovery rates and recovery times. *Journal of Banking & Finance*, *96*, 1–17.

Cockburn, I. M., Henderson, R., & Stern, S. (2018). *The impact of artificial intelligence on innovation* (Technical report). National Bureau of Economic Research.

Cui, W., Brabazon, A., & O'Neill, M. (2010). Evolving efficient limit order strategy using grammatical evolution. In *IEEE Congress on Evolutionary Computation* (pp. 1–6). IEEE.

de Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, *94*, 65–76.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. Buchanan, & A. Bryman (Eds.), *The Sage Handbook of Organizational Research Methods*. Sage Publications.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation. *Journal of Accounting and Economics*, *64*(2–3), 221–245.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*, 5228–5235.

Guo, W., Qiao, Y., & Hou, H. (2006). BP neural network optimized with PSO algorithm and its application in forecasting. In *2006 IEEE International Conference on Information Acquisition* (pp. 617–621). IEEE.

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, *13*(2), 586–632.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hawley, D. D., Johnson, J. D., & Raina, D. (1990). Artificial neural systems: A new tool for financial decision-making. *Financial Analysts Journal*, *46*(6), 63–72.

He, K., Tso, G. K., Zou, Y., & Liu, J. (2018). Crude oil risk forecasting: New evidence from multiscale analysis approach. *Energy Economics*, *76*, 574–583.

Heaton, J., Polson, N. G., & Witte, J. H. (2016). Deep learning in finance. *arXiv*. preprint arXiv:1602.06561.

Holland, J. H. (1992). Genetic algorithms. *Scientific American*, *267*(1), 66–73.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, *64*(6), 2833–2855.

Ito, T., Sakaji, H., Izumi, K., Tsubouchi, K., & Yamashita, T. (2017). Development of sentiment indicators using both unlabeled and labeled posts. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE.

Jang, J.-S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(3), 665–685.

Kamiya, S., Kim, Y. H., & Park, S. (2019). The face of risk: CEO facial masculinity and firm risk. *European Financial Management*, *25*(2), 239–270.

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, *36*(10), 1435–1457.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE.

Kim, S. (2018). Volatility forecasting for low-volatility portfolio selection in the US and the Korean equity markets. *Journal of Experimental & Theoretical Artificial Intelligence*, *30*(1), 71–88.

Kim, S.-S. (1998). Time-delay recurrent neural network for temporal correlations and prediction. *Neurocomputing*, *20*(1–3), 253–263.

Kim, Y. S., & Sohn, S. Y. (2004). Managing loan customers using misclassification patterns of credit scoring model. *Expert Systems with Applications*, *26*(4), 567–573.

Kolupaieva, I., Pustovhar, S., Suprun, O., & Shevchenko, O. (2019). Diagnostics of systemic risk impact on the enterprise capacity for financial risk neutralization: The case of Ukrainian metallurgical enterprises. *Oeconomia Copernicana*, *10*(3), 471–491.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70.

Le, T., Vo, B., Fujita, H., Nguyen, N. -T., & Baik, S. W. (2019). A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting. *Information Sciences*, *494*, 294–310.

Lee, J. (2020). Access to finance for artificial intelligence regulation in the financial services industry. *European Business Organization Law Review*, *21*(4), 731–757.

Lee, M.-C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, *36*(8), 10896–10904.

Lei, Y., & Qiu, X. (2020). Evaluating the investment climate for Chinaas cross-border e-commerce: The application of back propagation neural network. *Information*, *11*(11), 526.

Li, J., Meng, Q., Yang, Y., Mabu, S., Wang, Y., & Hirasawa, K. (2010). Trading rules on stock markets using genetic network programming with subroutines. In *Proceedings of SICE Annual Conference 2010* (pp. 3084–3088). IEEE.

Liao, S.-H., & Ho, S.-H. (2010). Investment project valuation based on the fuzzy real options approach. In *2010 International Conference on Technologies and Applications of Artificial Intelligence* (pp. 94–101). IEEE.

Liu, H. (2008). Study on flaws and improvement of discounted cash flow theory in mergers and acquisitions. In *2008 4th IEEE International Conference on Management of Innovation and Technology* (pp. 1337–1341). IEEE.

Lo, A. W. (1989). *Long-term memory in stock market prices* (Technical report), National Bureau of Economic Research.

Long, J., Chen, Z., He, W., Wu, T., & Ren, J. (2020). An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. *Applied Soft Computing*, *91*, 106205.

Lui, A., & Lamb, G. W. (2018). Artificial intelligence and augmented intelligence collaboration: Regaining trust and confidence in the financial sector. *Information & Communications Technology Law*, *27*(3), 267–283.

Mabu, S., Hirasawa, K., & Hu, J. (2007). A graph-based evolutionary algorithm: Genetic network programming (GNP) and its extension using reinforcement learning. *Evolutionary Computation*, *15*(3), 369–398.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3), 93–118.

McCluskey, J., & Liu, J. (2017). US financial market forecasting using data classification with features from global markets. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)* (pp. 965–969). IEEE.

Misra, P. K., & Goswami, K. (2015). Predictability of sugar futures: Evidence from the Indian commodity market. *Agricultural Finance Review*, *75*(4), 552–564.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314–1324.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Naranjo, R., Meco, A., Arroyo, J., & Santos, M. (2015). An intelligent trading system with fuzzy rules and fuzzy capital management. *International Journal of Intelligent Systems*, *30*(8), 963–983.

Nian, K., Coleman, T. F., & Li, Y. (2018). Learning minimum variance discrete hedging directly from the market. *Quantitative Finance*, *18*(7), 1115–1128.

Pai, G. V. (2018). Active portfolio rebalancing using multi-objective metaheuristics. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1845–1852). IEEE.

Pan, W.-T., Shao, Y.-M., Yang, T.-T., Luo, S.-H., & Li, X.-W. (2017). Prediction of mutual fund net value using backpropagation neural network. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (pp. 198–201). IEEE.

Piepenbrink, A., & Gaur, A. S. (2017). Topic models as a novel approach to identify themes in content analysis, *Academy of Management Proceedings* (Vol. 1, p. 11335). Academy of Management.

Piepenbrink, A., & Nurmammadov, E. (2015). Topics in the literature of transition economies and emerging markets. *Scientometrics*, *102*(3), 2107–2130.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.
Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.
Rahamneh, Z., Reyalat, M., Sheta, A., & Aljahdali, S. (2010). Forecasting stock exchange using soft computing techniques. In *ACS/IEEE International Conference on Computer Systems and Applications-AICCSA 2010* (pp. 1–5). IEEE.
Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market.*Journal of Banking & Finance*, *84*, 25–40.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.
Rout, M., Koudjonou, K. M., & Satapathy, S. C. (2020). Analysis of net asset value prediction using low complexity neural network with various expansion techniques. *Evolutionary Intelligence*, *14*, 643–655.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.
Senoguchi, J., & Kurahashi, S. (2013). Prediction of financial crises using statistic model and intelligent technologies in ubiquitous environments. *International Journal of Computer Applications in Technology*, *48*(2), 173–183.
Sezer, O. B., & Ozbayoglu, A. M. (2018). Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing*, *70*, 525–538.
Shivanna, A., & Agrawal, D. P. (2020). Prediction of defaulters using machine learning on Azure ML. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0320–0325). IEEE.
Sigrist, F., & Hirnschall, C. (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking & Finance*, *102*, 177–192.
Singer, Y. (1997). Switching portfolios. *International Journal of Neural Systems*, *8*(04), 445–455.
Sirignano, J. A. (2019). Deep learning for limit order books. *Quantitative Finance*, *19*(4), 549–570.
Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285*(1), 181–203.
Soto, J., Castillo, O., Melin, P., & Pedrycz, W. (2019). A new approach to multiple time series prediction using MIMO fuzzy aggregation models with modular neural networks. *International Journal of Fuzzy Systems*, *21*(5), 1629–1648.
Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*(5), 926–957.
Sun, Y., Liu, X., Chen, G., Hao, Y., & Zhang, Z. J. (2019). How mood affects the stock market: Empirical evidence from microblogs. *Information & Management*, *57*(5), 103181.
Takeda, A., & Kanamori, T. (2014). Using financial risk measures for analyzing generalization performance of machine learning models. *Neural Networks*, *57*, 29–38.
Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311.
Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479.
Turkson, R. E., Baagyere, E. Y., & Wenya, G. E. (2016). A machine learning approach for predicting bank credit worthiness. In *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)* (pp. 1–7). IEEE.
Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking & Finance*, *22*(10–11), 1421–1439.
Wang, F., Yan, X., & Zheng, L. (2020). Time-series and cross-sectional momentum in anomaly returns. *European Financial Management*.
Wang, K., & Huang, S. (2010). Using fast adaptive neural network classifier for mutual fund performance evaluation. *Expert Systems with Applications*, *37*(8), 6007–6011.
Wang, L., & Zhu, J. (2010). Financial market forecasting using a two-step kernel learning method for the support vector regression. *Annals of Operations Research*, *174*(1), 103–120.

Wang, Z.-J., & Zhao, L.-T. (2021). The impact of the global stock and energy market on EU ETS: A structural equation modelling approach. *Journal of Cleaner Production*, *289*, 125140.

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, *112*: 258–273.

Wong, B. K., & Selvi, Y. (1998). Neural network applications in finance: A review and analysis of literature (1990–1996). *Information & Management*, *34*(3), 129–139.

Xu, J., Liu, J., & Zhao, H. (2011). Financial forecasting: Comparative performance of volatility models in Chinese stock markets. In *2011 Fourth International Joint Conference on Computational Sciences and Optimization* (pp. 1220–1225). IEEE.

Yang, C.-H., Liao, M.-Y., Chen, P.-L., Huang, M.-T., Huang, C.-W., Huang, J.-S., & Chung, J.-B. (2009). Constructing financial distress prediction model using group method of data handling technique. In *2009 International Conference on Machine Learning and Cybernetics*, (Vol. 5, pp. 2897–2902). IEEE.

Yao, J., & Tan, C. L. (2000). A case study on using neural networks to perform technical forecasting of forex. *Neurocomputing*, *34*(1–4), 79–98.

Zadeh, L. A. (1988). Fuzzy logic. *Computer*, *21*(4), 83–93.

Zhang, R., Yuan, Z., & Shao, X. (2018). A new combined CNN-RNN model for sector stock price analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 546–551). IEEE.

Zhang, Z., Shen, Y., Zhang, G., Song, Y., & Zhu, Y. (2017). Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman neural network. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 225–228). IEEE.

How to cite this article: Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. *European Financial Management*, 28, 744–770. https://doi.org/10.1111/eufm.12326