Machine Learning and Finance: A Review using Latent Dirichlet Allocation Technique (LDA)

Ahmed Sameer El Khatib
Professor, Deptartament of Accounting, Centro Universitário FECAP,
São Paulo, Brasil.
ORCID: https://orcid.org/0000-0002-0764-8622
Email: ahmed.khatib@fecap.br

Abstract

The aim of this paper is provide a first comprehensive structuring of the literature applying machine learning to finance. We use a probabilistic topic modelling approach to make sense of this diverse body of research spanning across the disciplines of finance, economics, computer sciences, and decision sciences. Through the topic modelling approach, a Latent Dirichlet Allocation Technique (LDA), we can extract the 14 coherent research topics that are the focus of the 6,148 academic articles during the years 1990-2019 analysed. We first describe and structure these topics, and then further show how the topic focus has evolved over the last two decades. 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 modelling of topics for deep comprehension of a body of literature, especially when that literature has diverse multi-disciplinary actors.

Keywords: Machine Learning; topic modelling; structuring finance research; latent dirichlet allocation

1. Introduction

The techniques of Machine Learning (ML) and Artificial Intelligence (AI) offer significant benefits to financial decision makers in terms of new approaches to modelling and forecasting from data. This has been recognized by the finance industry, with about $41 billion annually expected to be spent globally by the finance firms on artificial intelligence technologies by 2022 (IDC, 2018; Sezer & Ozbayoglu, 2018; Zetzsche, Dirk; Arner, Buckley & Tang, 2020). Key current finance artificial intelligence applications are in algorithmic trading, risk management, and process automation. However, research on these topics among finance researchers has lagged practice. In this paper we therefore comprehensively structure this body of literature for the benefit of researchers seeking to understand the techniques and areas of interest of machine learning and artificial intelligence in finance.

For common understanding we first briefly give our working definitions of artificial intelligence techniques. Artificial intelligence (AI) is an umbrella term for a range of techniques involving intelligence demonstrated by machines, with this intelligence normally focused on prediction. In a finance context, due to its numerical focus, the most relevant AI methods have been machine learning (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 relationships within the input data. DL is an approach that has evolved from an earlier focus on artificial neural networks. Throughout this paper we generally refer to ML and DL under a combined term of ML, as DL is a sub-field of ML. 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 modelling, 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.

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, several 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, while 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 to- wards deep learning techniques and other advanced ML techniques. These recent applications include: the understanding of default recovery rates (Cheng & Cirillo, 2018); learning optimal option hedging rates (Nian, Coleman & Li, 2018); modelling investor sentiment (Renault, 2017); and the detection of stock price evolution based on order books (Zetzsche, Dirk; Arner, Buckley & Tang, 2020).

As finance journals have made tentative steps to acknowledging the potential of the new techniques of ML in finance, other disciplines have been making more stri- dent 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 in finance journals. Some recent examples of this outside-finance corpus of ML in finance 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, Lu, Wang, Megahed & Martinez, 2018; Aziz, Dowling, Hammami & Piepenbrink, 2019), and bank risk modelling (Cerchiello, Giudici & Nicola, 2017).

This multi-disciplinary 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 6,148 studies in ML and finance - it is beyond reasonable human capabilities to comprehensively understand the directions and topics of this research. For this reason, we adopt a topic modelling approach that enables probabilistic learning of the core topics of ML and finance across the main disciplines in which it is researched. The topic modelling approach also allows following the evolution of these topics over time. This thus 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 & Jordan, 2003; Blei, 2012; Aziz, Dowling, Hammami & Piepenbrink, 2019) to topic modelling. LDA works off the assumption that each document is a collection of (hidden) topics and the words used by authors represent those topics. Probabilistic inference is used to identify these topics from the collection of words in a document, and subsequently to identify the most relevant 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. Take a simple example where, from a sufficiently large corpus of documents, the extracted set of terms consists of just three words across the entire corpus: [river, bank, money]. LDA should be able distinguish this into two distinct topics of [river, bank] and [bank, money] recognizing the dual meaning of [bank]. The approach, which is detailed in the following section, is therefore recommended for its ability to arrive at distinct topics.

Previous topic modelling in other business disciplines has recently appeared in the literature, including on how emerging markets are researched (Piepenbrink and & 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 under- standing online brand discussions (Tirunillai & Tellis, 2014; Aziz et al., 2019). There has been a limited amount of topic modelling applications of closer relevance to finance, but not quite in finance. These include Moro et al. (2015) who grouped a small number of articles (219) 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. Our study makes several contributions. It is the first comprehensive structuring of the ML and finance literature. There have been some more limited prior reviews of this literature, including Heaton et al. (2016) who structure the deep learning and finance prior research, and Wong and Selvi (1998) who provide an early review of neural network applications in finance. There have also been some comprehensive reviews on ML as applied to economics (Mullainathan & Spiess, 2017; Athey, 2018). 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 modelling to the finance literature. This is important as the technique offers benefits to financial research comprehension. Financial 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 modelling is an important means of addressing this spread problem. To finance researchers, we also contribute to an overall sense as the application of ML to finance is of clear business importance but is relatively understudied by financial researchers. Instead, it is the primary domain of computer scientists. This imbalance is not a problem but does suggest that pertinent financial insights to the research might be being missed. By providing a structure to the topics of this research field we allow finance researchers to become familiar with the areas' structure as well as the approaches to research within each topic.

2. Background

2.1. Topic modelling through Latent Dirichlet Allocation (LDA)

Topic modelling is a process for discovering the 'hidden' or latent topics in a corpus of documents. There are a range of feasible approaches for this, including Latent Semantic Indexing, Probabilistic Latent Semantic Analysis, and Latent Dirichlet Al- location. The most widely accepted current approach is that of Latent Dirichlet Allocation (LDA) developed in Blei et al. (2003), 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 umixture of topics, and terms can 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 modelling is that it assumes a fixed number of hidden topics in a corpus of documents, uses word co-occurrence 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 will reflect this topic mixture. Thus, an article applying neural network approaches to credit risk modelling in banks might be written with a topic mixture of 50% credit risk modelling, 30% neural network methodologies, and 20% banking context. The choice of words in a document will then reflect this mixture of topics and their relative emphasis. The key task for the topic modelling researcher is therefore to work backwards from the observed words to uncover the latent topics.

A necessary first step in topic modelling 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). Common 'stop words' that have offered 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 common stem terms, e.g., 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 more relatively 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 not be a useful term for understanding sub-topics within that area. In a research context this will also normally involve removing research structural words such as results, hypothesis, and analysis, which while not common in a general word usage frequency are very common in a body of academic research studies.

The final dataset after the TF-IDF stage, referred to as the document-term matrix (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, from a corpus of 5,000 documents with 25,000 terms remaining after cleaning will result in a 5,000 * 25,000 dimension DTM. Topic modelling 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 5,000 * 20 matrix with values being the probabilistic weighting of each topic within each of the 5,000 documents. In turn the 20 topics will be a probabilistic weighting of each of the 25,000 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 ten most important terms to describe a topic aiming for these ten terms to explain about 80-90% of the occurrence of that topic. 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 1, 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.

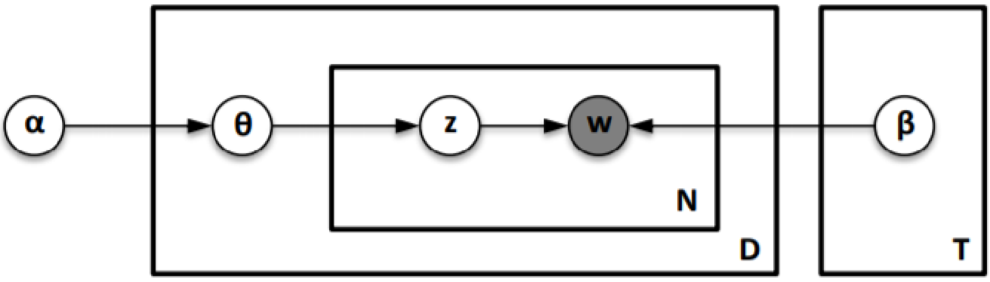

Figure 1: Latent Dirichlet Allocation plate diagram.

Sources: Kaplan and Vakili (2015); Aziz, Dowling, Hammami and Piepenbrink (2019)

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 Dirich- let distribution is a distribution that can be used to produce probability vectors1 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, while 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 involves Bayesian updating to these priors based on actual word distribution across topics and documents.

Referring to Figure 1 again, α informs the initial θ which is the proportion of topics per document. The initial proportions are semi-randomly 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 semi-randomly 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 semi- random allocation of topics to words and documents to a converged determination of probabilistic topics per document and across the corpus.

We now proceed, in the next section, to describe our corpus of ML and finance articles, our pre-processing choices undertaken to generate the DTM, and the LDA modelling options applied to the topic extraction.

3. Method

We use the Elsevier Scopus database to identify relevant prior research on ML in finance based on term searching within title, abstract, and keywords. This was a process of experimentation in terms of identifying suitable search terms where we had to iterate towards the best search terms. Our search criteria required a match of both a machine learning and a finance term, but in initial searches we were retrieving significant intruder articles with the combined terms of, for example, neural network and bank, where the resulting articles were about the application of neural networks within image banks for improved image recognition. An advantage of the topic modelling approach is that the presence of intruder articles is clear from the topics generated, as these articles clustered in topics that are clearly not related to ML in finance. Our final search criteria, which minimized, but didn't fully eliminate, intruder articles were as follows in Table 1:

Table 1. Search criteria

Machine learning search terms (4)	1) machine learning 2) artificial intelligence 3) support vector 4) deep learning 5) neural network
Finance search terms (14)	1) finance 2) investment 3) stock market 4) stock market 5) investor 6) equity 7) commercial bank 8) investment bank 9) credit institution 10) fixed income 11) debt 12) financial derivatives 13) financial crisis 14) risk management
Date Range	1990-2019
Publication type (2)	1) articles 2) conference papers
Source type (2)	1) journals 2) conference proceedings
Language	English language only
Subjects (5)	1) computer science 2) business 3) management and accounting 4) decision sciences 5) economics, econometrics and finance

After determining good term combinations, we experimented with some other variations of search terms for both ML and finance and did not find significant article count variation, suggesting a reasonable convergence on a suitable body of identified research. The date range start period was a feature of the research available, and the end point is October 2018, when we ran our last search update. We included conference papers published in conference proceedings in our sample as this is the primary mode of research communication in computer science, in contrast to finance practice, due to the need 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 dataset from this search process is 6,148 abstracts of articles which is the starting point for our analysis. 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, and Journal of Banking & Finance. For pre-processing we used the R packages tm and Quantitative Analysis of Textual Data (*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 (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 (low TF-IDF terms). Lastly, as the processing of the vocabulary leads to substantial shortening of the abstracts, so we only keep abstracts with at least five remaining terms. These pre-processing steps lead to reduction of the vocabulary in our DTM from 16,208 initial terms to 3,113 terms. Due to the length restriction on abstracts we keep 5,123 out of 6,148 abstracts. The final DTM therefore has dimensions of 5,123 and 3,113. 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 modelling itself we used the R package topic models (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 R topic models 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 estimated topics for models for 15, 20, 25, 30 and 35 topics. As the resulting models do not necessarily find global extrema, for each k we run ten models with randomly chosen (but fixed across the various k) starting points.

This thus results in 50 sets of topic models (5 k parameters * 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. We are guided by this value in choosing the final topic models to analyze, although not exclusively as we also manually inspect among the lower α topic models to choose which seems to be the most coherent in terms of individual topic composition. For this we inspected the top ten 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 three 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 is 20 topics and random seed 5 with an α of 0.1062.

4. Topics of Machine Learning and Finance

The topics extracted are contained in Table 1. Twenty topics are extracted in the preferred model. Of these we classify 14 as strong machine learning and finance topics, while the remaining six have either low coherency or are best classified as pure machine learning topics, and one is classified as a pure finance topic. In order 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 three categories: (1) investment analysis, (2) asset modelling and forecasting, and (3) risk management. We now proceed to discuss the topics within each category as well as the overall category focus. Our analysis focuses on the top matched articles to each topic to provide an illustration of the topic content.

Table 2. Topics of machine learning in finance

No	Area	Topic Label	Top Ten Words	Cohesion	Top 3 articles and θ %	Fit	Average θ %
1	ML & FIN	Market Sentiment	detect, social, sentiment, event, news, media, fraud, text, document, opinion.	H	Smailovic (2014) – 95 Mukwazvure (2015) – 94 Ito (2018) – 93	H H H	5.61
2	ML & FIN	Neural network financial forecasting	firm,debt,list,rough,attribut,chines,per,ratio,rd,size	L	Lian (2015) – 92 Below (2000) – 82 Al-Qaheri (2008) – 81	L H H	4.61
3	ML & FIN	Forex forecasting	fuzzi,deep,logic,currenc,neurofuzzi,anfi,dimension,liquid,match,dolla	H	Yao (2000) – 95 Parida (2016) – 94 Abraham (2002) – 92	H H H	5.04
4	ML & FIN	Technical analysis	trade,ann,cluster,signal,trader, som,predictor,arima,selforgan,time seri	H	Ng (2011) – 92 Chavarnakul (2008) – 89 Kumar (2009) – 86	H H H	7.32
5	ML & FIN	Investor behaviour	agent,stage,analyt,phase,nn,distanc,big,multiag,cycl,cooper	M	Araujo (2010) – 88 Miglietta (2009) – 86 Moerland (2018) – 86	H H M	4.58
6	ML & FIN	Techniques for financial forecasting	ensembl,wavelet,transform,swarm, speed,particl,pso,	H	Ghasemiyeh (2017) – 87 Yiwen (2000) – 87	H H M	5.24

			propag,converg, oil		Nikolic (2016) – 87		
7	ML & FIN	Risk assessment	project,custom,game,ai,strateg, flow,cash,budget, player,schedul	M	Kumagai (2016) – 96 Yu (2011) – 95 Marmier (2014) – 91	L H H	4.81
8	ML & FIN	Energy forecasting	servic,energi,electr,manufactur, cloud,capac,mainten,grid,safeti, load	H	Huang (2013) – 97 Mishra (2012) – 95 Oprea (2016) – 94	H H H	4.88
9	ML	Environmental resource modeling	dss,region,water,environment,sensor,transport,infrastructur,urban, citi,spatial	H	Assaf (2006) – 89 Coventry (2011) – 88 Tulbure (2016) – 88	H H H	4.49
10	ML	Health diagnostics	health,imag,care,insur,medic,patient,recognit,digit,detect,hospit	H	Li (2012) – 94 Chen (2007) – 93 Dadayan (2007) – 92	H H H	4.14
11	ML	Innovative learning	softwar,univers,student,educ,question,behaviour,collabor,cours,feedback,acquir	H	Crain (2012) – 94 Wagner (2006) – 90 Villaverde (2006) - 87	H H H	4.64
12	ML & FIN	Portfolio optimization	portfolio,asset,alloc,hedg,stochast, svr,constraint,multiobject,bound, optimi	H	Chellaboina (2013) – 93 Shen (2017) – 92 Steiner (1997) – 91	H H H	5.89
13	ML & FIN	Investment decision support	expert,fund,properti,scheme,transact,recommend,profil,avail,privat, mutual	H	Kampouridis (2010) – 87 Tsang (2004) – 86 Jeon (2016) – 79	H H M	4.16
14	ML & FIN	Chaos and financial forecasting	option,layer,recurr,filter,valuat, neuron,onlin,nonparametr,evolv, recognit	H	Kim (1998) – 97 Medeiros (2007) – 85 Niranjan (1997) – 83	H H H	4.59
15	ML &	Operational	secur,identif,quali	H	Goh (2010) – 94	H	4.61

	FIN	risk management	t,modul,attack, ontolog,graph,hierarch,element, node		Zadeh (2006) – 91 Bhattacharya (2007) – 90	M H	
16	FIN	Financial analysis	enterpris,crisi,countri,china, economi,polici,growth,corpor, govern,macroeconom	H	Wang (2011) – 93 Shaaf (1999) – 91 Huang (2011) – 91	H H H	6.07
17	ML & FIN	Credit risk modeling	credit,bank,score, commerci,loan, distress,bankruptci,discrimin,failur, custom	H	Waad (2011) – 94 Abdou (2014) – 92 Kim (2004) – 91	H H H	5.93
18	ML & FIN	Volatility modeling	volatil,svm,fluctuat,equiti,estat,multivari,kernel,sequenc,garch, svm	H	D’Cruz (2005) – 90 Liu (2012) – 87 Kumar (2012) – 86	H H H	5.32
19	ML & FIN	Risk forecasting	bp,logist,suppli,chain,ga,sector, bayesian,net,probabilist,cur	M	Mogre (2016) – 85 Wang (2009) – 82 Xu (2014) – 79	H H H	3.88
20	ML	Robotic processes	emerg,robot,mobil,internet,disast, electron,platform, locat,et,patent	H	Modelski (2010a) – 97 Modelski (2010b) – 97 Modelski (2010c) – 97	H H H	4.16

Notes: Table of topics determined based on extraction of 20 topics from 6,148 articles that match machine learning and finance search terms in Scopus. Grey rows are unused topics either due to low topic cohesion or topic not focusing on both machine learning and finance. Topic label is assigned by the authors. Top ten words are the stemmed words that best describe the topic. Cohesion is authors’ assessment of how well the top ten words describe the proposed topic label. Top three articles are the articles that best match the topic label based on θ. Fit is authors assessment of how well each of the three articles fit within the topic. Average θ is overall coherency of the topics mapped to the articles. See Section 3 for further description on the methodology and explanation of choices made in the topic extraction.

4.1 Investment analysis topics

The first category, investment analysis, is the broadest of the three categories, per- haps reflecting general research in this category. We identify four topics in this category, all of which offer an interesting insight into potential future research directions. The four topics are: (1) market sentiment, (2) technical analysis, (3) investor behaviour, and (4) investment decision support.

The first topic is market sentiment (Topic 1 in Table 1). A popular machine learning technique in this topic is support vector machines (SVM) which are particularly useful for classifying data into groups. This, of course, is a common financial research task. Two of the top papers matching this topic use SVMs as part of their toolkit to classify the sentiment and tone of financial news stories (Ferreira et al., 2014) and news comments (Mukwazvure & Supreethi, 2015), although the latter study doesn't confine itself to financial news comments.

The highest matched article to this topic is Smailović et al. (2014) which shows that public sentiment based on Twitter feeds can be used to predict stock price movements. This has been a popular finance research area since the original Bollen et al. (2011) study. The model that Smailović et al. (2014) apply is interesting in terms of it being 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. This approach moves beyond traditional machine learning 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.

Other interesting studies related to market sentiment include Ito et al. (2017) who propose how to use unlabeled text data from stock message boards. The need to label text data for its sentiment tone is a particular and costly problem with using text. Also, Lumezanu et al. (2012) who model the behaviour of ’propagandists’ on social media. They study an economic event of the US debt-ceiling crisis of 2011, but this could similarly be applied to distinguish between paid promoters of equities on social media and genuine investors.

Technical analysis (Topic 4) is, in contrast to the text-based market sentiment topic, primarily based on quantitative learning. The techniques are neural networks and increasingly deep learning variations of neural networks. Ng et al. (2011) approach the classic candlestick patterns of technical analysis with a combination of a Radial Basis Function (RBF) neural network and an improved error minimization technique. Their method seeks to distinguish between real and false bullish candle- stick patterns. Bullish candlesticks, in their most basic form, are where prices close the day near the daily high price, although Ng et al. (2011) apply ten different variations of bullish candlesticks, including multi-day patterns. The RBF in combination with the error minimization approach are used to train the model to distinguish be- tween bullish candlesticks that subsequently deliver either positive or negative 3-day returns. Testing on 10 years of Hong Kong stock data does show outperformance compared to a regular candlestick approach, although the testing is limited. Other papers applying neural networks to the technical analysis of stock prices include Chavarnakul and Enke (2008) who use a generalized regression neural network to predict from two technical indicators that combine volume and price signals, and Tilakaratne et al. (2007) who use intermarket indicators.

Variations of reinforcement learning for technical analysis are applied by Dempster et al. (2001) and Hussain et al. (2016). Dempster et al. (2001) races reinforcement learning and genetic programming for popular technical indicators used in forex trading (such as relative strength and momentum oscillators). Reinforcement learning proffers the advantage over more basic forms of machine learning by allowing recurrent learning through goal setting. This study finds that genetic programming marginally outperforms in trading. The more recent Hussain et al. (2016) study applies recurrent neural networks, which are a form of reinforcement learning that is classified as deep learning. The study's particular technique is not to showcase the ability of recurrent neural networks to predict financial time series, but rather a new regularization technique developed to allow better organization of the hidden layer in the neural network. This illustrates the extent to which the application of ML to financial phenomena outside of finance research has moved on from a more basic conversation on the merits of ML techniques towards a fine-tuning stage. They show their technique results in significant predictive improvements over alternative ML techniques including multilayer perceptron neural networks in the prediction of a range of financial times series.

The topic of investor behaviour (Topic 5) fits somewhat to the extant literature on theoretical modelling of agent behaviour in finance. The fit for this topic is classified by us in Table 1 as 'medium' indicating that the topic extracted from the LDA is not fully coherent. Moerland et al. (2018) is perhaps the most useful of the top-matched articles to the topic, as it provides a review of prior literature on the role of emotion in agent / machine interaction, as well as delving into how to integrate emotional understanding in reinforcement learning models. With the projected growth of robo- advising, and the issues here around the need to develop an emphatic relationship between the advisee / robo-advisor, some aspects of this article would be useful in that regard. The article also highlights the extent to which AI researchers are fully aware of the need to integrate emotion in future technology. This idea is applied specifically to behavioural finance in Miglietta and Remondino (2009). They address a range of behavioural finance theories, but primarily concentrate on prospect theory. The issue they explore - for a computer science audience - is how to improve agent modelling in simulated behaviour considering these behavioural biases. The article itself is more exploratory than prescriptive indicating that these issues remained open for the intended audience at the time of the articles' publication, thus showing potential for collaboration with behavioural finance researchers. The last article we explore for this topic is Araújo et al. (2010) which is an attempt, based on acknowledging uncertain agent behaviour, to allow model flexibility to account for this uncertainty. They propose a variation on a multilayer perceptron neural network for this purpose.

The final topic in this group is investment decision support (Topic 13). This topic concentrates on the application of AI (in part ML, but also other aspects of AI) to generate information that supports investment decision making. The two highest matched articles to this topic both cover software called EDDIE, standing for Evolutionary Dynamic Data Investment Evaluator (Tsang et al., 2004; Kampouridis & Tsang, 2010). We concentrate on a more recent Kampouridis and Otero (2017) study in the subsequent discussion as it refers to the most recent iteration of the EDDIE algorithm (EDDIE 9). EDDIE builds on a genetic programming approach, so is situated more within general AI rather than ML. The software is designed to support investment decisions by the application of genetic decision trees to arrive at simple yes/no or buy/don't buy answers to questions posed by the user such as whether a stock will rise by X% within n days. Of interest with this software, which is one of many examples of such software, is the extent to which it has grown in user flexibility with each iteration. The most recent version allows significant user customizability, while earlier versions were constrained by the strongly structured abilities of the initial algorithms.

Another application in this topic is Wang and Huang (2010) which addresses the issue of how to rapidly detect changes in mutual fund performance. For this they develop what they term a fast adaptive neural network classifier. The purpose of the 'fast adaptive' part of the neural network classifier is to allow new information to be constantly fed into the model and for it to subsequently update quickly. They show this speed output improvement in their study, comparing their model to a more standard neural network that needs to wait for significantly more information before it can efficiently update. A last interesting application under this topic by Mahalingam and Vivek (2016) is aimed at supporting a much smaller financial decision - that of automating how much a person can save given their spending behaviour. This illustrates some of the range in the investment decision support topic, and suggests some of the clear crossovers towards the FinTech industry.

4.2 Asset modelling and forecasting topics

Topics related to either asset modelling or asset price forecasting are the largest category of topics, with six of the 14 topics in this category. This is quite natural due to the emphasis on forecasting and prediction in machine learning techniques. The six topics are: (1) portfolio optimization, (2) volatility modelling, (3) forex forecasting, (4) energy forecasting, (5) chaos and financial forecasting, and (6) techniques for financial forecasting.

Portfolio optimization (Topic 12) is the first topic we explore in this category. Shen and Wang (2017) are one of the top matched articles to this topic. They address a classic issue with Markowitz-based portfolio selection that mean-variance portfolio estimation in practice requires long time periods of data to estimate input parameters. This, in turn, introduces the risk that data will not be relevant for current portfolios. Researchers such as Michaud (1989) have sought to address this issue by resampling recent pricing data through bootstrapping to avoid the use of older data, but without convincing success (Harvey et al., 2010). Shen and Wang (2017) address this issue using ensemble learning, an ML approach where multiple algorithms / sub-samples are tested to improve learning compared to a single ML application. A known advantage of ensemble learning is that even if individual algorithm applications are weaker than a single-application algorithm, their combined findings can outperform the single-application algorithm. The technique in Shen and Wang (2017) is a Bag of Little Bootstraps (BLB) to bootstrap sub- samples of equity data using short time periods, which when combined improve estimation errors for Markowitz input parameters. Other papers in this topic take alternative approaches to portfolio optimization, with Steiner and Wittkemper (1997) applying an artificial neural network, Ma et al. (2014) using a genetic algorithm, and Chellaboina et al. (2013) applying both these approaches to hedged option portfolios. The next two topics in this category - volatility modelling, forex forecasting - have elements of similarity with their focus on forecasting asset pricing and volatility. We concentrate first on volatility modelling and forecasting (Topic 18). As is evidenced from the keywords for the topic, the ML approach to volatility modelling is to provide alternative methods for estimating volatility compared to standard techniques such as GARCH. Various neural network approaches are the most common ML techniques (Liao & Wang, 2010; Liu & Wang, 2012). Hossain and Nasser (2011) is a useful example of this topic, with their comparison of GARCH, neural networks, and SVM. While SVMs were originally developed for classification, as noted in the previous section, they can also be applied to regression type problems. In Hossain and Nasser (2011) they argue and demonstrate how SVMs applied to volatility forecasting strikes a better balance between training accuracy and testing accuracy than neural networks, and overall outperforms GARCH and neural network approaches.

Foreign exchange (forex) forecasting (Topic 3) is particularly attractive for ML practical application due to the clean structure of the data available. Thus, it has been used as a data source for testing a wide number of ML techniques, while there has also been a ML focus within finance research due to the common industry application of algorithms to trading forex markets. As demonstrated by the keywords for the topic, deep learning based fuzzy neural networks are a core focus of the application of ML to forex markets (Abraham, 2002; AmirAskari & Menhaj, 2016; Kodogiannis & Lolis, 2002; Parida et al., 2015). Yao and Tan (2000) also present a simpler neural networks model without incorporating fuzzy logic. Taking the most recent paper of the papers closely matching the forex topic, AmirAskari and Menhaj (2016), we see some of the benefits of the fuzzy neural network approach in forex forecasting. The paper builds on the core idea of blending fuzzy logic and neural networks through testing a new model called the Modified Fuzzy Relational Model (MRFM) which they argue should allow better modelling of dynamic systems such as forex relationships. The paper shows this (weakly) by comparison with a feed for- ward neural network, an RBF neural network, and an earlier type of fuzzy neural network (ANFIS). It is a notable paper as it highlights how far the literature has moved on from traditional finance techniques, where even the contrast methods to the new method being showcased are considerably advanced compared to a standard finance forex study.

The energy forecasting topic (Topic 8) is only briefly discussed here. The topic itself is very much at the edge between a finance, accounting, and pure energy topic. This itself is interesting as it shows how the methods of ML can blend these topics. The topic concentrates on estimating energy usage and therefore energy costs and how to minimize these costs (Huang et al., 2013; Mishra et al., 2012; Oprea, 2015). While this has some benefits in terms of modelling energy usage and as a contribution towards understanding economic environmental impact, perhaps the topic is also useful at demonstrating how well the LDA topic modelling approach can cluster similar documents together from a large disparate literature.

The last two topics in this category are: chaos and financial forecasting (Topic 14), and techniques for financial forecasting (Topic 6). These two topics have some similarities due to the focus on particularly advanced techniques. In the first, the benefits of ML approaches to understanding chaotic time series are demonstrated. Ince and Trafalis (2008), among the highly matched papers to the topic, discusses the general issue with chaotic financial data and how ML approaches can help improve understanding. Kim (1998) and Rather et al. (2015) focus on recurrent neural networks as a means of modelling. Recurrent neural networks are proposed as useful as they allow some form of memory within the neural network, so there can be temporal connections, even if the overall time series is chaotic. Echoing this, the popular Heaton et al. (2016) paper (not a matched paper in our database as it has not been published) which overviews deep learning techniques in finance places an emphasis on a modern type of recurrent neural network: LSTM (Long-Short Term Memory) models. Other strongly matched techniques for this topic focus on the non-linearity aspect, with the testing of multilayer perceptron's (Medeiros & Barreto, 2007) and RBF neural networks (Niranjan, 1996) to help better model chaotic time series.

In the last topic in the asset modelling and forecasting category, techniques for financial forecasting (Topic 6), a variety of advanced techniques have been clustered together by the topic modelling. This includes varieties of genetic algorithms such as cuckoo searches: where the modelling approach is based on a cuckoo bird randomly laying eggs in a variety of foreign nests initially and then learning the best nests over time. Ghasemiyeh et al. (2017) applies this technique to stock price prediction. Worasucheep (2015) integrate an alternative optimization technique - Artificial

Bee Colony algorithm to model forex pricing, while particle swarm optimization is applied to model electricity pricing in Junyou (2007), and wavelets are applied in a short study by Yiwen et al. (2000).

4.3 Risk management topics

The third and final category of the topics relates to the area of risk management, for which we have four of the 14 topics. The topics are (1) risk assessment, (2) risk forecasting, (3) operational risk management, and (4) credit risk modelling. As noted in the introduction, risk modelling was one of the original focus on ML in finance, due to the strong methods in ML for classification and clustering. Classification is a particularly pertinent task for risk management, due to the broad need to classify acceptable and unacceptable risks.

The first two topics that we explore are risk assessment (Topic 7) and risk fore- casting (Topic 19). These two interconnected topics were a core focus following the global financial crisis, amid the seeming clear failure of traditional risk assessment and forecasting techniques. Our papers, however, cluster particularly around the nar- row field of the application of ML based approaches to the assessment and forecasting of risk related to projects. There is quite a strong overlap with supply chain research in terms of method and theory focus. There is also a focus on simpler techniques such as decision trees, which are used for classification in ML, but have a much wider use. A lot of the focus is instead on understanding the complexity within systems and managing that complexity. For instance, Marmier et al. (2014) introduces an integrated process through a decision tree model in new product development based on project risk. This system takes into consideration several risk activities along with other product development activities as part of project risk assessment that leads to the creation of a range of possible scenarios for a project.

Another top matching paper by Yu (2011) discuss the effectiveness of neural networks and game theory-based models of project risk assessment and management for large-scale long-term and high-investment projects. They model how neural networks can help manage risk in such a project.

While the risk assessment topic is largely theoretical in the scope of the matched articles, the research clustered under the risk forecasting topic (Topic 19) is quite applied. Mogre et al. (2016) proposes a decision support system-based framework to mitigate risk. Their approach encompasses identification, estimation, assessment, and decision making for risks. They apply their modelling approach to a case study of UK offshore wind farms and argue how their risk mitigation modelling strategies can reduce construction delays. Consistent with the Yu (2011) study in the risk assessment topic, the authors also suggest the usefulness of their proposed frameworks in the case of complex projects, as an enhanced level of complexity will usually foster a greater exposure towards risks. In a more direct ML application, Wang and Huang (2009) and Xu et al. (2014) focus on the preciseness and practicability of various neural network-based models in the context of risk assessment and risk forecasting.

Operational risk management (Topic 15) is the third topic of this category and represents perhaps the oldest and the newest threat that both financial and non- financial firms deal with. Operational risk refers to potential losses emanating from a host of operational breakdown linked to either an internal (e.g., inadequate or failed internal processes, people, and systems) or an external event (e.g., fraud, operational error) (Moosa, 2007). A difficulty is the sheer breadth of possible operational events, as opposed to the more easily defined other risk categories such as market and credit risk. A variety of AI and ML based risk management approaches have been applied to operational risk management. This includes a case-based reasoning approach to assess, identify and analyse operational risk (e.g., Goh & Chua, 2009) and an AI based risk management process for enabling an organization to insulate itself from serious system related security breaches (Bhattacharya & Ghosh, 2007).

The last topic identified for this category is credit risk modelling (Topic 17), a sub-area of the risk management category that demonstrates the most practical ap- plication of ML techniques, for financial institutions. There is also some data availability for studying credit risk modelling, an issue that has prevented the growth in study of the other topics in this category. As noted in the introduction, we started witnessing the use of ML techniques in credit risk models 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 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), Kim and Sohn (2004) tests neural network models to identify the misclassification of customers (good and bad borrowers) observed in classical credit scoring models. Also, Abdou et al. (2014) examines the robustness of three credit risk models including discriminant analysis, logistic regression, and multilayer perceptron neural networks, in the decision-making process of the Islamic finance houses in the UK. Their findings show that the neural network approach outperforms the other credit scoring techniques. This success is on grounds of classification accuracy, on predicting the rejected credit applications, and on the cost of misclassifying borrowers.

The rapidly growing complexity of credit markets (such as the growth of CDS markets) and availability of new data (such as on consumer and small firm (SME) lending) has further widened the scope for ML techniques in the assessment and modelling of credit risk. Addressing this new complexity and new data is the focus of the last set of papers we discuss related to credit risk modelling. A good example is Son et al. (2016) who use a sample of daily CDS prices with varying maturities and quality over the period 2001 to 2014 and show that non-parametric ML models with deep learning outperform traditional benchmark models in terms of price prediction accuracy as well as in proposing practical hedging measures. Khandani et al. (2010) introduce an ML technique based on decision trees and SVM that leads to cost savings of up to 25 percent in the case of consumer lending, while Figini et al. (2017) propose a multivariate outlier detection ML technique that improves credit risk estimation for the SME lending portfolio of UniCredit Bank. As many of the studies already reviewed in the last three sections also indicate, they find that an ensemble approach to model building is particularly useful in improving model accuracy.

4.4 Topic evolution over time

As a last analysis, we track the evolution of our 14 topics over time. On a broad analysis of the chart, we see the most popular topics at the beginning of the time period were quite modelling-based. This modelling-based focus was seen in the discussion of the topics of operational risk management, investment decision support, and portfolio optimization. The recent most popular topics are much more data-based and forecasting-based, including market sentiment and technical analysis. 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 2 which showed very few ML and finance papers before about 2007. For that reason, it is probably best to compare 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 large growth in market sentiment studies over this period: Z 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. Technical analysis which has been strong in most periods, has dipped somewhat in popularity, and volatility modelling has also fallen in relative ranking. This might also be reflective of a move away from pure numerical analysis, as both technical analysis and volatility are almost exclusively based on numerical analysis.

This time period also shows large falls in relative popularity of topics for which there is still limited data availability including: investor behaviour, risk assessment, operational risk management, and even credit risk modelling (although that remains one of the more popular topics). This suggests that data availability is currently key in terms of topic popularity. Lastly, we see a strong recent rise in popularity in techniques for financial forecasting, perhaps due to the expansion in possible new techniques following deep learning breakthroughs that have expanded the universe of possible techniques that can be developed. The rise in energy forecasting is interesting and is possibly because of an increased global research focus on optimizing and reducing energy consumption, as well as the increased grid complexity introduced by new forms of renewable energy production.

5 DISCUSSION AND CONCLUSIONS

In this study, we have provided a first finance application of topic modelling - a probabilistic technique for matching similar documents. 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 6,148 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 modelling technique can cleanly structure this diverse corpus of research into coherent topics.

Apart from the introduction of topic modelling to finance, 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 innate need within finance to continuously improve forecasting and modelling techniques. To some extent this is an existential issue for finance research, as industry practice in finance moves far beyond the techniques of traditional finance. Our mapping, provided in summary form in this paper, but with the full mapping for individual papers to topics available in a provided online repository, shows the structure of this literature for researchers in these topics seeking to augment their current research with the techniques of ML.

We identify three overall categories of topics: investment analysis, asset modelling and forecasting, and risk management, as well as 14 topics spread across these topics. These three categories, while sharing some techniques, are progressing at different speeds. Risk management, thanks to an early research lead, is well advanced in machine learning application of the modelling of risk. This is seen in the practical field of credit risk modelling to support bank and other financial institution lending decisions. However, the area is currently suffering from the rise in importance of data, with most relevant data being held privately by firms for their own internal use. Asset modelling and forecasting is also advanced due to the early understanding of the suitability of neural networks for modelling financial times series. Also, of help in this category has been the use of the clean datasets of financial time series as sample applications of machine learning techniques by non-finance machine learning researchers. Less advanced, but possibly more promising, is the category of investment analysis. In this category lies the promise of new data generation, such as on investor sentiment, behaviour, and the forecasting of fundamental finance variables. 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 robot-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 machine learning, 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, like the problem with ML in risk management research. 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 datasets available for understanding the impact of networks on financial behaviour. 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 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. This is, of course, just a minor snapshot of what finance research has practically produced for the finance industry. Machine learning, as demonstrated in this research, is the logical next step of the finance researcher toolkit, due to its strong emphasis on practical understanding. We hope that the structure provided in this study offers some guidance as to how to incorporate these techniques suitably in future research.

REFERENCES

Abdou, H. A., Alam, S. T.; Mulkeen, J. (2014). Would credit scoring work for islamic finance? A neural network approach. *International Journal of Islamic and Middle Eastern Finance and Management*, 7(1):112–125.

Abraham, A. (2002). Analysis of hybrid soft and hard computing techniques for forex monitoring systems. In 2002 IEEE *World Congress on Computational Intelligence*, volume 2, pages 1616–1621. IEEE.

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.

AmirAskari, M.; Menhaj, M. B. (2016). A modified fuzzy relational model ap- proach to prediction of foreign exchange rates. In 2016 4th *International Conference on Control, Instrumentation, and Automation (ICCIA)*, pages 457–461. IEEE.

Araújo, R. d. A., de Oliveira, A. L.; Soares, S. C. (2010). A quantum-inspired hybrid methodology for financial time series prediction. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.

Athey, S. (2018). The impact of machine learning on economics. In Ajay K. Agrawal, J. G. and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda. University of Chicago Press.*

Aziz, S., Michael D., Helmi H.; A. Piepenbrink. Machine learning in finance: A topic modellng approach. In: *1st International Banking and Finance Research Conference*, Agadir, Morocco, October 2019.

Bhattacharya, S.; Ghosh, S. (2007). An artificial intelligence-based approach for risk management using attack graph. In 2007 International Conference on Computational Intelligence and Security, pages 794–798. IEEE.

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

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.

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

Chavarnakul, T.; Enke, D. (2008). Intelligent technical analysis based equiv- olume charting for stock trading using neural networks. *Expert Systems with Applications,* 34(2):1004–1017.

Chellaboina, V., Bhatia, A.; Bhat, S. P. (2013). Explicit formulas for optimal hedging stratergies for European contingent claims. In 2013 IEEE *Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 122–127. IEEE.

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

Dempster, M. A., Payne, T. W., Romahi, Y.; Thompson, G. W. (2001). Computational learning techniques for intraday FX trading using popular technical indicators. IEEE *Transactions on neural networks*, 12(4):744–754.

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.

Ferreira, J. Z., Rodrigues, J., Cristo, M.; de Oliveira, D. F. (2014). Multi-entity polarity analysis in financial documents. In Proceedings of the 20th *Brazilian Symposium on Multimedia and the Web*, pages 115–122. ACM.

Figini, S., Bonelli, F.; Giovannini, E. (2017). Solvency prediction for small and medium enterprises in banking. *Decision Support Systems*, 102:91–97.

Ghasemiyeh, R., Moghdani, R.; Sana, S. S. (2017). A hybrid artificial neural network with metaheuristic algorithms for predicting stock price. Cybernetics and Systems, 48(4):365–392.

Goh, Y. M.; Chua, D. (2009). Case-based reasoning approach to construction safety hazard identification: Adaptation and utilization. *Journal of Construction Engineering and Management*, 136(2):170–178.

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

Harvey, C. R., Liechty, J. C., Liechty, M. W.; Müller, P. (2010). Portfolio selection with higher moments. Quantitative Finance, 10(5):469–485.

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.

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

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

Hossain, A.; Nasser, M. (2011). Comparison of the finite mixture of ARMA- GARCH, back propagation neural networks and support-vector machines in fore- casting financial returns. Journal of Applied Statistics, 38(3):533–551.

Huang, D., Thottan, M.; Feather, F. (2013). Designing customized energy services based on disaggregation of heating usage. In 2013 IEEE PES Innovative Smart Grid Technologies (ISGT), pages 1–6. IEEE.

Hussain, A. J., Al-Jumeily, D., Al-Askar, H.; Radi, N. (2016). Regularized dynamic self-organized neural network inspired by the immune algorithm for financial time series prediction. Neurocomputing, 188:23–30.
Ince, H.; Trafalis, T. B. (2008). Short term forecasting with support vector ma- chines and application to stock price prediction. International Journal of General Systems, 37(6):677–687.

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), pages 1–8. IEEE.

Junyou, B. (2007). Stock price forecasting using PSO-trained neural networks. In IEEE Congress on Evolutionary Computation, pages 2879–2885. IEEE.

Kampouridis, M.; Otero, F. E. (2017). Heuristic procedures for improving the predictability of a genetic programming financial forecasting algorithm. Soft Computing, 21(2):295–310.

Kampouridis, M. and Tsang, E. (2010). EDDIE for investment opportunities forecasting: Extending the search space of the GP. In 2010 IEEE Congress on Evolutionary Computation (CEC), pages 1–8. IEEE.

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

Khandani, A. E., Kim, A. J.; Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11):2767–2787.

Kim, 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.

Kodogiannis, V.; Lolis, A. (2002). Forecasting financial time series using neural network and fuzzy system-based techniques. Neural Computing & Applications, 11(2):90–102.

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

Liao, Z.; Wang, J. (2010). Forecasting model of global stock index by stochastic time effective neural network. Expert Systems with Applications, 37(1):834–841.

Liu, F.; Wang, J. (2012). Fluctuation prediction of stock market index by Legendre neural network with random time strength function. Neurocomputing, 83:12– 21.

Lumezanu, C., Feamster, N.; Klein, H. (2012). # bias: Measuring the tweeting behavior of propagandists. In Sixth International AAAI Conference on Weblogs and Social Media.

Ma, Y., Gong, X.; Tian, G. (2014). A mean-semi-variance portfolio optimization model with full transaction costs. In 2014 International Conference on Computational Intelligence and Communication Networks (CICN), pages 623–627. IEEE.

Mahalingam, P.; Vivek, S. (2016). Predicting financial savings decisions using sigmoid function and information gain ratio. Procedia Computer Science, 93:19–25.

Marmier, F., Ioana, F. D., and Didier, G. (2014). Strategic decision-making in NPD projects according to risk: Application to satellites design projects. Computers in Industry, 65(8):1107 – 1114.

Medeiros, C. M.; Barreto, G. A. (2007). Pruning the multilayer perceptron through the correlation of backpropagated errors. In Seventh International Conference on Intelligent Systems Design and Applications, pages 64–69. IEEE.

Michaud, R. O. (1989). The Markowitz optimization enigma: Is ‘optimized’ optimal? Financial Analysts Journal, 45(1):31–42.

Miglietta, N.; Remondino, M. (2009). Modelling cognitive distortions of behavioural finance. In International Conference on Computational Intelligence, Modelling and Simulation, 2009., pages 204–209. IEEE.

Mishra, A., Irwin, D., Shenoy, P., Kurose, J.; Zhu, T. (2012). Smartcharge: Cutting the electricity bill in smart homes with energy storage. In Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet, page 29. ACM.

Moerland, T. M., Broekens, J., and Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: A survey. Machine Learning, 107(2):443–480.

Mogre, R., D’Amico, F., et al. (2016). A decision framework to mitigate supply chain risks: An application in the offshore-wind industry. IEEE Transactions on Engineering Management, 63(3):316–325.

Moosa, I. A. (2007). Operational Risk Management. Springer.

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.

Mukwazvure, A.; Supreethi, K. (2015). A hybrid approach to sentiment analysis of news comments. In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), pages 1–6. IEEE.

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

Ng, W. W., Liang, X.-L., Chan, P. P.; Yeung, D. S. (2011). Stock investment decision support for Hong Kong market using RBFNN based candlestick models. In 2011 International Conference on Machine Learning and Cybernetics (ICMLC), volume 2, pages 538–543. IEEE.

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

Niranjan, M. (1996). Sequential tracking in pricing financial options using model based and neural network approaches. In M.C. Mozer, M.I. Jordan, T. P., editor, Advances in Neural Information Processing Systems, pages 960–966. Cambridge: MIT Press.

Oprea, S. (2015). Informatics solutions for electricity consumption optimization. In 2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), pages 193–198. IEEE.

Parida, A., Bisoi, R., Dash, P., and Mishra, S. (2015). Financial time series prediction using a hybrid functional link fuzzy neural network trained by adaptive unscented kalman filter. In 2015 IEEE Power, Communication and Information Technology Conference (PCITC), pages 568–575. IEEE.

Piepenbrink, A.; Gaur, A. S. (2017). Topic models as a novel approach to identify themes in content analysis. In Academy of Management Proceedings, volume 2017, page 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.

Rather, A. M., Agarwal, A.; Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications, 42(6):3234–3241.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. Journal of Banking & Finance, 84:25–40.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5):513–523.

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.

Shen, W.; Wang, J. (2017). Portfolio selection via subset resampling. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pages 1517–1523.

Smailović, J., Grčar, M., Lavrač, N.; Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. Information Sciences, 285:181–203.

Son, Y., Byun, H.; Lee, J. (2016). Nonparametric machine learning models for predicting the credit default swaps: An empirical study. Expert Systems with Applications, 58:210–220.

Steiner, M.; Wittkemper, H.-G. (1997). Portfolio optimization with a neural network implementation of the coherent market hypothesis. European Journal of Operational Research, 100(1):27–40.

Suganuma, M., Shirakawa, S., and Nagao, T. (2017). A genetic programming approach to designing convolutional neural network architectures. In Proceedings of the Genetic and Evolutionary Computation Conference, pages 497–504. ACM.

Tilakaratne, C. D., Mammadov, M. A.; Morris, S. A. (2007). Effectiveness of using quantified intermarket influence for predicting trading signals of stock markets. In Proceedings of the sixth Australasian conference on Data mining and analytics, pages 171–179. Australian Computer Society.

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.

Tsang, E., Yung, P.; Li, J. (2004). EDDIE-Automation, a decision support tool for financial forecasting. Decision Support Systems, 37(4):559–565.

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. Journal of Banking & Finance, 22(10-11):1421–1439.

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, Y.; Huang, L. (2009). Risk assessment of supply chain based on BP neural network. In KAM'09. Second International Symposium on Knowledge Acquisition and Modelling, 2009, volume 2, pages 186–188. IEEE.

Weng, B., Lu, L., Wang, X., Megahed, F. M., and 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.

Worasucheep, C. (2015). Forecasting currency exchange rates with an Artificial Bee Colony-optimized neural network. In 2015 IEEE Congress on Evolutionary Computation (CEC), pages 3319–3326. IEEE.

Xu, W., Zhang, Z., Gong, D.; Guan, X. (2014). Neural network model for the risk prediction in cold chain logistics. International Journal of Multimedia and Ubiquitous Engineering, 9(8):111–124.

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.

Yiwen, Y., Guizhong, L.; Zongping, Z. (2000). Stock market trend prediction based on neural networks, multiresolution analysis and dynamical reconstruction. In Proceedings of the IEEE/IAFE/INFORMS 2000 Conference on Computational Intelligence for Financial Engineering, pages 155–156. IEEE.

Yu, Y. (2011). Risk management game method of the weapons project based on bp neural network. In 2011 International Conference on Information Technology, Computer Engineering and Management Sciences (ICM), volume 1, pages 113–117. IEEE.

Zetzsche, Dirk Andrea; Arner, Douglas W. and Buckley, Ross P. and Tang, Brian, Artificial Intelligence in Finance: Putting the Human in the Loop (February 1, 2020). CFTE Academic Paper Series: Centre for Finance, Technology and Entrepreneurship, no. 1., University of Hong Kong Faculty of Law Research Paper No. 2020/006, Available at SSRN: https://ssrn.com/abstract=3531711 .

Copyright Disclaimer