Gupta *et al. Financ Innov* (2020) 6:39
https://doi.org/10.1186/s40854-020-00205-1

3/10

Financial Innovation

REVIEW Open Access

Comprehensive review of text-mining applications in finance

Aaryan Gupta[1], Vinya Dengre[1], Hamza Abubakar Kheruwala[1] and Manan Shah[2*]

*Correspondence:
manan.shah@spt.pdpu.ac.in
[2] Department of Chemical Engineering, School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, Gujarat 382007, India
Full list of author information is available at the end of the article

- pretty bad writing
- not very comprehensive
- some interesting references ...

Abstract

Text-mining technologies have substantially affected financial industries. As the data in every sector of finance have grown immensely, text mining has emerged as an important field of research in the domain of finance. Therefore, reviewing the recent literature on text-mining applications in finance can be useful for identifying areas for further research. This paper focuses on the text-mining literature related to financial forecasting, banking, and corporate finance. It also analyses the existing literature on text mining in financial applications and provides a summary of some recent studies. Finally, the paper briefly discusses various text-mining methods being applied in the financial domain, the challenges faced in these applications, and the future scope of text mining in finance.

Keywords: Text mining, Machine learning, Financial forecasting, Sentiment analysis, Text classification, Corporate finance

Introduction

Today, technology is deeply integrated with everyone's lives. Nearly every activity in modern life, from phone calls to satellites sent into space, has evolved exponentially with technology (Patel et al. 2020a, b, c; Panchiwala and Shah 2020). The increasing ability to create and manage information has been an influential factor in the development of technology. According to the National Security Agency of the United States, 1826 petabytes on average are handled daily over the Internet (Hariri et al. 2019; Jaseena and David 2014). With the rapid increase in data and information communicated over the Internet, it has become necessary to regulate and ease the flow of the same (Ahir et al. 2020; Gandhi et al. 2020). A number of commercial and social applications have been introduced for these purposes. Aspects of data and information, such as security, research, and sentiment analysis, can be of great help to organisations, governments, and the public (Jani et al. 2019; Jha et al. 2019). There are various optimized techniques that aid us in tasks such as classification, summarisation, and ease of access and management of data, among others (Shah et al. 2020a, b; Talaviya et al. 2020). Algorithms related to machine learning and deep learning (DL) are just some of the many algorithms that can be used to process the available information (Kakkad et al. 2019; Kundalia et al. 2020). Even though there is a massive amount of available information, the use of computational techniques

can help us process information from top to bottom and analyse entire documents as well as individual words (Pandya et al. 2019; Parekh et al. 2020).

Human-generated 'natural' data in the form of text, audio, video, and so on are rapidly increasing (Shah et al. 2020a, b). This has led to a rise in interest in methods and tools that can help extract useful information automatically from enormous amounts of unstructured data (Jaseena and David 2014; David and Balakrishnan 2011). One crucial method is text mining, which is a combined derivative of techniques such as data mining, machine learning, and computational linguistics, among others. Text mining aims to extract information and patterns from textual data (Talib et al. 2016b; Fan et al. 2006). The trivial approach to text mining is manual, in which a human reads the text and searches for useful information in it. A more logical approach is automatic, which mines text in an efficient way in terms of speed and cost (Herranz et al. 2018; Sukhadia et al. 2020; Pathan et al. 2020).

According to the India Brand Equity Foundation (IBEF 2019), the Indian financial industry alone had US $340.48 billion in assets under management as of February 2019. This value only provides us with a limited indication of the actual size and reach of the global finance industry. Technology has paved the way for digitalisation in this rapidly growing behemoth. 'FinTech' is a developing domain in the finance industry, which has been defined as a union of finance and information technology (Zavolokina et al. 2016). Marrara et al. (2019) examined how FinTech relates to Italian small and medium-sized enterprises (SMEs), where FinTech has witnessed huge growth in terms of investment and development, and how it has proved fruitful for the SME market in a short amount of time. FinTech has popularised the use of data in the financial industry. This data is substantially in the form of structured or unstructured text. Therefore, traditionally and technically, textual data can be regarded as always having been a prevailing and essential element in the finance sector.

Unstructured textual data have been increasing rapidly in the finance industry (Lewis and Young 2019). This is where text mining has a lot of potential. Kumar and Ravi (2016) explored various applications in the financial domain in which text mining could play a significant role. They concluded that it had numerous applications in this industry, such as various kinds of predictions, customer relationship management, and cybersecurity issues, among others. Many novel methods have been proposed for analysing financial results in recent years, and artificial intelligence has made it possible to analyse and even predict financial outcomes based on historical data.

Finance has been an important force in human life since the earliest civilisations. It is noteworthy that from barter systems to cryptocurrencies, finance has always been associated with data, such as transactions, accounts, prices, and reports. Manual approaches to processing data have been reduced in use and significance over time. Researchers and practitioners have come to prefer digitised and automated approaches for studying and analysing financial data. Financial data contain a significant amount of latent information. If the latent information were to be extracted manually from a huge corpus of data, it might take years. Advancements in text mining have made it possible to efficiently examine textual data pertaining to finance. Bach et al. (2019) published a literature review on text mining for big-data analysis in finance. They structured the review in terms of three critical questions. These questions pertained to the intellectual core

of finance, the text-mining techniques used in finance, and the data sources of financial sectors. Kumar and Ravi (2016) discussed the model presented by Vu et al. (2012) that implemented text mining on Twitter messages to perform sentiment analysis for the prediction of stock prices. They also mentioned the model of Lavrenko et al. (2000), which could classify news stories in a way that could help identify which of them affected trends in finance and to what degree. We will further discuss text-mining applications in finance in subsequent sections.

Apart from finance, we present a brief overview of text mining in other industries. On social media, people generate text data in the form of posts, blogs, and web forum activity, among many others (Agichtein et al. 2008). Despite the vast quantity of data available, the relatively low proportion of content of significant quality is still a problem (Kinsella et al. 2011), which is an issue that can be solved by text mining (Salloum et al. 2017). In the biomedical field too, there is a need for effective text-mining and classification methods (Krallinger et al. 2011). On e-commerce websites, text mining is used to prevent the repetition of information to the same audience (Da-sheng et al. 2009) and improve product listings through reviews (Kang and Park 2016; Ur-Rahman and Harding 2012). In healthcare, researchers have worked on applications such as the identification of healthcare topics directly from personal messages over the Internet (Lu 2013), classification of online data (Srivastava et al. 2018), and analysis of patient feedback (James et al. 2017). The agriculture industry has also used text mining in, for example, the classification of agricultural regulations (Espejo-Garcia et al. 2018), ontology-based agricultural text clustering (Su et al. 2012), and analysis of agricultural network public opinions (Lee 2019). Text mining has also been utilised in the detection of malicious web URLs which evolve over time and have complex features (Li et al. 2020a; b, c).

This paper discusses the use of text mining in the financial domain in detail, taking into consideration three major areas of application: financial forecasting, banking, and corporate finance. We also discuss the widely used methodologies and techniques for text mining in finance, the challenges faced by researchers, and the future scope for text-mining methods in finance.

Overview of text-mining methodologies

Text mining is a process through which the user derives high-quality information from a given piece of text. Text mining has seen a significant increase in demand over the last few years. Coupled with big data analytics, the field of text mining is evolving continuously. Finance is one major sector that can benefit from these techniques; the analysis of large volumes of financial data is both a need and an advantage for corporates, government, and the general public. This section discusses some important and widely used techniques in the analysis of textual data in the context of finance.

Sentiment analysis (SA)

One of the most important techniques in the field is SA. It has applications in numerous sectors. This technique extracts the underlying opinions within textual data and is therefore also referred to as opinion mining (Akaichi et al. 2013). It is of prime use in a number of domains, such as e-commerce platforms, blogs, online social media, and microblogs. The motives behind sentiment analysis can be broadly divided into emotion

recognition and polarity detection. Emotion detection is focused on the extraction of a set of emotion labels, and polarity detection is more of a classifier-oriented approach with discrete outputs (e.g., positive and negative) (Cambria 2016).

There are two main approaches for SA, namely lexicon-based (dictionary-based) and machine learning (ML). The latter is further classified into supervised and unsupervised learning approaches (Xu et al. 2019; Pradhan et al. 2016). Lexicon-based approaches use SentiWordNet word maps, whereas ML considers SA as a classification problem and uses established techniques for it. In lexicon-based approaches, the overall score for sentiment is calculated by dividing the sentiment frequency by the sum of positive and negative sentiments. In ML approaches, the major techniques that are used are Naïve Bayes (NB) classifier and support vector machines (SVMs), which use labelled data for classification. SA using ML has an edge over the lexicon approach, as it doesn't require word dictionaries that are highly costly. However, ML requires domain-specific datasets, which can be considered as a limitation (Al-Natour and Turetken 2020). After data preprocessing, feature selection is performed as per the requirement, following which one obtains the final results after the analysis of the given data as per the adopted approach (Hassonah et al. 2019).

In the financial domain, stock market prediction is one of the applications in which SA has been used to predict future stock market trends and prices from the analysis of financial news articles. Joshi et al. (2016) compared three ML algorithms and observed that random forest (RF) and SVMs performed better than NB. Renault (2019) used StockTwits (a platform where people share ideas about the stock market) as a data source and applied five algorithms, namely NB, a maximum entropy method, a linear SVM, an RF, and a multilayer perceptron and concluded that the maximum entropy and linear SVM methods gave the best results. Over the years, researchers have combined deep learning methods with traditional machine learning techniques (e.g., construction of sentiment lexicon), thus obtaining more promising results (Yang et al. 2020).

Information extraction

Information extraction (IE) is used to extract predefined data types from a text document. IE systems mainly aim for object identification by extracting relevant information from the fragments and then putting all the extracted pieces in a framework. Post extraction, DiscoTEX (Discovery from TextEXtraction) is one of the core methods used to convert the structured data into meaningful data to discover knowledge from it (Salloum et al. 2018).

In finance, named-entity recognition (NER) is used for extracting predefined types of data from a document. In banking, transaction order documents of customers may come via fax, which results in very diverse documents because of the lack of a fixed template and creates the need for proper feature extraction to obtain a structured document (Emekligil et al. 2016).

Natural language processing (NLP)

NLP is a part of the artificial intelligence domain and attempts to help transform imprecise and ambiguous messages into unambiguous and precise messages. In the financial sector, it has been used to assess a firm's current and future performance, domain

standards, and regulations. It is often used to mine documents to obtain insights for developing conclusions (Fisher et al. 2016). NLP can help perform various analyses, such as NER, which further helps in identifying the relationships and other information to identify the key concept. However, NLP lacks a dictionary list for all the named entities used for identification (Talib et al. 2016a; b).

As NLP is a pragmatic research approach to analyse the huge amount of available data, Xing et al. (2017) applied it to bridge the gap between NLP and financial forecasting by considering topics that would interest both the research fields. Figure 1 provides an intuitive grasp of natural language-based financial forecasting (NLFF).

Chen et al. (2020) discussed the role of NLP in FinTech in the past, present, and future. They reviewed three aspects, namely know your customer (KYC), know your product (KYP), and satisfy your customer (SYC). In KYC, a lot of textual data is generated in the process of acquiring information about customers (corporate sector and retail). With respect to KYP, salespersons are required to know all the attributes of their product, which again requires data in order to know the prospects, risks, and opportunities of the product. In SYC, salespersons/traders and researchers try to make the financial activities more efficient to satisfy the customers in the business-to-customer as well as customer-to-customer business models. Herranz et al. (2018) discussed the role of NLP in teaching finance and reported that it enhanced the transfer of knowledge within an environment overloaded with information.

Text classification

Text classification is a four-step process comprising feature extraction, dimension reduction, classifier selection, and evaluation. Feature extraction can be done with common techniques such as term frequency and Word2Vec; then, dimensionality reduction is performed using techniques such as principal component analysis and linear discriminant analysis. Choosing a classifier is an important step, and it has been observed that deep learning approaches have surpassed the results of other machine learning algorithms. The evaluation step helps in understanding the performance of the model; it is conducting using various parameters, such as the Matthews correlation coefficient

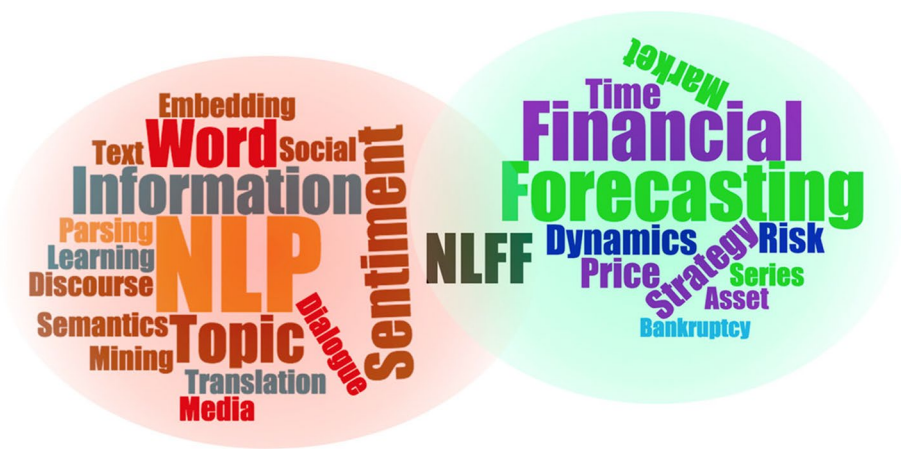

Fig. 1 An intersection of NLP and financial forecasting to illustrate the concept of NLFF (Xing et al. 2017)

(MCC), area under the ROC curve (AUC), and accuracy. Accuracy is the simplest of these to evaluate. Figure 2 shows an overview of the text classification process (Kowsari et al. 2019).

Brindha et al. (2016) compared the performance of various text classification techniques, namely NB, k-nearest neighbour (KNN), SVM, decision tree, and regression, and found that based on the precision, recall, and F1 measures, SVM provided better results than the others.

Deep learning

Deep learning is a part of machine learning, which trains a data model to make predictions about new data. Deep learning has a layered architecture, where the input data goes into the lowest level and the output data is generated at the highest level. The input is transformed at the various middle levels by applying algorithms to extract features, transform features into factors, and then input the factors into the deeper layer again to obtain transformed features (Heaton et al. 2016). Widiastuti (2018) focused on the input data, as it plays an important role in the performance of any algorithm. The author concluded that modification of the network architecture with deep learning algorithms can markedly affect performance and provide good results.

In finance, deep learning solves the problem of complexity and ambiguity of natural language. Kraus and Feuerriegel (2017) used a corpus of 13,135 German ad hoc announcements in English to predict stock market movements and concluded that deep learning was better than the traditional bag-of-words approach. The results also showed that the long short-term memory models outperformed all the existing machine learning algorithms when transfer learning was performed to pre-train word embeddings.

Review of text-mining applications in finance

As mentioned in earlier sections, this paper focuses on the applications of text mining in three sectors of finance, namely financial predictions, banking, and corporate finance. In the subsections, we review various studies. Some literature has been summarised in detail, and in the end, a tabular summary of some more studies is included. Figure 3 shows a summarised link between the text-mining techniques and their corresponding applications in the respective domains. Although the following subsections discuss the studies pertaining to each sector individually, there has also been research on techniques that can be applied to multiple financial sectors. One such system was proposed by Li et al. (2020a), which was a classifier based on adaptive hyper-spheres. It could be helpful in tasks such as credit scoring, stock price prediction, and anti-fraud analysis.

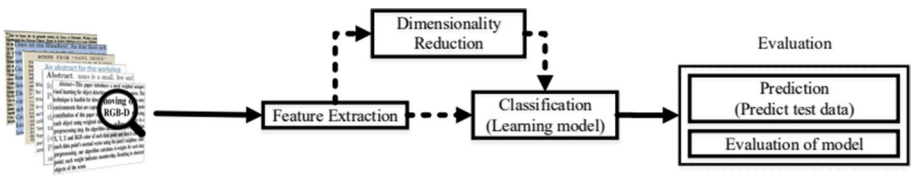

Fig. 2 A general overview of the text classification process (Kowsari et al. 2019)

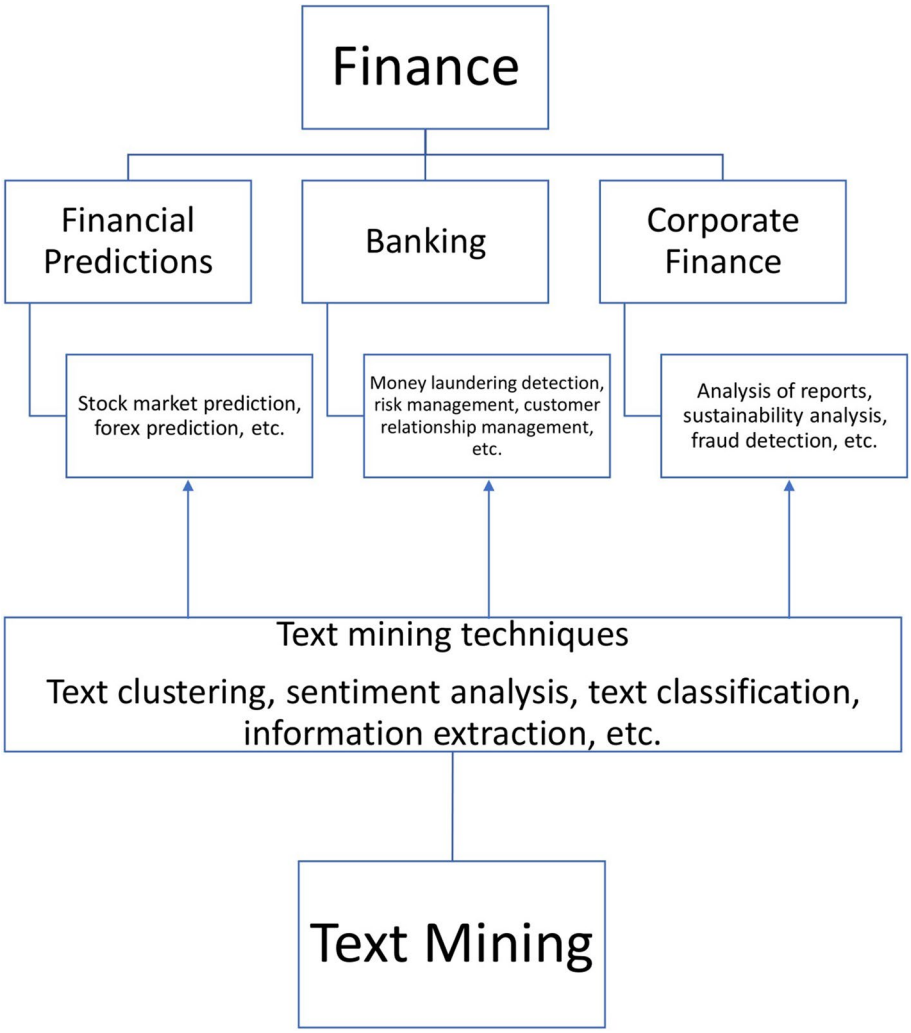

Fig. 3 An overview of how text mining can be used in the financial domain. This paper follows a systematic approach for reviewing text-mining applications, as depicted by the flowchart in the figure. The two independent entities, namely finance and text mining, are linked together to show the possible applications of various text-mining techniques in various financial domains

Prediction of financial trends

Using the ever-expanding pool of textual data to improve the dynamics of the market has long been a practice in the financial industry. The increasing volume of press releases, financial data, and related news articles have been motivating continued and sophisticated analysis, dating back to the 1980s, in order to derive a competitive advantage (Xing et al. 2017). Abundant data investigated with text mining can deliver an advantage in a variety of scenarios. As per Tkáč and Verner (2016) and Schneider and Gupta (2016), among the many ideas covered in financial forecasting, from credit scoring to inflation rate prediction, a large proportion of focus is on stock market and forex prediction. Wen et al. (2019) proposed an idea regarding how retail investor attention can be used for evaluation of the stock price crash risk.

Wu et al. (2012) proposed a model that combined the features of technical analysis of stocks with sentiment analysis, as stock prices also depend on the decisions of investors who read stock news articles. They focused on obtaining the overall sentiment behind each news article and assigned it the respective sentiment based on the weight it carried. Next, using different indicators, such as price, direction, and volume, technical analysis was performed and the learning prediction model was generated. The model was used to predict Taiwan's stock market, and the results proved to be more promising than models that employed either of the two. This indicates an efficient system that can be integrated with even better features in the future.

Al-Rubaiee et al. (2015) analysed the relationship between Saudi Twitter posts and the country's stock market (Tadawul). They used a number of algorithms such as SVM, KNN, and NB algorithms to classify Arabic text for the purpose of stock trading. Their major focus was on properly preprocessing data before the analysis. By comparing the results, they found that SVM had the best recall, and KNN had the best precision. The one-to-one model that they built showcased the positive and negative sentiments as well as the closing values of the Tadawul All Share Index (TASI). The relationship between a rise in the TASI index and an increase in positive sentiments was found out to be greater than that of a decline in the index and negative sentiments. The researchers mentioned that in future work they would incorporate the Saudi stock market closing values and sentiment features on tweets to explore the patterns between the Saudi stock index and public opinion on Twitter.

Vijayan and Potey (2016) proposed a model based on recent news headlines that predicted the forex trends based on the given market situations. The information about the past forex currency pair trends was analysed along with the news headlines corresponding to that timeline, and it was assumed that the market would behave in the future as it had done in the past. The researchers focused on the elimination of redundancy, and their model focused on news headlines rather on entire articles. Multilayer dimension reduction algorithms were used for text mining, the Synchronous Targeted Label Prediction algorithm was used for optimal feature reduction, and the J48 algorithm was used for the generation of decision trees. The main focus was on fundamental analysis that targeted unstructured textual data in addition to technical analysis to make predictions based on historical data. The J48 algorithm resulted in an improvement in the accuracy and performance of the overall system, better efficiency, and less runtime. In fact, the researchers reported that the algorithm could be applied to diverse subjects, such as movie reviews.

Nassirtoussi et al. (2015) proposed an approach for forex prediction wherein the major focus was on strengthening text-mining aspects that had not been focused upon in previous studies. Dimensionality reduction, semantic integration, and sentiment analysis enabled efficient results. The system predicted the directional movement of a currency pair based on news headlines in the sector from a few hours before. Again, headlines were taken into consideration for the analysis, and a multilayer algorithm was used to address semantics, sentiments, and dimensionality reduction. This model's process was highly accurate, with results of up to 83%. The strong results obtained in that study demonstrate that the studied relationships exist. The models can be applied to other contexts as well.

Nikfarjam et al. (2010) discussed the components that constitute a forecasting model in this sector and the prototypes that had been recently introduced. The main components were compared with each other. Feature selection and feature weighting were used to select a piece of news and assign weights to them, used either individually or in combination for feature selection. Next, feature weighting was used to calculate the weights for the given terms. The feature weighting methodology was based on the study by Fung et al. (2002), who had assigned more weights to enhance the term frequency-inverse document frequency (TF-IDF) weighting. For text classification, most researchers have applied SVMs to classify the input text into either good or bad news. Some researchers have used Bayesian classifiers, and some others have used a combination of binary classifiers to achieve the final classification decision. Many authors have focused on news features but not equally addressed the available market data. The focus of most studies has been on the analysis of news and indicator values separately, which has proved to be less efficient. The combination of both market news and the status of market trends at the same time is expected to provide stronger results.

Gupta et al. (2019) proposed a combination of two models: the primary model obtained the dataset for prediction, preprocessed the dataset using logistic regression to remove redundancy, and employed a genetic algorithm, KNN, and support vector regression (SVR). In a comparison of all three, KNN was the basis for their predictions, with an efficiency of more than 50%. The genetic algorithm was used next in search for better accuracy. In an attempt to further support the genetic algorithm, SVR was used, which gave the opening price for any day in the future. For sentiment analysis, Twitter was used, as it was considered the most popular source for related news. The model divided the tweets into two categories, and the rise or fall of the market was predicted taking into consideration the huge pool of keywords. In the end, the model had an accuracy of about 70–75%, which seems reasonable for a dynamic environment.

Nguyen et al. (2015) focused on sentiment analysis of social media. They obtained the sentiments behind specific topics of discussion of the company on social media and achieved promising results in comparison with the accuracy of stocks in the preceding year. Sentiments annotated by humans on social media with regards to stock prediction were analysed, and the percentage of desired sentiments was calculated for each class. For a remaining lot of messages without explicit sentiments, a classification model was trained using the annotated sentiments on the dataset. For both of these tasks, an SVM was used as the classification model. In another study, after lemmatisation by CoreNLP, latent Dirichlet allocation (LDA) (Blei et al. 2003) was used as the generative probabilistic model. The authors also implemented the JST model (Lin and He 2009) and Aspect-based Sentiment Analysis for analysing topic sentiments for stock prediction. The study's limitation was that the topics and models were selected beforehand. The accuracy was around 54%; however, the overall prediction in the model passed only if the stock went up or down. As the model just focused on sentiments and historical prices, the authors intended to add more factors to build a more accurate model.

Li et al. (2009) approached financial risk analysis through the available financial data on sentiments and used machine learning and sentiment analysis. The uniqueness of their study was the volume of data and the information sentiments. A generalised autoregressive conditional heteroskedasticity modelling (GARCH)-based artificial

neural network and a GARCH-based SVM were used. A special training process, named the 'dynamic training technique', was applied because the data was non-stationary and noisy and could have resulted in overfitting. For analysing news, the semantic orientation-based approach was adopted, mainly because of the number of articles that were analysed in the study. The future work on this model was expected to include more input data and better sentiment analysis algorithms to obtain better results.

The use of sentiment analysis as a tool to facilitate investment and risk decisions by stock investors was demonstrated by Wu et al. (2014). Sina Finance, an experimental platform, was the basis for the collection of financial data for this model. The method incorporated machine learning based on SVM and GARCH with sentiment analysis. At the specific opening and closing times for each day, the GARCH-based SVM was used to identify the relations between the obtained information's sentiment and stock price volatility. This model showed better results when predicting individual stocks rather than at the industry level. The machine learning approach was about 6% more accurate than the lexicon-based semantic approach, and it performed better with bigger datasets. The model performed better on datasets relating to small companies, as small companies were observed to be more sensitive to online reviews. The authors mentioned their future scope as expanding their dataset and attempting to create a more efficient sentiment calculation algorithm to increase the overall accuracy, similar to the one made by Li et al. (2009).

A slightly different approach was used by Ahmad et al. (2006), who focused on sentiment analysis of financial news streams in multiple languages. Three widely spoken languages, namely Arabic, Chinese, and English, were used for replication for automatic sentiment analysis. The authors adopted a local grammar approach using a local archive of the three languages. A statistical criterion in the training collection of texts helped in the identification of keywords. The most widely available corpus was for English, followed by Chinese and Arabic. Based on the frequencies of various words, the most widely utilised words were ranked and selected. Through manual evaluation, the accuracy of extraction ranged from 60 to 75%. A more robust evaluation of this model would be necessary for use in real-time markets, with the inclusion of more than one news vendor at a time.

Over the years, deep learning has become acknowledged as a useful machine learning technique that enables state-of-the-art results. It uses multiple layers to create representations and features from the input data. Text-mining analysis has also continuously evolved. The early basic model used lexicon-based analysis to account for a particular entity (sentiment analysis). Considering the complexity of language, a complete understanding of what any piece of text aims to convey requires a more complex analysis to identify and target relevant entities and related aspects (Dohaiha et al. 2018). The most important aspect is the relationship between the words in the text, and how the same is dominant in determining the meaning of the content. Several language elements, such as implications (Ray and Chakrabarti 2019) and sarcasm, require high-level methods for handling. This problem requires the use of deep learning models that can help completely understand a given piece of text. Deep learning may incorporate time series analysis and aspect-based sentiment analysis, which enhances data mining, feature selection, and fast information retrieval. Deep learning models learn features during the process

of learning. They create abstract representations of the given data and therefore are unchanged with local changes to the input data (Sohangir et al. 2018). Word embeddings target words that are similar in context. By the measurement of similarities between words (e.g., cosine similarity in the case of vectors), one can employ word embeddings in the initial data preprocessing layers for faster and more efficient NLP execution (Young et al. 2018).

The huge amount of streaming financial news and articles are impossible to be processed by humans for interpretation and application on a daily basis. In a number of uses, such as portfolio construction, forecasting a financial time series is essential. The application of DL techniques on such data for forecasting purposes is of interest to industry professionals. It has been reported that repeated patterns of price movements can be estimated using econometric and statistical models (Souma et al. 2019). Even though the market is dynamic, a combination of deep learning models and past market trends is very useful for accurate predictions. In a comparison of real trades with the generated market trades with the use of SA, Kordonis et al. (2016) found a considerable effect of sentiments on the predictions. Because of the promising results, the use of artificial intelligence and deep learning has attracted the interests of many researchers and practitioners to improve forecasting.

With the use of deep learning, one has to perform little work by hand, while being able to harness a large amount of computation and data. DL techniques that use distributed representation are considered state-of-the-art methods for a large variety of NLP problems. We expect these models to improve and get better at handling unlabelled data through the development and use of approaches such as reinforcement learning.

Owing to the advancements in technology, there are several factors that can be used in models that aim to predict market movements. Not only the price models but also a number of different related models include macroeconomic variables (e.g., investment). Although macroeconomic indicators are important, they tend to be updated infrequently. Unlike such economic factors, public mood and sentiments (Xing et al. 2018a, b) are dynamic and can be instantaneously monitored. For instance, behavioural science researchers have found that the stock market is affected by the investors' psychology (Daniel et al. 2001). Depending on their mood states, investors make numerous decisions, a big proportion of which are risky. The impact of sentiment and attention measures on stock market volatility (Audrino et al. 2018) can be gauged through news articles, social media, and search engine results. The models that incorporate technical indicators of the market with sentiments obtained from the aforementioned sources outperform those that rely on only one of the two (Li et al. 2009). In a study pertaining to optimal portfolio allocation, Malandri et al. (2018) used historical data of the New York Stock Exchange and combined it with sentiment data to get comparatively better returns for the portfolios taken under consideration.

Empirical studies have shown that current market prices are a reflection of recently published news; this has been clearly shown by the Efficient Market Hypothesis (Fama 1991). Rather than being dependent on the existing information, price changes are markedly affected by new information or news. ML and DL methods have allowed data scientists to play a part in financial sector analysis and prediction (Picasso et al. 2019). There has been an increasing use of text-mining methods to make trading decisions (Wu

et al. 2012). Different kinds of models, including neural networks, are used for sentiment embeddings from news, tweets, and financial blogs. Mudinas et al. (2019) studied the change of Granger-caused stocks based on sentiments alone—although this did not provide promising results, the integration with prediction models gave better results. This is because sentiments cannot be determinant factors alone, but they can be used with prediction models to lead to better and dynamic results.

As discussed above, a plethora of proposals and approaches in relation to financial forecasting have been studied, the two main applications of which have been stock prediction and forex. The main focus of these studies was on obtaining sentiments from news headlines and not from entire articles. Researchers have used a variety of text-mining approaches to integrate the abundant amount of useful information with financial patterns. Table 1 summarises some more research studies that have been conducted in recent years on the subject of text mining in financial predictions.

Banking and related applications

Banking is one of the largest and fastest-growing industries in this era of globalisation. The industry is heading towards adopting the most efficient practices for each of its departments. The total lending in the financial year 2017–2018 increased from US $429.92 billion to $1347.18 billion at a CAGR of 10.94% (Ministry of Commerce and Industry, Government of India, 2019). This huge rise is promoting strong economic growth, increasing incomes, enhancing trouble-free access to bank credit, and increasing consumerism. In the midst of an IT revolution, competitive reasons have led to the rising importance and adoption of banking automation. IT enables the implementation of various techniques for risk controls and smooth flow of transactions over electronic mediums and supports financial product innovation and development.

Gao and Ye (2007) proposed a framework for preventing money laundering with the help of the transaction histories of customers. They did this by identifying suspicious data from various textual reports from law enforcement agencies. They also mined unstructured databases and text documents for knowledge discovery in order to automatically extract the profiles of the entities that could be involved in money laundering. They employed SVM, decision trees, and Bayesian inference to develop a hierarchical structure of the suspicious reports and regression to identify hidden patterns.

Bholat et al. (2015) analysed the utility of text mining in central banks (CB), as a wide range of data sources is required for evaluating monetary and financial stability and for achieving policy objectives. Therefore, text-mining techniques are more powerful than manual means. The authors elucidated two major approaches: the use of text as data for research purposes in CB, and the various text-mining techniques for this purpose. For the former, they suggested that textual data in the form of social narratives can be used by central banks as financial indicators for risk and uncertainty management by employing topic clustering on the narratives. The latter aspect involved preprocessing of data to de-duplicate it, convert it into text files, and reduce it into tokens by various tokenisation techniques. Thereafter, text-mining techniques, such as dictionary techniques, vector space models, latent semantic analysis, LDA, and NB algorithm, were applied to the tokenised data. The authors concluded that aggregately, these can be a very useful addition to the efficient functioning of the CB.

Table 1 Summary of recent research studies of text-mining applications for financial predictions

Study	Datasets	Techniques/algorithms used	Evaluation parameters	Performance (on the basis of evaluation parameters)	References
Stock index forecasting	German ad hoc announcements and stock indices (DAX, CDAX, and STOXX)	Lasso, ridge regression, elastic net, gradient boosting, random forest	Root mean squared error (RMSE)	DAX: 409.662 CDAX: 35.273 STOXX: 12.170	Feuerriegel and Gordon (2018)
Stock prices prediction	Apple's stock prices, July 2017	SVR	Comparison of kernels	RBF kernel performed best	Shah et al. (2018b)
AZFinText system for stock market prediction	Financial news articles	SVR	Closeness, directional accuracy, simulated trading engine	Closeness: 0.04261 Directional accuracy: 57.1% Simulated trading: 2.06% return	Schumaker and Chen (2009)
Financial time series forecasting	Security companies' quarterly and annual reports	ARIMA, SVR	Mean absolute error (MAE), mean absolute per cent error (MAPE), RMSE	MAE: 1.07 MAPE: 23.06 RMSE: 1.34	Wang et al. (2012)
Stock price prediction	Online Flemish newspaper articles	SVM	Accuracy, AUC, return rate, Sharpe ratio	Varied results across the defined metrics for different techniques; suggestion for a new evaluation parameter	Junqué de Fortuny et al. (2014)
Stock price prediction by automated news reading	Corporate announcements, stock price data	SVM, SVR	Accuracy, Squared correlation coefficient (R^2)	Accuracy: 76.3% R^2: 20.2%	Hagenau et al. (2013)
Forex intraday trend prediction	News headlines, currency pair rates	J48 classifier, synchronous target feature reduction	Accuracy	80%	Vijayan and Potey (2016)
Mining critical indicators of stock market movements	Financial news articles	RF	Accuracy	98.34% ???	Elagamy et al. (2018)
Using news sentiments for stock price prediction	News articles related to the Nifty Pharma Index	Dictionary-based sentiment analysis model	Directional accuracy	70.59%	Shah et al. (2018a, b)
Exchange rates prediction	News data, social media information by Market Psych	Multivariate linear regression, multilayer perceptron	Directional accuracy	60.26%	Yusuuf and Shihabeldeen (2019)

Bach et al. (2019) stated that a huge amount of unstructured data from various sources has created a requirement for the extraction of keywords in the banking sector. They mentioned four different procedures for the extraction of keywords, which were obtained from the study by Bharti and Babu (2017). Bach et al. further discussed how keyword extraction can be implemented to extract related useful comments and documents and to compare the banking institutions as well. They also reviewed some other text-mining techniques that can be utilised by banks. NER was used on large datasets for the extraction of entities such as a person, location, and organisation. Sentiment analysis was done to analyse customer opinions, which is crucial for a bank's functioning. Topic extraction was found to be useful mainly in credit banking. Social network analysis, a graph theory-based methodology to study the social media user structure, provided an outlook on how the customers are connected on the social media and how impactful they were in sharing information to the network of interests. This social network analysis could then be coupled with text mining to identify the keywords which correspond to the customers' common interest.

Yap et al. (2011) discussed the issue faced by recreational clubs with respect to potential defaulters and non-defaulters. They proposed a credit scoring model that utilised text mining for estimating the financial obligations of credit applicants. A scorecard was built with the help of past performance reports of the borrowers wherein different clubs used different criteria for evaluating the historic data. The data was split into a 70:30 ratio for training and validating, respectively. They used three different models, namely a credit scorecard model, logistic regression model, and decision tree model, with an accuracy rate of 72.0%, 71.9%, and 71.2% respectively. Although the model benefitted the club administration, it also had a few limitations, such as poor quality of the scorecard and biased samples used to evaluate new applicants, as the model was built on historic data.

Xiong et al. (2013) devised a model for personal bankruptcy prediction using sequence mining techniques. The sequence results showed good prediction ability. This model has potential value in many industries. For clustering categorical sequences, a model-based k-means algorithm was designed. A comparative study of three models, namely SVM, credit scoring, and the one proposed by them, found that the accuracies were 89.3%, 80.54%, and 94.07% respectively. The sequence mining used in the proposed model outperformed the other two models. In terms of loss prediction, the KNN algorithm had the potential to identify bad accounts with promising predictive ability.

Bhattacharyya et al. (2011) explored the use of text mining in credit card fraud detection by evaluating two predictive models: one based on SVM, and the other based on a combination of random forest with logistic regression. They discussed various challenges and problems in the implementation of the models. They recommended that the models should always be kept updated to account for the growing malpractices. The original dataset used in the study comprised more than 50 million real-time credit card transactions. The dataset was split into multiple datasets as per the requirements of different techniques. Because of imbalanced data, the performance was not solely measured by the overall accuracy but also by sensitivity, specificity, and area under the curve. Although the random forest model showed the highest overall accuracy of 96.2%, the study provided some other noteworthy observations. The accuracy of each model varied

according to the proportion of the fraudulent cases, with all of them having more than 99% accuracy for a dataset with 2% fraud rates. The authors concluded with suggestions for future exploration: modifying the models to make them more accurate and devising a more reliable approach to split datasets into training and testing sets.

Kou et al. (2014) used data regarding credit approval and bankruptcy risk from credit card applications to analyse financial risks using clustering algorithms. They made evaluations based on 11 performance measures using multicriteria decision-making (MCDM) methods. A previous study by Kou et al. (2012) had proposed these MCDM methods for the evaluation of classification algorithms. In a later study (Kou et al. 2019), they employed these methods for assessing the feature selection methods for text classification.

In addition to the above-discussed literature in this section, Table 2 provides a summary of some more studies related to the banking finance industry. As visible in Table 2, banking has a lot of different text-mining applications. Risk assessment, quality assessment, money laundering detection, and customer relationship management are just a few examples from the wide pool of possible text-mining applications in banking.

Applications in corporate finance

Corporate finance is an important aspect of the financial domain because it integrates a company's functioning with its financial structure. Various corporate documents such as the annual reports of a company have a lot of hidden financial context. Text-mining techniques can be employed to extract this hidden information and also to predict the company's future financial sustainability.

Guo et al. (2016) implemented text-mining algorithms that are widely used in accounting and finance. They merged the Thomson Reuters News Archive database and the News Analytics database. The former provides original news, and the latter provides sentiment scores ranging from −1 to 1 with positive, negative, and neutral scores. To balance the dataset, 3000 news articles were randomly selected for training and 500 for testing. Three algorithms, namely NB, SVM, and neural network, were run on the dataset. The overall output accuracies were 58.7%, 78.2%, and 79.6%, respectively. With the neural network having the highest accuracy, it was concluded that it can be used for text mining-based finance studies. Another model based on semantic analysis was also implemented, which used LDA. LDA was used to extract document relationships and the most relevant information from the documents. According to the authors, in accounting and finance, this technique has proven to be advantageous for examining analyst reports and financial reporting.

Lewis and Young (2019) discussed the importance of text mining in financial reports. They preferred NLP methods. They highlighted the exploding growth of unstructured textual data in corporate reporting, which opens numerous possibilities for financial applications. According to the authors, NLP methods for text mining provide solutions for two significant problems. One, they prevent overload through automated procedures to deal with immense amounts of data. Two, unlike human cognition, they are able to identify the underlying important latent features. The authors reviewed the widely used methodologies for financial reporting. These include keyword searches and word counts, attribute dictionaries, NB classification,

Table 2 Summary of recent research studies on some text-mining applications for banking

Study	Datasets	Techniques/algorithms used	Evaluation parameters	Performance (on the basis of evaluation parameters)	References
Topic extraction for Italian banks	Data from Twitter API	Topic modelling, LDA	Graphical representations	–	Krstić et al. (2019)
Bank risk report quality assessment	Annual risk reports of German banks	Regressions	MAE, RMSE	MAE: 0.0338 RMSE: 0.0436	Fritz and Tows (2018)
Anti-money laundering reduction	Money laundering cases from the Internal Revenue Service	SAS visual text analytics	Defining use cases for text analytics	–	Cook and Herron (2018)
Analysis of central bank documents	Bank of Italy documents	Vector space model, sentiment analysis	Automated readability index (ARI), Formality score	ARI: 12–15, Formality score: 70–75	Bruno (2016)
Bank service quality assessment	Client reviews from www.banki.ru	Rule-based classifier	Precision, recall, F-measure	Average F-measure: 0.86	Bidulya and Brunova (2016)
Bank failure prediction	Bank 10-K reports	SVM	Accuracy, precision, recall, F1 score	Accuracy: 84.34% Precision: 10.81 Recall: 40 F1: 17.02	Gupta et al. (2016)
Risk detection in the banking system	CEO letters and outlook sections in banks' annual reports	NB, SVM	Accuracy, precision, recall	Accuracy (best out of all): 79.2%	Nopp and Hanbury (2015)
Distress prediction by evaluating sentiments	Annual reports and financial statements	Feed forward neural network (FFNN), SVM	Accuracy, F-measure, MCC	FFNN Accuracy: 92.22% F-measure: 0.923 MCC: 0.835 SVM Accuracy: 92.40% F-measure: 0.924 MCC: 0.838	Hájek and Olej (2013)
Sentiment analysis of the bank's customers	Bank reviews from mouthshut.com and myBankTracker.com	Ontology-driven sentiment analysis	Accuracy	73.118%	Chaturvedi and Chopra (2014)
Assessment of consumer financial complaints	Consumer complaints from the CFPB	SAS Contextual Analysis, SAS Visual Analytics, SAS Visual Statistics	Defining methods for text and visual analytics	–	Sabo (2017)
Opinion mining analysis in the banking system	Social media text from Twitter and Facebook	Rough set theory combined with a decision system	Precision, recall, accuracy	Accuracy (best out of all): 79%	Sumathi and Sheela (2017)
Analysis of bank reviews	Citibank reviews from Twitter, mouthshut.com, and myBankTracker.com	Opinion mining, sentiment analysis	Positive/negative	–	Gulaty (2016)

cosine similarity, and LDA. Some factors, such as limited access to the text data resources and insufficient collaboration between various sectors and disciplines, were identified as challenges that are hindering progress in the application of text mining to finance.

Arguing that corporate sustainability reports (CSR) have increased dramatically, become crucial from the financial reporting perspective, and are not amenable to manual analysis processes, Shahi et al. (2014) proposed an automated model based on text-mining approaches for more intelligent scoring of CSR reports. After pre-processing of the dataset, four classification algorithms were implemented, namely NB, random subspace, decision table, and neural networks. Various parameters were evaluated and the training categories and feature selection algorithms were tuned to determine the most effective model. NB with the Correlation-based Feature Selection (CFS) filter was chosen as the preferred model. Based on this model, software was designed for CSR report scoring that lets the user input a CSR report to get its score as an automated output. The software was tested and had an overall effectiveness of 81.10%. The authors concluded that the software could be utilised for other purposes such as the popularity of performance indicators as well.

Holton (2009) implemented a model for preventing corporate financial fraud with a different and interesting perspective. The author considered employee disgruntlement or employee dissatisfaction as a hidden indicator that is responsible for fraud. A minimal dataset of intra-company communication messages and emails on online discussion groups was prepared. After using document clustering for estimating that the data possess sufficient predictive power, the NB classifier was implemented to classify the messages into disgruntled/non-disgruntled classes, and an accuracy of 89% was achieved. The author proposed the use of the model for fraud risk assessment in corporations and organisations with the motivation that it can be used to prevent huge financial losses. The performance of other models such as neural networks and decision trees was to be compared in future work.

Chan and Franklin (2011) developed a new decision-support system to predict the occurrence of an event by analysing patterns and extracting sequences from financial reports. After text preprocessing, textual information generalisation was performed with the help of a shallow parser, which had an F-measure of 85%. The extracted information was stored in a separate database. From this database, the event sequences were identified and extracted. A decision tree model was then implemented on these sequences to create an inference engine that could predict the occurrence of new events based on the training sequences. With an 85: 15% training-to-testing split, the model achieved an overall accuracy of 89.09%. The authors concluded by highlighting that their model had better and robust performance compared to the prevailing models.

Humpherys et al. (2011) reviewed various text-mining methods and theories that have been proposed for the detection of corporate fraud in financial statements and subsequently devised a methodology of their own. Their dataset comprised the Management's Discussion and Analysis section of corporate annual financial reports. After basic analysis and reduction, various statistical and machine learning algorithms were implemented on the dataset, among which the NB and C4.5 decision

tree models both gave the highest accuracy of 67.3% for classifying 10-K reports into fraudulent and non-fraudulent. The authors suggested that their model can be used by auditors for detecting fraudulent statements in reports with the aid of the Agent99 analyser tool.

Loughran and McDonald (2011) came up with the argument that the word lists contained in the Harvard Dictionary, which is commonly used for textual analysis, are not suitable for financial text classification because a lot of negative words in the Harvard list are not actually considered a negative in the financial context. Corporate 10-K reports were taken as data sources to create a new dictionary with new word lists for financial purposes. The authors advised the use of term weighting for the word lists. The new word lists were compared with the Harvard word lists on multiple financial data items, such as 10-K filing returns, material weaknesses, and standardised unexpected earnings. Although a significant difference between the word lists was not observed for classification, the authors still suggested the use of their lists in order to be more careful and prevent any erroneous results.

Whereas other researchers have mostly focused on fraud detection and financial predictions from corporate financial reports, Song et al. (2018) focused on sentiment analysis of these reports with respect to the CSR score. The sentences in the sample reports were manually labelled as positive and negative in order to create sample data for the machine learning algorithm. SVM was implemented on the dataset with a 3:1 training to test split, which achieved a precision ratio of 86.83%. Following this, an object library was created, with objects referring to the internal and external environment of the company. Sentiment analysis was conducted on these objects. Then, six regression models were developed to get the CSR score, with the model comprising of the Political, Economic, Social, Technological, Environmental and Legal (PESTEL), Porter's Five Forces, and Primary and Support Activities showing the best performance in predicting the CSR score. The authors concluded that CSR plays a vital role in a company's sustainability, and their research could aid stakeholders in their company-related decision-making.

There have been more studies on CSR reports and sustainability. Liew et al. (2014) analysed process industries for their sustainability trends with the help of CSR and sustainability reports of a large number of big companies. The RapidMiner tool was used for text preprocessing followed by generating frequency statistics, pruning, and further text refinement, which generated sustainability-related terms for analysis. The most occurring terms were taken into consideration to create a hierarchical tree model. Environment, health and safety, and social were identified as the key concepts for sustainability. Based on term occurrence and involvement, the authors classified the sustainability issues as specific, critical, rare, and general.

Table 3 presents some more studies on the applications of text mining in corporate finance. As evident from the table and the above-mentioned studies, the annual corporate reports are the most commonly used data source for text-mining applications.

Challenges and future scope

The financial sector is a significant driver of broader industry, and the increasing amount of data in this field has given rise to a number of applications that can be used to improve the field and achieve commercial objectives.

Table 3 Summary of recent research studies on some text-mining applications for corporate finance

Study	Datasets	Techniques/algorithms used	Evaluation parameters	Performance (on the basis of evaluation parameters)	References
Identifying text patterns for financial performance	Annual reports of US-listed companies (10-K)	Clustering, sentiment analysis	Correlations between text patterns in company's reports and its sales performance	–	Lee et al. (2018)
Company sustainability report analysis	Corporate disclosures	n-grams, content analysis	Variations in disclosure	−8.89 to 1.57%	Aureli (2017)
Competitive analysis from social media	Social media data from Social Mention, SABI database	Social mention tool	Pearson correlation, F-ratio	F-ratio: 3.361 (Good fit)	Gemar and Jiménez-Quintero (2015)
Automatic classification of accounting literature	Articles from EbscoHost online academic database	Bayes classifier, decision tree, rule-based classifier	Accuracy	87.27%	Chakraborty et al. (2014)
Financial reports analysis	Quarterly reports of telecom manufacturers	Prototype-matching text clustering	Change in tone of financial reports with respect to the company's performance	–	Kloptchenko et al. (2004)
Computer-aided analysis of corporate disclosures	Annual reports of publicly listed DAX companies	Dictionary model, topic extraction, link analysis	Frequency of keywords, trend analysis	–	Matthies and Coners (2015)
Deviation detection in financial statements	Financial statements	Link Grammar Parser, conceptual graphs	Cumulative similarity	–	Kamaruddin et al. (2007)
Corporate bankruptcy prediction	Japanese annual reports	Conditional probability, chi-square test	Word frequency statistical metrics	–	Shirata et al. (2011)
Financial statement fraud detection	Financial statements	Bag-of-words, SVM	Accuracy, recall, precision, purity, F-measure	Conceptual paper; hence, performance not evaluated	Gupta and Gill (2012)
Financial footnote analysis	Income tax footnotes from financial reports	NB, k-means, KNN, SVM, decision tree	Runtime, accuracy, RMSE, absolute error	NB performed best Runtime: 4 s Accuracy: 82.86% Absolute error: 0.171 RMSE: 0.414	Heidari and Felden (2015)

Figure 4 shows some common challenges faced by various text-mining techniques in the financial sector. The huge amount of data available is highly unstructured and has explicit meanings in addition to implicit ones. The data needs to undergo proper preprocessing before it can be used for analysis. Although lexicon lists are available for various domains, the financial sector has to have a specific dictionary for such approaches, so as to assign proper weights to corresponding aspects in the document. In addition to

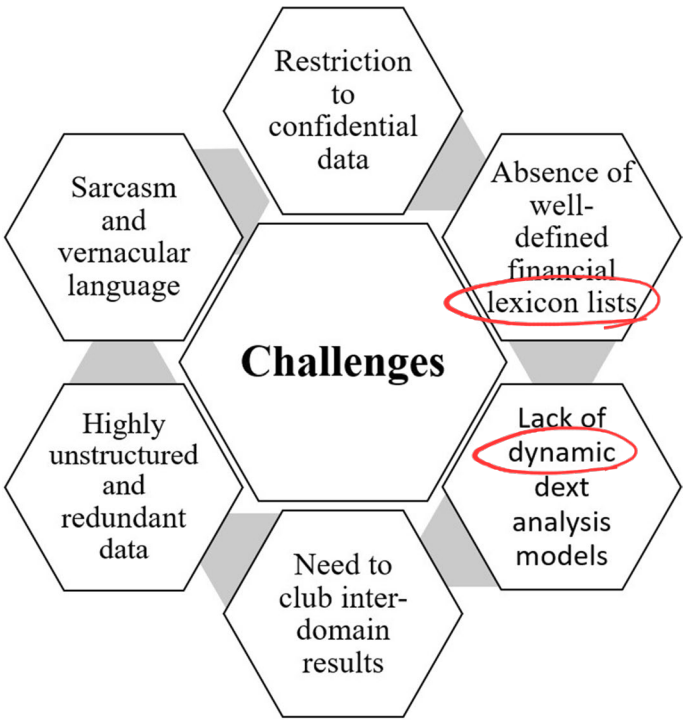

Fig. 4 Major challenges to text mining in finance

this, there is still restricted access to classified information, which is a significant obstacle. Lastly, the current techniques focus on obtaining static results statically that are true for a given period of time. There is a need for a system that performs text-mining techniques on dynamically obtained data to output real-time results to enable even better insights.

The combination of text-mining techniques and financial data analytics can produce a model that can potentially be the most efficient model for this problem domain. The results obtained from mining textual data can be integrated with those from financial analysis, thereby providing models that focus on historical data as well as opinions from diverse sources.

Conclusion

This paper conducted an organised qualitative review of recent literature pertaining to three specific sectors of finance. First, this paper analysed the growing importance of text mining in predicting financial trends. While the prior consensus may have been that financial markets are unpredictable, text mining has challenged this notion. The second area of study was banking, which has seen constant growth in technological innovation over the years, especially in digitisation. Text mining has played a key role in supporting these advancements both directly and indirectly through combination with other technologies. Corporate finance was the third study area. We discussed the importance of text mining in enabling the utilisation of corporate reports and financial statements for serving various purposes in addition to supporting corporate sustainability goals. The use of text mining in financial applications is not limited to these sectors. Researchers

are increasingly showing interest in text-mining applications and constantly seeking to build more accurate models. There are still many unexplored possibilities in the financial domain, and the related research can help develop more robust and accurate predictive and analytic systems.

Acknowledgements
The authors are grateful to Nirma University and Department of Chemical Engineering, School of Technology, Pandit Deendayal Petroleum University for the permission to publish this research.

Authors' contributions
All the authors make substantial contribution in this manuscript. AG, VD, HA and MS participated in drafting the manuscript. AG, VD and HA wrote the main manuscript, all the authors discussed the results and implication on the manuscript at all stages. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
All relevant data and material are presented in the main paper.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Computer Science, Nirma University, Ahmedabad, Gujarat, India. [2] Department of Chemical Engineering, School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, Gujarat 382007, India.

Received: 29 January 2020 Accepted: 17 September 2020
Published online: 02 November 2020

References

Agichtein E, Castillo C, Donato D, Gionis A, Mishne G (2008) Finding high-quality content in social media. In: Proceedings of the international conference on web search and web data mining—WSDM '08. https://doi.org/10.1145/1341531.1341557

Ahir K, Govani K, Gajera R, Shah M (2020) Application on virtual reality for enhanced education learning, military training and sports. Augment Hum Res 5:7

Ahmad K, Cheng D, Almas Y (2006) Multi-lingual sentiment analysis of financial news streams. In: Proceedings of science, pp 1–8

Akaichi J, Dhouioui Z, López-Huertas Pérez MJ (2013) Text mining facebook status updates for sentiment classification. In: 2013 17th international conference on system theory, control and computing (ICSTCC), Sinaia, 2013, pp 640–645. https://doi.org/10.1109/ICSTCC.2013.6689032

Al-Natour S, Turetken O (2020) A comparative assessment of sentiment analysis and star ratings for consumer reviews. Int J Inf Manage. https://doi.org/10.1016/j.ijinfomgt.2020.102132

AL-Rubaiee H, Qiu R, Li D (2015) Analysis of the relationship between Saudi twitter posts and the Saudi stock market. In: 2015 IEEE seventh international conference on intelligent computing and information systems (ICICIS). https://doi.org/10.1109/intelcis.2015.7397193

Audrino F, Sigrist F, Ballinari D (2018) The impact of sentiment and attention measures on stock market volatility. Available at SSRN: https://ssrn.com/abstract=3188941 or https://doi.org/10.2139/ssrn.3188941

Aureli S (2017) A comparison of content analysis usage and text mining in CSR corporate disclosure. Int J Digit Account Res 17:1–32

Bach MP, Krsti Z, Seljan S, Turulja L (2019) Text mining for big data analysis in financial sector: a literature review. Sustainability 2019(11):1277

Bharti SK, Babu KS (2017) Automatic keyword extraction for text summarization: a survey. CoRR. abs/1704.03242.

Bhattacharyya S, Jha S, Tharakunnel K, Westland JC (2011) Data mining for credit card fraud: a comparative study. Decis Support Syst 50(3):602–613

Bholat D, Hansen S, Santos P, Schonhardt-Bailey C (2015) Text mining for central banks: handbook. Centre Cent Bank Stud 33:1–19

Bidulya Y, Brunova E (2016) Sentiment analysis for bank service quality: a rule-based classifier. In: 2016 IEEE 10th international conference on application of information and communication technologies (AICT). https://doi.org/10.1109/icaict.2016.7991688

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(2003):993–1022

Brindha S, Prabha K, Sukumaran S (2016) A survey on classification techniques for text mining. In: 2016 3rd international conference on advanced computing and communication systems (ICACCS), Coimbatore, 2016, pp 1–5. https://doi.org/10.1109/ICACCS.2016.7586371
Bruno G (2016) Text mining and sentiment extraction in central bank documents. In: 2016 IEEE international conference on big data (big data). https://doi.org/10.1109/bigdata.2016.7840784
Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107. https://doi.org/10.1109/MIS.2016.31
Chakraborty V, Chiu V, Vasarhelyi M (2014) Automatic classification of accounting literature. Int J Account Inf Syst 15(2):122–148
Chan SWK, Franklin J (2011) A text-based decision support system for financial sequence prediction. Decis Support Syst 52(1):189–198
Chaturvedi D, Chopra S (2014) Customers sentiment on banks. Int J Comput Appl 98(13):8–13
Chen CC, Huang HH, Chen HH (2020) NLP in FinTech applications: past, present and future
Cook A, Herron B (2018) Harvesting unstructured data to reduce anti-money laundering (AML) compliance risk, pp 1–10
Daniel K, Hirshleifer D, Teoh S (2001) Investor psychology in capital markets: evidence and policy implications. J Monet Econ 49:139–209. https://doi.org/10.1016/S0304-3932(01)00091-5
Da-sheng W, Qin-fen Y, Li-juan L (2009) An efficient text classification algorithm in E-commerce application. In: 2009 WRI world congress on computer science and information engineering. https://doi.org/10.1109/csie.2009.346
David JM, Balakrishnan K (2011) Prediction of key symptoms of learning disabilities in school-age children using rough sets. Int J Comput Electr Eng Hong Kong 3(1):163–169
Dohaiha H, Prasad PWC, Maag A, Alsadoon A (2018) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2018.10.003
Elagamy MN, Stanier C, Sharp B (2018) Stock market random forest-text mining system mining critical indicators of stock market movements. In: 2018 2nd international conference on natural language and speech processing (ICNLSP). https://doi.org/10.1109/icnlsp.2018.8374370
Emekligil E, Arslan S, Agin O (2016) A bank information extraction system based on named entity recognition with CRFs from noisy customer order texts in Turkish. In: Knowledge engineering and semantic web, pp 93–102
Espejo-Garcia B, Martinez-Guanter J, Pérez-Ruiz M, Lopez-Pellicer FJ, Javier Zarazaga-Soria F (2018) Machine learning for automatic rule classification of agricultural regulations: a case study in Spain. Comput Electron Agric 150:343–352
Fama EF (1991) Efficient capital markets: II. J Finance 46(5):1575–1617. https://doi.org/10.2307/2328565
Fan W, Wallace L, Rich S, Zhang Z (2006) Tapping the power of text mining. Commun ACM 49(9):76–82
Feuerriegel S, Gordon J (2018) Long-term stock index forecasting based on text mining of regulatory disclosures. Decis Support Syst 112:88–97
Fisher I, Garnsey M, Hughes M (2016) Natural language processing in accounting, auditing and finance: a synthesis of the literature with a roadmap for future research. Intell Syst Account Finance Manag. https://doi.org/10.1002/isaf.1386
Fritz D, Tows E (2018) Text mining and reporting quality in German banks—a cooccurrence and sentiment analysis. Univers J Account Finance 6(2):54–81
Fung G, Yu J, Lam W (2002) News sensitive stock trend prediction. Adv Knowl Discov Data Min. https://doi.org/10.1007/3-540-47887-6_48
Gandhi M, Kamdar J, Shah M (2020) Preprocessing of non-symmetrical images for edge detection. Augment Hum Res 5:10. https://doi.org/10.1007/s41133-019-0030-5
Gao Z, Ye M (2007) A framework for data mining-based anti-money laundering research. J Money Laund Control 10(2):170–179
Gemar G, Jiménez-Quintero JA (2015) Text mining social media for competitive analysis. Tour Manag Stud 11(1):84–90
Gulaty M (2016) Aspect-based sentiment analysis in bank reviews. https://doi.org/10.13140/RG.2.1.2072.3445
Guo L, Shi F, Tu J (2016) Textual analysis and machine leaning: crack unstructured data in finance and accounting. J Finance Data Sci 2(3):153–170
Gupta R, Gill NS (2012) Financial statement fraud detection using text mining. Int J Adv Comput Sci Appl 3(12):189–191
Gupta A, Simaan M, Zaki MJ (2016) Investigating bank failures using text mining. In: 2016 IEEE symposium series on computational intelligence (SSCI). https://doi.org/10.1109/ssci.2016.7850006
Gupta A, Bhatia P, Dave K, Jain P (2019) Stock market prediction using data mining techniques. In: 2nd international conference on advances in science and technology, pp 1–5
Hagenau M, Liebmann M, Neumann D (2013) Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Support Syst 55(3):685–697
Hájek P, Olej V (2013) Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In: Communications in computer and information science, pp 1–10.
Hariri RH, Fredericks EM, Bowers KM (2019) Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. https://doi.org/10.1186/s40537-019-0206-3
Hassonah M, Al-Sayyed R, Rodan A, Al-Zoubi A, Aljarah I, Faris H (2019) An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.105353
Heaton JB, Polson NG, Witte JH (2016) Deep learning in finance. arXiv:1602.06561
Heidari M, Felden C (2015) Financial footnote analysis: developing a text mining approach. In: Int'l conf. data mining, pp 10–16
Herranz S, Palomo J, Cruz M (2018) Building an educational platform using NLP: a case study in teaching finance. J Univ Comput Sci 24:1403
Holton C (2009) Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem. Decis Support Syst 46(4):853–864
Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF (2011) Identification of fraudulent financial statements using linguistic credibility analysis. Decis Support Syst 50(3):585–594
IBEF (2019) https://www.ibef.org/download/financial-services-april-2019.pdf

James TL, Calderon EDV, Cook DF (2017) Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. Expert Syst Appl 71:479–492
Jani K, Chaudhuri M, Patel H, Shah M (2019) Machine learning in films: an approach towards automation in film censoring. J Data Inf Manag. https://doi.org/10.1007/s42488-019-00016-9
Jaseena KU, David JM (2014) Issues, challenges, and solutions: big data mining. In: Natarajan Meghanathan et al. (eds) NeTCoM, CSIT, GRAPH-HOC, SPTM—2014, pp 131–140
Jha K, Doshi A, Patel P, Shah M (2019) A comprehensive review on automation in agriculture using artificial intelligence. Artif Intell Agric 2:1–12
Joshi K, Bharathi N, Jyothi R (2016) Stock trend prediction using news sentiment analysis. Int J Comput Sci Inf Technol 8:67–76. https://doi.org/10.5121/ijcsit.2016.8306
Junqué de Fortuny E, De Smedt T, Martens D, Daelemans W (2014) Evaluating and understanding text-based stock price prediction models. Inf Process Manag 50(2):426–441
Kakkad V, Patel M, Shah M (2019) Biometric authentication and image encryption for image security in cloud framework. Multiscale Multidiscip Model Exp Des. https://doi.org/10.1007/s41939-019-00049-y
Kamaruddin SS, Hamdan AR, Bakar AA (2007) Text mining for deviation detection in financial statements. In: Proceedings of the international conference on electrical engineering and informatics. Institut Teknologi Bandung, Indonesia, 2007, June 17–19
Kang T, Park DH (2016) The effect of expert reviews on consumer product evaluations: a text mining approach. J Intell Inf Syst 22(1):63–82
Kinsella S, Passant A, Breslin JG (2011) Topic classification in social media using metadata from hyperlinked objects. Adv Inf Retr. https://doi.org/10.1007/978-3-642-20161-5_20
Kloptchenko A, Eklund T, Karlsson J, Back B, Vanharanta H, Visa A (2004) Combining data and text mining techniques for analysing financial reports. Intell Syst Account Finance Manag 12(1):29–41
Kordonis J, Symeonidis S, Arampatzis A (2016) Stock price forecasting via sentiment analysis on twitter. https://doi.org/10.1145/3003733.3003787.
Kou G, Lu Y, Peng Y, Shi Y (2012) Evaluation of classification algorithms using MCDM and rank correlation. Int J Inf Technol Decis Mak. https://doi.org/10.1142/S0219622012500095
Kou G, Peng Yi, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12. https://doi.org/10.1016/j.ins.2014.02.137
Kou G, Yang P, Peng Yi, Xiao F, Chen Y, Alsaadi F (2019) Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Appl Soft Comput 86:105836. https://doi.org/10.1016/j.asoc.2019.105836
Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. Information 10:150
Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-aryamontri A, Winter A, Perfetto L, Briganti L, Licata L, Iannuccelli M, Castagnoli L, Cesareni G, Tyers M, Schneider G, Rinaldi F, Leaman R, Gonzalez G, Matos S, Kim S, Wilbur WJ, Rocha L, Shatkay H, Tendulkar AV, Agarwal S, Liu F, Wang X, Rak R, Noto K, Elkan C, Lu Z, Dogan RI, Fontaine JF, Andrade-Navarro MA, Valencia A (2011) The protein–protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinform 12(Suppl 8):S3. https://doi.org/10.1186/1471-2105-12-s8-s3
Kraus M, Feuerriegel S (2017) Decision support from financial disclosures with deep neural networks and transfer learning. Decis Support Syst. https://doi.org/10.1016/j.dss.2017.10.001
Krstić Ž, Seljan S, Zoroja J (2019) Visualization of big data text analytics in financial industry: a case study of topic extraction for Italian banks (September 12, 2019). In: 2019 ENTRENOVA conference proceedings. https://ssrn.com/abstract=3490108 or https://doi.org/10.2139/ssrn.3490108
Kumar BS, Ravi V (2016) A survey of the applications of text mining in financial domain. Knowl Based Syst 114:128–147
Kundalia K, Patel Y, Shah M (2020) Multi-label movie genre detection from a movie poster using knowledge transfer learning. Augment Hum Res 5:11. https://doi.org/10.1007/s41133-019-0029-y
Lavrenko V, Schmill M, Lawrie D, Ogilvie P, Jensen D, Allan J (2000) Mining of concurrent text and time series. In: KDD-2000 Workshop on text mining, vol 2000. Citeseer, pp 37–44
Lee CT (2019) Early warning mechanism of agricultural network public opinion based on text mining. Revista De La Facultad De Agronomia De La Universidad Del Zulia, 36
Lee B, Park JH, Kwon L, Moon YH, Shin Y, Kim G, Kim H (2018) About relationship between business text patterns and financial performance in corporate data. J Open Innov Technol Market Complex. https://doi.org/10.1186/s40852-018-0080-9
Lewis C, Young S (2019) Fad or future? Automated analysis of financial text and its implications for corporate reporting. Account Bus Res 49(5):587–615
Li N, Liang X, Li X, Wang C, Wu DD (2009) Network Environment and Financial Risk Using Machine Learning and Sentiment Analysis. Human Ecol Risk Assess Int J 15(2):227–252. https://doi.org/10.1080/10807030902761056
Li T, Kou G, Peng Y, Shi Y (2020a) Classifying with adaptive hyper-spheres: an incremental classifier based on competitive learning. IEEE Trans Syst Man Cybern Syst 50(4):1218–1229. https://doi.org/10.1109/TSMC.2017.2761360
Li X, Wu P, Wang W (2020b) Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong. Inf Process Manag. https://doi.org/10.1016/j.ipm.2020.102212
Li T, Kou G, Peng Yi (2020c) Improving malicious URLs detection via feature engineering: linear and nonlinear space transformation methods. Inf Syst 91:101494. https://doi.org/10.1016/j.is.2020.101494
Liew WT, Adhitya A, Srinivasan R (2014) Sustainability trends in the process industries: a text mining-based analysis. Comput Ind 65(3):393–400
Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceeding of the 18th ACM conference on information and knowledge management—CIKM '09. https://doi.org/10.1145/1645953.1646003
Loughran T, Mcdonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Finance 66(1):35–65

Lu Y (2013) Automatic topic identification of health-related messages in online health community using text classification. SpringerPlus 2(1):309
Malandri L, Xing F, Orsenigo C, Vercellis C, Cambria E (2018) Public mood-driven asset allocation: the importance of financial sentiment in portfolio management. Cogn Comput. https://doi.org/10.1007/s12559-018-9609-2
Marrara S, Pejic Bach M, Seljan S, Topalovic A (2019) FinTech and SMEs—the Italian case. https://doi.org/10.4018/978-1-5225-7805-5.ch002
Matthies B, Coners A (2015) Computer-aided text analysis of corporate disclosures—demonstration and evaluation of two approaches. Int J Digit Account Res 15:69–98
Mudinas A, Zhang D, Levene M (2019) Market trend prediction using sentiment analysis: lessons learned and paths forward. arXiv:1903.05440
Nan L, Xun L, Xinli L, Chao W, Desheng DW (2009) Network environment and financial risk using machine learning and sentiment analysis. Hum Ecol Risk Assess Int J 15(2):227–252
Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DC (2015) Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Syst Appl 42(1):306–324. https://doi.org/10.1016/j.eswa.2014.08.004
Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42(24):9603–9611
Nikfarjam A, Emadzadeh E, Muthaiyah S (2010) Text mining approaches for stock market prediction. In: 2010 the 2nd international conference on computer and automation engineering (ICCAE). https://doi.org/10.1109/iccae.2010.5451705
Nopp C, Hanbury A (2015) Detecting risks in the banking system by sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 591–600
Panchiwala S, Shah MA (2020) Comprehensive study on critical security issues and challenges of the IoT world. J Data Inf Manag. https://doi.org/10.1007/s42488-020-00030-2
Pandya R, Nadiadwala S, Shah R, Shah M (2019) Buildout of methodology for meticulous diagnosis of K-complex in EEG for aiding the detection of Alzheimer's by artificial intelligence. Augment Hum Res. https://doi.org/10.1007/s41133-019-0021-6
Parekh V, Shah D, Shah M (2020) Fatigue detection using artificial intelligence framework. Augment Hum Res 5:5
Patel D, Shah Y, Thakkar N, Shah K, Shah M (2020a) Implementation of artificial intelligence techniques for cancer detection. Augment Hum Res. https://doi.org/10.1007/s41133-019-0024-3
Patel D, Shah D, Shah M (2020b) The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. Ann Data Sci. https://doi.org/10.1007/s40745-019-00239-y
Patel H, Prajapati D, Mahida D, Shah M (2020c) Transforming petroleum downstream sector through big data: a holistic review. J Petrol Explor Prod Technol. https://doi.org/10.1007/s13202-020-00889-2
Pathan M, Patel N, Yagnik H, Shah M (2020) Artificial cognition for applications in smart agriculture: a comprehensive review. Artif Intell Agric. https://doi.org/10.1016/j.aiia.2020.06.001
Pejic Bach M, Krstić Ž, Seljan S, Turulja L (2019) Text mining for big data analysis in financial sector: a literature review. Sustainability 11:1277. https://doi.org/10.3390/su11051277
Picasso A, Merello S, Ma Y, Oneto L, Cambria E (2019) Technical analysis and sentiment embeddings for market trend prediction. Expert Syst Appl 135:60–70. https://doi.org/10.1016/j.eswa.2019.06.014
Pradhan MV, Vala J, Balani P (2016) A survey on sentiment analysis algorithms for opinion mining. Int J Comput Appl 133:7–11. https://doi.org/10.5120/ijca2016907977
Ray P, Chakrabarti A (2019) A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Appl Comput Inform. https://doi.org/10.1016/j.aci.2019.02.002
Renault T (2019) Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. Digit Finance. https://doi.org/10.1007/s42521-019-00014-x
Sabo T (2017) Applying text analytics and machine learning to assess consumer financial complaints. In: Proceedings of the SAS global forum 2017 conference. SAS Institute Inc., Cary NC. https://support.sas.com/resources/papers/proceedings17/SAS0282-2017.pdf
Salloum S, Al-Emran M, Monem A, Shaalan K (2017) A survey of text mining in social media: facebook and twitter perspectives. Adv Sci Technol Eng Syst J 2:127–133. https://doi.org/10.25046/aj020115
Salloum S, Mostafa A, Monem A, Shaalan K (2018) Using text mining techniques for extracting information from research articles. https://doi.org/10.1007/978-3-319-67056-0_18
Schneider MJ, Gupta S (2016) Forecasting sales of new and existing products using consumer reviews: a random projections approach. Int J Forecast 32(2):243–256
Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news. ACM Trans Inf Syst 27(2):1–19
Shah D, Isah H, Zulkernine F (2018a) Predicting the effects of news sentiments on the stock market. In: 2018 IEEE international conference on big data (big data). https://doi.org/10.1109/bigdata.2018.8621884
Shah T, Shaikh I, Patel A (2018b) Comparison of different kernels of support vector machine for predicting stock prices. Int J Eng Technol 9(6):4288–4291
Shah G, Shah A, Shah M (2019) Panacea of challenges in real-world application of big data analytics in healthcare sector. Data Inf Manag. https://doi.org/10.1007/s42488-019-00010-1
Shah D, Dixit R, Shah A, Shah P, Shah M (2020) A Comprehensive analysis regarding several breakthroughs based on computer intelligence targeting various syndromes. Augment Hum Res 5:14. https://doi.org/10.1007/s41133-020-00033-z
Shah K, Patel H, Sanghvi D, Shah M (2020) A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment Hum Res 5:12. https://doi.org/10.1007/s41133-020-00032-0
Shahi AM, Issac B, Modapothala JR (2014) Automatic analysis of corporate sustainability reports and intelligent SCORING. Int J Comput Intell Appl 13(01):1450006. https://doi.org/10.1142/s1469026814500060

Shirata CY, Takeuchi H, Ogino S, Watanabe H (2011) Extracting key phrases as predictors of corporate bankruptcy: empirical analysis of annual reports by text mining. J Emerg Technol Account 8(1):31–44
Sohangir S, Wang D, Pomeranets A et al (2018) Big data: deep learning for financial sentiment analysis. J Big Data 5:3. https://doi.org/10.1186/s40537-017-0111-6
Song Y, Wang H, Zhu M (2018) Sustainable strategy for corporate governance based on the sentiment analysis of financial reports with CSR. Financ Innov. https://doi.org/10.1186/s40854-018-0086-0
Souma W, Vodenska I, Aoyama H (2019) Enhanced news sentiment analysis using deep learning methods. J Comput Soc Sci 2:33–46. https://doi.org/10.1007/s42001-019-00035-x
Srivastava SK, Singh SK, Suri JS (2018) Healthcare text classification system and its performance evaluation: a source of better intelligence by characterizing healthcare text. J Med Syst. https://doi.org/10.1007/s10916-018-0941-6
Su Y, Wang R, Chen P, Wei Y, Li C, Hu Y (2012) Agricultural ontology based feature optimization for agricultural text clustering. J Integr Agric 11(5):752–759
Sukhadia A, Upadhyay K, Gundeti M, Shah S, Shah M (2020) Optimization of smart traffic governance system using artificial intelligence. Augment Hum Res 5:13. https://doi.org/10.1007/s41133-020-00035-x
Sumathi N, Sheela T (2017) Opinion mining analysis in banking system using rough feature selection technique from social media text. Int J Mech Eng Technol 8(12):274–289
Talaviya T, Shah D, Patel N, Yagnik H, Shah M (2020) Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artif Intell Agric. https://doi.org/10.1016/j.aiia.2020.04.002
Talib R, Hanif MK, Ayesha S, Fatima F (2016a) Text mining: techniques. Appl Issues 7(11):414–418
Talib R, Kashif M, Ayesha S, Fatima F (2016b) Text mining: techniques, applications and issues. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2016.071153
Tkáč M, Verner R (2016) Artificial neural networks in business: two decades of research. Appl Soft Comput 38:788–804
Ur-Rahman N, Harding JA (2012) Textual data mining for industrial knowledge management and text classification: a business oriented approach. Expert Syst Appl 39(5):4729–4739
Vijayan R, Potey MA (2016) Improved accuracy of FOREX intraday trend prediction through text mining of news headlines using J48. Int J Adv Res Comput Eng Technol 5(6):1862–1866
Vu TT, Chang S, Ha QT, Collier N (2012) An experiment in integrating sentiment features for tech stock prediction in twitter. In: Workshop on information extraction and entity analytics on social media data, COLING, Mumbai, India, pp 23–38
Wang B, Huang H, Wang X (2012) A novel text mining approach to financial time series forecasting. Neurocomputing 83:136–145
Wen F, Xu L, Ouyang G, Kou G (2019) Retail investor attention and stock price crash risk: evidence from China. Int Rev Financ Anal 65:101376. https://doi.org/10.1016/j.irfa.2019.101376
Widiastuti N (2018) Deep learning—now and next in text mining and natural language processing. IOP Conf Ser Mater Sci Eng 407:012114. https://doi.org/10.1088/1757-899X/407/1/012114
Wu JL, Su CC, Yu LC, Chang PC (2012) Stock price predication using combinational features from sentimental analysis of stock news and technical analysis of trading information. Int Proc Econ Dev Res. https://doi.org/10.7763/ipedr
Wu DD, Zheng L, Olson DL (2014) A decision support approach for online stock forum sentiment analysis. IEEE Trans Syst Man Cybern Syst 44(8):1077–1087
Xing FZ, Cambria E, Welsch RE (2017) Natural language based financial forecasting: a survey. Artif Intell Rev 50(1):49–73
Xing FZ, Cambria E, Welsch RE (2018a) Natural language based financial forecasting: a survey. Artif Intell Rev 50:49–73. https://doi.org/10.1007/s10462-017-9588-9
Xing F, Cambria E, Welsch R (2018b) Intelligent asset allocation via market sentiment views. IEEE Comput Intell Mag 13:25–34. https://doi.org/10.1109/MCI.2018.2866727
Xiong T, Wang S, Mayers A, Monga E (2013) Personal bankruptcy prediction by mining credit card data. Expert Syst Appl 40(2):665–676
Xu G, Yu Z, Yao H, Li F, Meng Y, Wu X (2019) Chinese text sentiment analysis based on extended sentiment dictionary. IEEE Access 7:43749–43762. https://doi.org/10.1109/ACCESS.2019.2907772
Yang Li, Li Y, Wang J, Sherratt R (2020) Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access 8:1–1. https://doi.org/10.1109/ACCESS.2020.2969854
Yap BW, Ong SH, Husain NHM (2011) Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst Appl 38(10):13274–13283
Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing [review article]. IEEE Comput Intell Mag 13(3):55–75. https://doi.org/10.1109/MCI.2018.2840738
Yusuuf H, Shihabeldeen A (2019) Using text mining to predicate exchange rates with sentiment indicators. J Bus Theory Pract 7(2):60–75
Zavolokina L, Dolata M, Schwabe G (2016) The FinTech phenomenon: antecedents of financial innovation perceived by the popular press. Financ Innov. https://doi.org/10.1186/s40854-016-0036-7

Publisher's Note