Expert Systems With Applications 164 (2021) 114053

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Neural network embeddings on corporate annual filings for portfolio selection

George Adosoglou [a,*], Gianfranco Lombardo [b], Panos M. Pardalos [a,1]

[a] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA
[b] Department of Engineering and Architecture, University of Parma, Parma, Italy

ARTICLE INFO

Keywords:
Neural network embedding
Document embedding
Machine learning
Asset pricing
Inattention
Annual reports

ABSTRACT

In recent years, there has been an increased interest from both academics and practitioners in automatically analyzing the textual part of companies' financial reports to extract meaning rich in information for future outcomes. In particular, tracking textual changes among companies' reports can have a large and significant impact on stock prices. This impact happens with a lag implying that investors only gradually realize the implications of the news hinted by document changes. However, the length of these documents as well as their complexity in terms of structure and language have been increasing dramatically making this process more and more difficult to perform. In this paper, we analyzed how to face this complexity by learning arbitrary dimensional vector representations for US corporate filings (10-Ks) from 1998 to 2018, exploiting and comparing different neural network embedding techniques which take into account words' semantics through vectors proximity. We also compared their ability to capture changes associated with future risk-adjusted abnormal returns with other more commonly used approaches in literature. Finally, we propose a novel investment strategy named Semantic Similarity Portfolio (SSP) that exploits these neural network embeddings. We show that firms that do not change their 10-Ks in a semantically important way from the previous year tend to have large and statistically significant future risk-adjusted abnormal returns. We, also document an amplifying effect when we incorporate a momentum-related criterion, where the companies selected must also have had positive previous year returns. Specifically, a portfolio that buys "non-changers" based on this strategy earns up to 10% in yearly risk-adjusted abnormal returns (alpha).

1. Introduction

Published in 1998 by the Securities and Exchange Commission (SEC), the Plain English Handbook was the first publication providing guidelines to help public companies create clear SEC disclosure documents. This publication and the Sarbanes–Oxley Act of 2002, which was constructed to supervise the financial reporting, have made corporate filings an increasingly reliable source of information. It can be argued that the detailed summary of a company's financial performance required by the SEC in a 10-K make it the most comprehensive corporate filing with a regular cadence and repeated use. Given the huge number of forms of this kind filed annually, however, it has become extremely cumbersome for investors to analyze and make informed decisions based on them. Specifically, firms that experience economic changes are mandated to update the risk factor item of their 10-Ks and include the most recent information available to them. A lot of these changes, however, go unattended by investors for long periods of time (as long as one year) as Cohen et al. showed in their paper "Lazy Prices".

In Lazy Prices, Cohen et al. (2019) argued that a simple comparison of consecutive 10-Ks hides a lot of valuable information. It is true that while tables in financial statements are always presented with the current year's numbers accompanied by several previous years' corresponding numbers, the same is untrue for the text. The management being "lazy" a lot of times uses last years filings verbatim in constructing the current year's 10-Ks while making only the necessary changes so as to be within the boundaries of fiduciary responsibility. Observing these changes yields an important, and robust indication for future firm performance. L. Cohen et al. showed this by computing quintiles from the distribution of the similarity scores from all companies and then constructing long-short equally and capitalization weighted portfolios. The first Quintile (Q1) in this framework includes companies with the biggest differences between their documents, referred to as "Changers", while Quintile 5 (Q5) represents firms whose fillings have the biggest similarities ("Non-Changers"). The portfolio that they constructed has a

* Corresponding author.
E-mail addresses: g.adosoglou@ufl.edu (G. Adosoglou), gianfranco.lombardo@unipr.it (G. Lombardo), pardalos@ufl.edu (P.M. Pardalos).
[1] This author was supported by a Humboldt Research award (Germany).

https://doi.org/10.1016/j.eswa.2020.114053
Received 26 June 2020; Received in revised form 19 September 2020; Accepted 23 September 2020
Available online 25 September 2020
0957-4174/Published by Elsevier Ltd.

holding period of 3 months while the re-balancing occurs every month. Specifically, they found that going long the "non-changers" and short the "changers" yields statistically significant 5-factor alphas proving that breaks from previous standardized reporting can have significant implications for firms' future stock returns.

It is normal, however, for managers to be incentivized to minimize (maximize) the effect on their companies' stock prices from negative news (positive) news about their firms respectively (Laughran & Mcdonald, 2011). Previous works Dyer et al. (2017) showed that the managers provide boilerplate information and avoid giving accurate signals of the company's status by extending the document length. However, the SEC prohibits any misleading statement or omission under Rule 10b-5 and demands a company's CEO and CFO to certify the accuracy of the 10-K. This means that even though valuable information about the company and the industry does exist in the 10-Ks, the management has incentives to hide it. We argue that since the methods used in the "Lazy Prices" paper to measure document similarity ignore syntax or semantics, these differences can be better captured with a model that does, especially in cases where the CEO/CFO strategically obfuscate risks and corporate issues (Li, 2008).

The novelty in our approach is to represent each company, its activities and current affairs as a vector by applying neural network embedding techniques to the financial annual reports of these companies. In this way, we are able to capture changes associated with future risk-adjusted abnormal returns by taking into account semantics and temporal dynamics. We also demonstrate that incorporating a momentum-related component into our portfolio selection method provides significant synergies further adding to the originality of our paper.

We tested two different approaches: Word embedding with Word2-Vec algorithm by averaging these vectors to embed the entire document, and Doc2Vec algorithm that is able to directly learn document embeddings following different approaches than averaging word vectors. Both methodologies are based on shallow neural networks with linear activation function and unsupervised learning approach that can preserve words' order and semantics. Then, the similarity between documents can be measured using the cosine similarity measure. Namely, compared to "Lazy Prices" that focuses more on exploiting the unattended disclosed information like adding or deleting sentences in the document, our model focuses more on the changes in the topic covered and writing style in the 10-Ks by representing arbitrarily the entire documents with vectors in a fixed-dimensional semantic space.

This paper has two goals. First, we show that the neural network embedding techniques represent an interesting approach that is able to address the increasing complexities of annual reports' textual analysis. In light of this, we construct a portfolio, named Semantic Similarity Portfolio (SSP), that exploits the Distributed Memory Model of Paragraph Vectors (PV-DM) mode of Doc2Vec which we found to be the best performing technique for this task. The neural network embedding approach produced a superior result compared to the popular alternative Bag-of-Words (BoW) model (Salton et al., 1975) in capturing changes in consecutive 10-Ks found significant to future abnormal portfolio returns. The second goal is to show that incorporating a momentum-related criterion, based on a "non-struggling" companies attribute computed on prior companies' returns, can have a significant amplifying effect on excess risk-adjusted returns. It can be argued that this criterion can signal the nature of these changes since a struggling company would keep it's 10-K semantically unchanged if it's management believed that the challenges they are currently facing will persist in the upcoming year as well. In other words, in this portfolio setting, we also avoid companies with persistent risks and difficulties that are documented in the 10-K-s but are not being removed by the CEO/CFO leaving the 10-Ks semantically unchanged.

2. State of the art

Machine learning applications on text have almost four decades of history. However, only in the last decades a set of machine learning techniques known as neural networks (NNs) have continued to advance and start to prove highly effective for a great number of natural language processing (NLP) tasks.

Financial news, in particular, have been extensively exploited to make predictions regarding the markets. While, more recently, social media and corporate disclosures have also been utilized in various applications.

Khadjeh Nassirtoussi et al. (2015), for example, produced a multi-layer algorithm testing three machine learning models, namely: SVM, k-nearest neighbors (*k*-NN) and Naïve Bayes, that exploit semantics and sentiment of news-headlines for a FOREX market prediction task. Van De Kauter et al. (2015) proposed a novel fine-grained approach that captures explicit and implicit topic-dependent sentiment in company-specific news text. Gunduz and Cataltepe (2015) trained a Naive Bayes classifier with daily news articles to predict the direction of the BIST100 Index for the day following. Classification techniques such as Naïve Bayes and SVM have also been exploited by Nizer and Nievola (2012) on news text to predict the volatility of financial assets. Market volatility related to news and exploited with neural networks (NNs) has also been studied extensively by Zopounidis et al. (2010) Wang et al. (2011), on the other hand, proposed an ontology based framework to mine dependence relationships between financial instruments and news. Finally, Lupiani-Ruiz et al. (2011) presents a semantic search engine for financial news using Semantic Web technologies customized on the Spanish stock market.

Neural networks (NNs) have, also, been applied to financial news by Day and Lee (2016), and for sentiment analysis tasks and predictors of volatility by Tetlock (2007). A version of Kohonen's self-organizing map, called spiral spherical neural network, has been applied by Jagric et al. (2015) to investigate the European Union banking sector and proving interesting insights about their reciprocal action and integration. Word Embeddings have been used by Peng and Jiang (2016) in leveraging financial news to predict stock prices. Neural networks have also been used to improve the performance of sentiment analysis for StockTwits by Sohangir et al. (2018). Finally, Cerchiello et al. (2017) apply Doc2Vec (Djuric et al., 2015) to detect bank distress by mining news and financial data. News analytics for buy and sell decisions have also been studied extensively in Doumpos et al. (2012) where models such as k-nearest neighbor, feed-forward NNs, SVM and Naive Bayesian classifiers are compared in classification tasks and sentiment analysis. Various other computational approaches for asset trading have also been compared with sentiment analysis and news text analytics by Andriosopoulos et al. (2019).

Furthermore, motivated by the works of Brown and Tucker (2011), who showed in their paper using the Bag-of-Words (BoW) model that firms that undergo significant economic changes modify the Management Discussion and Analysis (MD&A) section of the 10-K reports in a much greater way than the ones that do not. Li (2011) uses Naıve Bayesian machine learning algorithm to associate MD&A tone with future firm performance. Bandiera et al. (2020) apply the Latent Dirichlet allocation (LDA) model to a large panel of CEO diary data to estimate behavioral types and predict firm performance.

More recently, neural networks (NNs) have been also used in analyzing corporate filings. Rawte et al. (2018) use deep learning on the item 1A (Risk Factors) of various banks' 10-Ks for the classification task of predicting bank failures. Deep learning models have also been used on disclosures to predict corporate bankruptcies (Mai et al., 2019). Furthermore, Tsai et al. (2016) uses the Word2Vec model to learn the continuous-vector word representations in order to discover new finance keywords and update a financial dictionary.

This increasing interest in neural network based solution to financial text analysis is due in particular to the ability of these techniques to

leverage information in an unsupervised or supervised way by learning an internal representation of documents as learned weights during pattern recognition. Furthermore, in several applications, like the one we are describing in this paper, the ability to capture semantics plays a vital role in detecting changes that actually do have a meaning and are not related to the use of different words that have similar meanings like synonyms or in this regard to a different author trying to convey the same point.

In these cases, models such as the BoW model that ignore the semantics and syntax are deemed useless.

2.0.1. Neural language models

In the Natural Language Processing (NLP) field several approaches have been introduced to deal with text documents of all kinds for information retrieval and prediction tasks. The common goal of these techniques is to build a statistical model that is able to learn the joint probability function of sequences of words in a language. The key problems that previous methods, such as Bag-of-Words (BoW), failed to address are the absence of word ordering, lack of context and the curse of dimensionality: a new sentence on which the model is tested is likely to differ from all the word sequences that were seen while training with a resulting data-sparsity problem due to the increasing number of unique words, the vocabulary size and thus the representation size for each word or document. In Bengio et al. (2003) the authors first proposed a neural network language model, known as probabilistic feed-forward neural network language model or Distributed Representation, that is able to learn the joint probability while learning a word feature vector in R^n. It consists of input words, a shared projection matrix, hidden and output layers. Several other architectures based on neural networks have been introduced in literature to solve computational issues related to Distributed Representation in the case of large text mining, for further details see Jing et al. (2019). The one that gained the most promising results both in performance and complexity is Word2Vec, presented in Mikolov et al. (2013).

2.0.2. Word2Vec

Word2Vec is a word embedding algorithm that exploits a shallow neural network with a linear activation function to embed words into distributed low dimensional vectors. It provides two different models: the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. Both architectures are able to embed words in such a way that similar words should have similar embeddings in terms of spatial proximity. The CBOW model tries to predict a word given its context. It is called bag-of-words since the projection is not dependent on the order of the words in the history. The Skip-gram model is trained to predict neighbor words in the sentence. Here, since the more distant words are usually less related to the current word, the more distant the words are, the less they are sampled. In both architectures, the objective function is optimized using Stochastic Gradient Descent (SGD) in the form of back-propagation on just a single hidden-layer feed-forward neural network. One-hot encoded words are fed into the network, while the hidden layer has no activation function. The output layer is implemented with a hierarchical Softmax function. Values from the hidden layers are then the resulting node embedding vectors. This means that syntax and semantics are captured as the indirect result of predicting the next word in a sentence. In the Skip-gram given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, the objective is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right)$$

where:

$$p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

In other words, it learns a word feature vector by predicting its context in a window of surrounding other words preserving the order.

Each output of y_i is the unnormalized log-probability for each output word i and it is computed as:

$$y = b + U h\left(w_{t-k}, \ldots, w_{t+k}; W\right) \quad (1)$$

where U, b are the softmax parameters. h is constructed by a concatenation or an average of the word vectors extracted from W. On the other hand, the CBOW learns feature vectors by predicting a word missing from its context. For this reason CBOW results to be faster than Skip-gram but less accurate in capturing some semantic aspects of the words. Thus, in this paper we considered only the Skip-gram model as the word embedding technique for our analysis. Depending on the final application where text mining has to be applied, the resulting word embeddings have to be combined to get a document representation. Usually for classification tasks, documents are represented as a matrix where each word is represented with the learned embedding. In our case, being interested in detecting changes among documents over the years with cosine similarity, we represent a document as the average vector of words that compound it.

2.0.3. Doc2Vec

In Le and Mikolov (2014), the authors came up with an extension of Word2Vec that is directly able to learn feature vectors for documents in an unsupervised way: Doc2Vec. This algorithm learns distributed vector representations for paragraphs, regardless of their length, while learning word feature vectors. Practically, it exploits the same logic as the Word2Vec architecture but it introduces also the concept of the paragraph token. This token acts as an additional word and as memory to remember what is missing from the current context. Thus, the paragraph token has an associated paragraph vector to be learned. This vector is shared with all the windows context that the algorithm uses to learn the other word embeddings for the same document. Instead each word feature vectors that compounds the W word vector matrix is shared across the other paragraphs. Also, in this case, Doc2Vec provides two different architectures. The first, Distributed Memory version of Paragraph Vector (PV-DM) is an extension of the CBOW model in Word2Vec. The only thing that changes in compared to the Word2Vec architecture is in Eq. (1), where h is constructed from both W and D (the document matrix that contains all of the paragraph vectors). Then it takes the concatenation of W and D to predict the next word. The other version, as an extension of the Skip-Gram is the Distributed Bag of Words version of Paragraph Vector (PV-DBOW). In both architectures, for each document vectors are unique, while the W are shared. Each document is mapped to a unique vector represented by a column in matrix D and each word is also mapped to a unique vector, represented by a column in matrix W. The D and W are concatenated to predict the next word in the context.

3. Methodology

The proposed methodology can be resumed in five main steps:

- **Data collection**: We collected all the available SEC 10-K filings for the years from 1998 to 2018. For the firms where the 10-Ks were available we also collected all the monthly returns and market capitalizations.
- **Data selection**: In order to avoid bias in our dataset and to avoid extreme returns as outliers, we filtered our collection on the basis of market value and annual return
- **Text pre-processing**: Every SEC filing has been processed in order to clean it from tables, urls, HTML tags. Finally, we applied english stopwords removal and Stemming to get the root form for each word and reduce the globally size of the dictionary.
- **Models training**: Both Word2Vec and Doc2Vec have been evaluated separately to get SEC filings embeddings
- **Portfolio construction and evaluation**: We build different weighted and equally weighted portfolios using cosine similarity among the documents to compare the different models. We also evaluated the impact of combining cosine similarity with Momentum strategies

3.1. Data collection

We collected all 10-K and 10-K related (10-K, 10-K405, 10-KSB, 10-KT) SEC filings from the Loughran–McDonald dataset for years from 1999 to 2018 ([dataset] (Bill McDonald, 2019)). We selected these year range as 2018 marks the end of the second decade after the SEC published the Plain English Handbook in 1998. For these firms we collected also monthly stock data from the Center for Research in Security Prices (CRSP) using the WRDS linking tables since the SEC identifies companies only through the CIK codes. With this data we compute monthly returns and market capitalizations for all the firms. We then computed annual returns (including dividends) for the fiscal year starting on April 1st which is when the majority of US 10-Ks have already been filed.

3.2. Data selection

For each year we selected data on the basis of two policies: (1) We kept companies where the market capitalization is above than 300 million dollars. This is necessary since otherwise our results would be largely dominated by micro-caps, given that these companies encompass more than half of the publicly traded stocks while also tend to have more extreme returns (see discussion in Fama and French (2008)); (2) companies whose annual return value crosses the 1000% annual return threshold have been excluded from the analysis in order to avoid outliers created by small scalars. After the screening, we are left with 45,516 firm-year observations. Our resulting dataset is very similar to the CRSP stock universe both in value-weighted and equally-weighted returns which verifies that no bias of any kind has been introduced. This will be further discussed in the results section (see Fig. 1).

3.3. Models training

We use the corpus of all the firms' 10-Ks to train both the PV-DM and PV-DBOW Doc2Vec model with various vector dimensions and epochs. In all experiments, we use concatenation as the method to combine the vectors. The vector size, number of epochs and other hyper-parameters were selected based on the suggestions of Lau and Baldwin (2016). We end up selecting the PV-DM Doc2Vec model trained with 256 dimensions and 10 epochs as the best model. For these hyperparameters we also train a Word2Vec model and develop word embeddings for all the words in all 10-Ks. We take the average of these words in each filing and use it as the representation of the document.

3.4. Portfolio construction

In an attempt to measure the semantic differences between two consecutive financial reports of a company we experimented with various metrics. After we represented all companies' 10-Ks with vectors, we tried to measure consecutive changes by considering the cosine similarity, the euclidean distance, the Radius of Gyration and finally the Jaccard similarity. In the light of results, we chose at the end to only use the cosine similarity as it proved to be the most effective one in capturing semantic changes with neural network embeddings. Our hypothesis is that because of the embedding vectors' nature, their orientation is much more stable and reliable than their magnitude which suffers from the random initialization of the weights of the neural networks. We then compute for each of the companies the cosine similarity measures between their year-on-year 10-K fillings' embeddings generated by the three neural network embedding models discussed: PV-DM, PV-DBOW and the Word2Vec-based model. For each of these three cases we built a long-only portfolio consisting of stocks whose cosine similarity measure was higher than 0.95. In all cases, stocks are held for a year and the re-balancing occurs annually as well.

After computing the corresponding calendar time portfolios, we find that the PV-DM version of the Doc2Vec model is the best model out of the neural network embedding models we tested in capturing semantic changes in 10-Ks associated with future risk-adjusted abnormal returns. We term this strategy "Semantic Similarity" and report its performance against a respective strategy that uses instead the bag-of-words model to represent these documents.

Furthermore, driven by an effort to reduce selecting companies whose 10-Ks remain semantically unchanged but the companies themselves are facing persisting challenges, we incorporate a momentum-related criterion where the companies selected must also have had positive previous year returns ($Ret(-12,0) > 0$). This criterion attempts to exclude struggling companies whose CEO/CFO have reported the persisting challenges in the previous year's 10-K and have not removed them in the current 10-K leaving these reports semantically unchanged. It could be argued that this momentum amplified strategy, termed as "Non-struggling" attempts to select well performing stable companies that face no great risks. We, finally, report its performance against the same strategy using the bag-of-words model, while we also compute returns for the raw-momentum strategy (buying stocks with positive previous year stock returns and holding them for one year) and find that there are no statistically significant abnormal returns associated with it.

4. Results and discussion

In this section we first report the performance of various neural network embedding techniques in capturing future abnormal returns associated with 10-Ks consecutive changes. We, then report, the performance of the "Semantic Similarity" and the "Non-changers" portfolios, termed SSP and NSP respectively and compare it to the corresponding portfolios built using the BoW model instead. We term this portfolios the "BoW" portfolio and the "BoW-Mom" portfolio. We also compare our performance results with the results derived from the "Lazy Prices" analysis which uses the BoW model to capture changes in consecutive 10-Ks.

For the performance evaluation we use multi-factor alphas since the large returns found in this study might have resulted from large exposures to systematic risk factors. We investigate this hypothesis by adding to the Capital Asset Pricing Model (CAPM) the two most influential systematic risk factors: the size based factor small-minus-big (SMB) and the high-minus-low book-to-market factor (HML) (Fama & French, 1996). Furthermore, for a 5-factor analysis we also include the up-minus-down momentum factor (UMD), as well as the Pástor and Stambaugh's traded liquidity factor (PS_VWF).

The CAPM, Fama–French and 5-Factor alphas along with the corresponding betas are empirically estimated via a linear regression as:

Capital Asset Pricing Model (CAPM):

$$R_t - r_t^f = \alpha + \beta_{MKT}\left(R_t^M - r_t^f\right) + \varepsilon_t \tag{2}$$

3-Factor Fama and French Model:

$$R_t - r_t^f = \alpha + \beta_{MKT}\left(R_t^M - r_t^f\right) + \beta_{HML}HML_t + \beta_{SMB}SMB_t + \varepsilon_t \tag{3}$$

5-Factor Model:

$$R_t - r_t^f = \alpha + \beta_{MKT}\left(R_t^M - r_t^f\right) + \beta_{HML}HML_t + \beta_{SMB}SMB_t + \beta_{UMD}UMD_t + \beta_{PS\_VWF}PS\_VWF_t + \varepsilon_t \tag{4}$$

where $R_t - r_t^f$ is the excess return from each strategy, $R_t^M - r_t^f$ is the market risk premium, r_t^f is the risk free rate based on the one-month Treasury bill rate, HML_t is the difference between the high book-to-market value companies' returns minus low book-to-market value companies' returns, SMB_t is the difference between the small capitalization and large capitalization portfolios' returns and UMD_t is the momentum factor , i.e. the returns or the winners minus losers

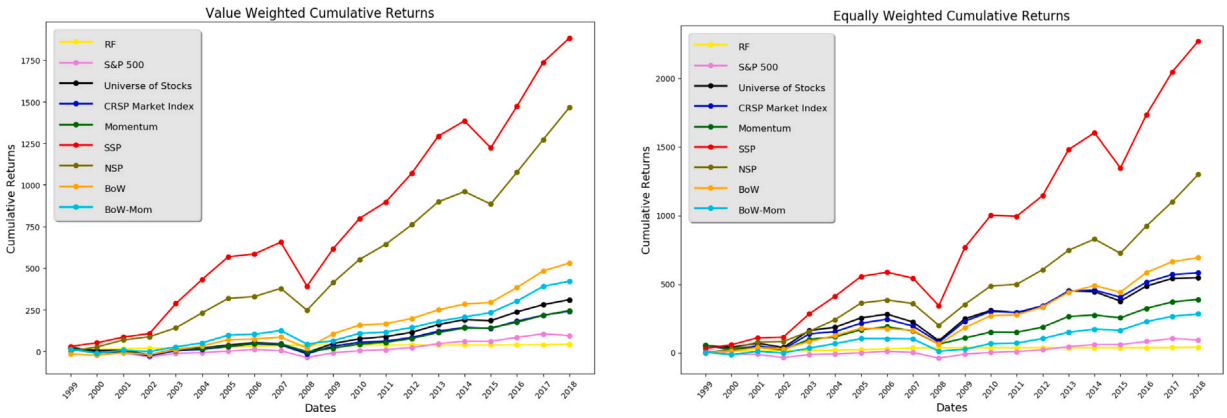

Fig. 1. Value (left) and equally (right) weighted cumulative returns from 1999–2018 withs the "Semantic-similarity" (SSP) and the "Non-struggling" (NSP) portfolios. We compare the returns with the buy-and-hold cumulative return of S&P 500, the BoW-based with and without momentum respective portfolios, our universe of available stocks portfolio (Universe of Stocks) as well as the raw momentum (Momentum) portfolio and the CRSP market index. The momentum strategy in the portfolios selects stocks with positive past year return.

portfolio based on the past 11 months and PS_VWF_t the the Pástor–Stambaugh liquidity traded factor constructed from the returns of the top decile liquidity beta portfolio minus the returns of the bottom-decile liquidity beta portfolio.[2] The parameter α is the measure of the abnormal risk-adjusted return that captures the excess return above what is expected based just on the risk of the portfolio. The five risk factors' time series as well as the risk-free rates are gathered from Whartons Research Data Services (WRDS), Fama–French Portfolios and Factors dataset.

In all portfolios we hold stocks for 12 months (from April 1st to March 30th) and re-balance every 12 months on April 1st. Note that for the value-weighted portfolio returns each stock in the portfolio is weighted by its lagged market capitalization. We, also, display the "Lazy Prices" respective results for a direct comparison.

Finally, in Fig. 1 we display the cumulative returns of the "Semantic-similarity" and "Non-struggling" strategies over the two decades. We display two figures, one with equal weighted and one with value weighted cumulative returns. The returns are compared with the buy-and-hold cumulative return of the S&P 500, the available stock universe, the CRSP stock universe and for a more direct comparison, with the corresponding BoW-based with and without momentum strategy returns. Note that all of the reported returns include dividends.

4.1. Neural network embeddings performance evaluation

Table 1 presents the performance of the PV-DM and PV-DBOW versions of Doc2Vec as well as the document embedding with Word2Vec model in terms of capturing changes associated with these future risk-adjusted abnormal returns. The table includes the equal-weighted and value-weighted annual portfolio abnormal returns as well as the statistical significance. These are computed by regressing in each case the twenty years of compounded returns on the market, the SMB and HML factors as well as the UMD and PS_VWF factors. The average number of companies selected each year is 60 for the PV-DM model, 240 for the PV-DBOW model and 850 for Word2Vec.

We see that the best performance lies with the PV-DM model. Specifically, the long-only portfolio using PV-DM model earns a large and significant abnormal return of 11% per year (t= 2.75). This proves the superiority of the PV-DM model. Furthermore, our results with the PV-DM model are mostly unaffected from controlling for the three Fama–French factors (market, size, and value). This suggests that the returns we see between the portfolio is not driven by systematic loadings on the most commonly used risk factors. Furthermore, controlling for two additional factors: momentum and liquidity, the equally-weighted portfolio earns significant abnormal return of 7.45% per year but the value-weighted portfolio return is statistically insignificant.

4.2. Main results

In an effort to improve our five-factor alphas and for further reasons discussed in Section 3.4, we added a simple momentum-related criterion to the portfolio selection. We termed this strategy as "non-struggling" and the portfolio associated with it as "non-struggling" portfolio or NSP. In this framework we select companies with very similar consecutive 10-Ks (cosine similarity of the PV-DM paragraph vectors is higher than 0.95) but also with positive previous year returns (starting from April 1st and ending March 30th) It is important to keep in mind that the momentum strategy by itself does not yield any excess returns, meaning there is no momentum premium (see Fig. 1). In fact, over the two decades under study the cumulative returns with the momentum strategy were slightly less than the available universe of stocks returns implying that the criterion has actually a small value effect, also referred to as the mean reversion effect. This means that our portfolio results are not driven by momentum effects which can also be further validated by the fact that the portfolio does not have a statistically significant momentum beta (see Table 3).

As seen in Table 2 all excess returns, 3-factor alphas and 5-factor alphas of the NSP are higher and statistically more significant compared to the "Semantic Similarity" strategy. Specifically, the value-weighted portfolio reaches a statistically significant 9.75% per year in 3-factor alpha (t= 3.24) and 8.45% per year in 5-factor alpha (t = 2.61). These are extraordinary alphas. In fact, for a comparison, a regression on the highly used, both in academia and the industry, UMD portfolio's returns for the same dates produces a smaller and less statistically significant 3-Factor annual alpha of 6.29 (t=1.84).

In the same table (Table 2) we also compare our performance with the corresponding BoW-based with and without momentum strategy performances. In these cases, the only difference is that the companies are selected if the cosine similarity of the BoW vector representations of the previous year's 10-K with the current year's is higher than 0.95. The BoW-based portfolio with and without the same momentum-related criterion show, however, no significant abnormal returns in either case. This further validates the superiority of the PV-DM model in capturing semantic changes in the 10-Ks as well as a synergistic value created by incorporating the previous year's returns

The equal-weighted and value-weighted annual portfolio abnormal returns as well as the statistical significance in Table 2 are computed by

[2] The risk-free rate is not deducted from the SMB, HML, UMD or PS_VFW portfolios since these factor returns are the difference between two portfolios (each having the risk-free rate deducted) making the risk-free rates cancel out.

Table 1
Portfolio Returns Exploiting Neural Network Embeddings: This Table reports the annual portfolio excess return, 3-Factor alphas, and 5-factor alphas (market, size, value, momentum, and liquidity) for the three long-only portfolios constructed based on the three similarity measures: Doc2Vec's two versions (PV-DM and PV-DBOW), Word2Vec average. All portfolios select companies whose cosine similarity of the vector representations is higher than 0.95. Returns are annualized and multiplied by 100. The left part of the table presents value-weighted portfolio returns and the right part presents equal-weighted portfolio returns. The t-statistics are shown below the estimates, while the statistical significance is indicated by ***, **, and * for the 1%, 5%, and 10% levels, respectively.

	Value-weighted			Equally-weighted		
Portfolio	CAPM alpha	3-Factor alpha	5-Factor alpha	CAPM alpha	3-Factor alpha	5-Factor alpha
PV-DM	11.04**	7.93**	5.87	10.94**	6.60**	7.45**
t-stat	(2.75)	(2.27)	(1.59)	(2.56)	(2.50)	(2.69)
Word2Vec	3.17	1.35	−0.98	1.26	1.25	−0.66
t-stat	(1.54)	(0.87)	(−0.84)	(0.79)	(0.89)	(−0.52)
PV-DBOW	1.34	1.27*	1.13	5.48*	2.83**	3.8***
t-stat	(1.20)	(1.71)	(1.27)	(1.90)	(2.18)	(3.22)

Table 2
"Semantic Similarity" and "Non-struggling" annual portfolio returns: This Table reports the annual portfolio excess return, 3-Factor alphas, and 5-factor alphas (market, size, value, momentum, and liquidity) of the long-only portfolio, termed "Non-struggling" portfolio (NSP) which selects companies whose previous year's returns were positive and whose cosine similarity of the Doc2Vec (PV-DM version) vector representations is higher than 0.95. SSP refers to the "Semantic-similarity" portfolio. The performance is compared to portfolios constructed using the BoW model instead. Returns are annualized and multiplied by 100. The t-statistics are shown below the estimates, and the statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	Value-weighted			Equally-weighted		
Portfolio	CAPM alpha	3-Factor alpha	5-Factor alpha	CAPM alpha	3-Factor alpha	5-Factor alpha
SSP	11.04**	7.93**	5.87	10.94**	6.60**	7.45**
t-stat	(2.75)	(2.27)	(1.59)	(2.56)	(2.50)	(2.69)
BoW	3.88	3.84	5.1	5.09	3.14	3.57
t-stat	(1.20)	(1.25)	(1.54)	(1.60)	(1.29)	(1.30)
NSP	11.11***	9.75***	8.45**	9.88**	7.40**	6.28**
t-stat	(3.30)	(3.24)	(2.61)	(2.57)	(2.45)	(2.15)
BoW-Mom	3.92	4.33	1.26	2.16	1.74	−1.35
t-stat	(1.51)	(1.56)	(0.44)	(0.70)	(0.52)	(0.71)

regressing in each case the twenty years of compounded returns, while the average number of companies selected each year is 40 for the NSP, 190 for the Bow-based without momentum portfolio and 125 for the BoW-based with momentum portfolio

For the portfolio with the best performance, the NSP portfolio, we also report in Table 3 all the factor loadings derived from the time-series regressions using the capital asset pricing model (CAPM), 3-factor and 5-factor models. These loadings are measures of the exposure to the market, size, value, momentum and liquidity risks. We observe statistically significant very small betas which suggest much lower risk as well as statistically significant exposure to the HML factor which shows that our strategy has a value tilt. These observations show that our strategy avoids high beta, high growth stocks while it also selects the least struggling value stocks. This could mean that these companies have moats, i.e. sustainable competitive advantages protecting them from external threats such as rivals or industry disruption. It is extraordinary to see information derived from text to relate to the Fama and French value premium. The rest of the factor loadings, SMB, UMD and PS_VWF are statistically not significant.

Table 3
Regression of 3 Factor and 5 Factor model with the "Non-struggling" portfolio: This table reports the factor exposure of the long-only "Non-struggling" portfolio. This portfolio selects companies whose previous year's returns were positive and whose cosine similarity of the Doc2Vec (PV-DM version) vector representations is higher than 0.95. Returns are annualized and multiplied by 100. The t-statistics are shown underneath the estimates, while the statistical significance is indicated by ***, **, and * for the 1%, 5%, and 10% levels, respectively.

	Value-weighted		Equally-weighted	
Factors	3-Factor	5-Factor	3-Factor	5-Factor
Intercept (α)	9.75***	8.45**	7.40**	6.28**
t-stat	(3.24)	(2.61)	(2.45)	(2.15)
MKTRF	0.48***	0.45**	0.61***	0.59***
t-stat	(3.30)	(2.66)	(4.23)	(3.34)
SMB	0.04	−0.19	0.31	0.10
t-stat	(0.14)	(−0.62)	(1.08)	(0.30)
HML	0.45***	0.34**	0.61***	0.49**
t-stat	(2.96)	(1.74)	(3.94)	(2.48)
UMD	–	−0.12	–	−0.12
t-stat		(−0.50)		(−0.54)
PS_VWF	–	0.49*	–	0.46
t-stat		(1.84)		(1.69)

4.3. Comparison with the Lazy Prices paper's results

Before we compare our results with the "Lazy Prices" results it is important to note several differences between our portfolios and the "Lazy Prices" portfolios: (1) The period under study is from 1995 to 2014 while in this paper we study the 1999–2018 period, (2) the holding period of the portfolio is 9 months compared to our 12-month holding period for our portfolios (3) "Lazy Prices" takes into consideration firms with "off-cycle" fiscal year-ends (firms whose 10-Ks are reported after April 1st) so they are invested all year long too, (4) the Q5 portfolio even though a long-only portfolio, it selects the quintile of companies with the least year-on-year changes on 10-Ks compared to a threshold being used in our model, (5) the Q5-Q1 portfolio is a long-short portfolio; no such portfolio has been constructed in this paper, meaning in our portfolios we do not go short. We believe that the comparison is fair as all the portfolios are invested all year long and the difference between the period under study does not affect the alphas. In fact, our strategies have lower transaction costs compared to "Lazy Prices" which selects a much larger proportion of the market while also have no shorting costs. Looking at Table 4, in a direct comparison of the portfolios which only use 10-Ks, do not use previous stock returns and do not go short stocks, the Semantic Similarity strategy considerably outperforms "Lazy Prices" in terms of value and equal-weighted 3-factor and 5-factor alphas (compare with Table 1). In a

Table 4
Lazy Prices's 10-K annualized Portfolio Returns: This table reports Lazy Prices' annualized portfolio excess return, 3-Factor alphas, and 5-factor alphas (market, size, value, momentum, and liquidity). Returns are annualized and multiplied by 100 and the similarity measure used is the cosine similarity measure. Q1 and Q5 represent the quintiles of firms with the least and most similarity correspondingly between documents this year and last year. Q5-Q1 refers to the long-short portfolio which goes long the Q5 and short the Q1 whereas Q1 refers to the portfolio that goes only long non-changers. The t-statistics are shown below the estimates, while the statistical significance is indicated by ***, **, and * for the 1%, 5%, and 10% levels, respectively.

	Value-weighted			Equally-weighted		
Portfolio	CAPM alpha	3-Factor alpha	5-Factor alpha	CAPM alpha	3-Factor alpha	5-Factor alpha
Q5	12***	5.28**	5.16**	11.52***	2.88**	2.76**
t-stat	(3)	(2.78)	(2.81)	(3.05)	(2.76)	(2.7)
Q5-Q1	7.68***	8.88***	8.16***	1.92	2.88***	2.28**
t-stat	(3.55)	(4.17)	(3.77)	(1.5)	(2.82)	(2.24)

more general comparison of the best strategies, compared to the long-short "Lazy Prices" portfolio, the equal-weighted alphas again in the Semantic Similarity strategy are significantly higher, whereas the value-weighted alphas slightly outperform only when we add the momentum criterion in the "Non-struggling" strategy (Table 2). Specifically the PV-DM Doc2Vec-based portfolio earns value-weighted 3-Factor and 5-Factor annual alphas of up to 8.45 and 9.75 whereas the Q5 ("Lazy prices" long-only portfolio) earns 5.16% and 5.28% respectively. This implies an outperformance of about 400 bps in annual excess returns.

4.4. Cumulative returns

Finally, Fig. 1 plots the value and equally weighted cumulative returns for the various portfolios that were constructed. Specifically, it plots the cumulative returns for the "Semantic Similarity" and "Non-struggling" strategies as well as the BoW-based with and without momentum-related criterion portfolios, the whole universe of stocks under consideration portfolio as well as the raw-momentum (Momentum) portfolio and the CRSP market index. For each of these portfolios we plot one chart with the value weighted and one with the equally weighted cumulative returns. In both charts we also add the cumulative returns of the S&P500 and the risk free cumulative returns for comparison. The first thing to notice is that the available universe of stocks is a representative data set with no survivor-biases or other biases of any sort as the returns do not deviate from the CRSP market index. Second, the momentum strategy by itself (buying stocks with positive prior year returns and holding them for the next year) by itself does not present any excess returns whatsoever. This shows that "Non-struggling" results are not driven by the momentum effect rather by the changes in year-on-year 10-K-s and the synergistic value that is created.

The final and main thing to notice in Fig. 1 is the historical performance of both the "Semantic-similarity" and the "Non-struggling" portfolios, SSP and NSP relative to the S&P 500 benchmark, the available universe of stocks and the raw-momentum portfolio. Over a 20-year backtest, these two strategies exhibit significant outperformance to these benchmarks. During this period, $10,000 invested with the SSP at the end of 1999 would have yielded over $200,000 in 2018 compared to only $41,000 for the whole available universe of stocks ($64,820 for the equally-weighted) and $29,233 for the S&P 500. Additionally, $10,000 would have yielded over $140,000 for the NSP, compared to only $35,000 for the value-weighted raw-momentum strategy and $48,900 for the equally-weighted. These results further validate our models' measures performance in capturing changes in 10-K-s associated with future abnormal returns. Another thing to notice is that even though the SSP cumulative returns are larger than the NSP, the 3-factor and 5-factor alphas are higher for the NSP. This is due to the fact that the NSP carries less risk as the abnormal returns are more consistent over time and the occasional drawdowns are smaller.

5. Conclusion

Measuring modifications and semantic changes from the previous year 10-Ks is challenging because the disclosures are qualitative. Even though our measures are not perfect, they are a step forward in understanding and quantifying these hard to identify changes.

We can assert, in light of our results, that the PV-DM version of the Doc2Vec model outperforms in capturing semantic changes associated with future abnormal returns in year-on-year 10-Ks the more widely used state-of-the-art bag-of-words (BoW) as well as the PV-DBOW version of Doc2Vec and the average of the Word2Vec embeddings. This was expected since treating words and phrases as discrete symbols fails to take into account the word order and the semantics of the words, while it also suffers from frequent near orthogonality due to its high dimensional sparse representation. We, also, found that the PV-DBOW performs slightly better than the BoW model. Previous year returns proved to be a strong contributor to abnormal future returns associated with these changes in year-on-year 10-Ks. Specifically, a portfolio that selects companies whose cosine similarity of the year-on-year PV-DM Doc2Vec representations is higher than 0.95 and the previous year stock return is positive earns statistically significant three-factor and five-factor alphas up to 10% per year. The main limitations of our approach are related to the computational time required to train these kind of models and to the occasional changes in the companies' executives that should be taken into account to better understand the nature of the changes in the financial reports. Both issues are objects for further analysis that we plan to present in our future works. Our measures are applicable to lots of other cases in which the disclosure is narrative, but the content is unrestricted, the timing is routine, such as CEO letters to shareholders, proxy statements, earnings press releases, and the prepared part of earnings conference calls.

CRediT authorship contribution statement

George Adosoglou: Conceptualization, Methodology, Formal analysis, Writing. **Gianfranco Lombardo:** Conceptualization, Methodology, Software, Writing. **Panos M. Pardalos:** Supervision of the research and the writing of the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Andriosopoulos, D., Doumpos, M., Pardalos, P. M., & Zopounidis, C. (2019). Computational approaches and data analytics in financial services: A literature review. *The Journal of the Operational Research Society*, *70*(10), 1581–1599. http://dx.doi.org/10.1080/01605682.2019.1595193, https://doi.org/10.1080/01605682.2019.1595193.

Bandiera, O., Prat, A., Hansen, S., & Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, *128*(4), 1325–1369. http://dx.doi.org/10.1086/705331.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, *3*(Feb), 1137–1155.

Bill McDonald (2019). *Stage one 10-x parse files*. Data retrieved from World Development Indicators, https://sraf.nd.edu/data, [dataset].

Brown, S. V., & Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year md&a modifications. *Journal of Accounting Research*, *49*(2), 309–346. http://dx.doi.org/10.1111/j.1475-679X.2010.00396.x.

Cerchiello, P., Nicola, G., Ronnqvist, S., & Sarlin, P. (2017). Deep learning bank distress from news and numerical financial data. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3292485.

Cohen, L., Axelson, U., Barberis, N., Chalmers, J., Diether, K., Dimerci, I., Engelberg, J., Gurun, U., Hoberg, G., Huang, X., Jones, H., Kelly, B., Karpoff, J., Kiku, D., Ledesma, P., Lou, D., Manela, A., Maug, E., Merrill, C., ... Cohen, L. (2019). *Lazy prices 2019*.

Day, M. Y., & Lee, C. C. (2016). Deep learning for financial sentiment analysis on finance news providers. In *Proceedings of the 2016 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2016 (No. 1)* (pp. 1127–1134). http://dx.doi.org/10.1109/ASONAM.2016.7752381.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Doc2vec. In *Proceedings of the 24th international conference on World Wide Web - WWW '15 companion (Vol. 32)* (pp. 29–30). arXiv:1405.4053. https://cs.stanford.edu/{~}quocle/paragraph{_}vector.pdf{%}0Ahttp://dl.acm.org/citation.cfm?doid=2740908.2742760.

Doumpos, M., Zopounidis, C., & Pardalos, P. (2012). *Financial decision making using computational intelligence*. Springer US, https://www.springer.com/gp/book/9781461437727.

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. http://dx.doi.org/10.1016/j.jacceco.2017.07.002, https://doi.org/10.1016/j.jacceco.2017.07.002.

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, *51*(1), 55–84. http://dx.doi.org/10.1111/j.1540-6261.1996.tb05202.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb05202.x.

Fama, E. F., & French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, *63*(4), 1653–1678. http://dx.doi.org/10.1111/j.1540-6261.2008.01371.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2008.01371.x.

Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications*, *42*(22), 9001–9011. http://dx.doi.org/10.1016/j.eswa.2015.07.058, http://dx.doi.org/10.1016/j.eswa.2015.07.058.

Jagric, T., Bojnec, S., & Jagric, V. (2015). Optimized spiral spherical self-organizing map approach to sector analysis–the case of banking. *Expert Systems with Applications*, *42*(13), 5531–5540.

Jing, K., Xu, J., & He, B. (2019). A survey on neural network language models. arXiv preprint arXiv:1906.03591.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. http://dx.doi.org/10.1016/j.eswa.2014.08.004.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. (pp. 78–86). http://dx.doi.org/10.18653/v1/w16-1609, http://arxiv.org/abs/1607.05368.

Laughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65. http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning (pp. 1188–1196)*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2–3), 221–247. http://dx.doi.org/10.1016/j.jacceco.2008.02.003.

Li, F. (2011). Do stock market investors understand the risk sentiment of corporate annual reports?. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.898181.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., & Camón-Herrero, J. B. (2011). Financial news semantic search engine. *Expert Systems with Applications*, *38*(12), 15565–15572. http://dx.doi.org/10.1016/j.eswa.2011.06.003, http://dx.doi.org/10.1016/j.eswa.2011.06.003.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743–758. http://dx.doi.org/10.1016/j.ejor.2018.10.024, https://doi.org/10.1016/j.ejor.2018.10.024.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nizer, P. S., & Nievola, J. C. (2012). Predicting published news effect in the Brazilian stock market. *Expert Systems with Applications*, *39*(12), 10674–10680. http://dx.doi.org/10.1016/j.eswa.2012.02.162, http://dx.doi.org/10.1016/j.eswa.2012.02.162.

Peng, Y., & Jiang, H. (2016). Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In *2016 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL HLT 2016 - proceedings of the conference* (pp. 374–379). http://arxiv.org/abs/arXiv:1506.07220v1.

Rawte, V., Gupta, A., & Zaki, M. J. (2018). Analysis of year-over-year changes in risk factors disclosure in 10-k filings. In *Proceedings of the 4th international workshop on data science for macro-modeling, DSMM 2018 - in conjunction with the ACM SIGMOD/PODS conference* (pp. 2–5). http://dx.doi.org/10.1145/3220547.3220555.

Salton, G., Wong, A., & Yang, C. S. (1975). Vector space model for automatic indexing. information retrieval and language processing. *Communications of the ACM*, *18*(11), 613–620.

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, *5*(1), http://dx.doi.org/10.1186/s40537-017-0111-6.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. http://dx.doi.org/10.1111/j.1540-6261.2007.01232.x.

Tsai, M. F., Wang, C. J., & Chien, P. C. (2016). Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems*, *7*(3), http://dx.doi.org/10.1145/2948072.

Van De Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, *42*(11), 4999–5010. http://dx.doi.org/10.1016/j.eswa.2015.02.007.

Wang, S., Xu, K., Liu, L., Fang, B., Liao, S., & Wang, H. (2011). An ontology based framework for mining dependence relationships between news and financial instruments. *Expert Systems with Applications*, *38*(10), 12044–12050. http://dx.doi.org/10.1016/j.eswa.2011.01.148.

Zopounidis, C., Doumpos, M., & Pardalos, P. M. (2010). *Handbook of financial engineering*, Springer Science & Business Media, https://www.springer.com/gp/book/9780387766812.