9-2016

Textual analysis and machine leaning: Crack unstructured data in finance and accounting

Li GUO

Singapore Management University, liguo.2014@pbs.smu.edu.sg

Feng SHI

University of International Business and Economics

Jun TU

Singapore Management University, tujun@smu.edu.sg

DOI: https://doi.org/10.1016/j.jfds.2017.02.001

Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 2 (2016) 153−170
http://www.keaipublishing.com/en/journals/jfds/

Textual analysis and machine leaning: Crack unstructured data in finance and accounting☆

Li Guo [a], Feng Shi [b], Jun Tu [a,*]

[a] *Singapore Management University, Singapore*
[b] *University of International Business and Economics, China*

Received 12 December 2016; revised 9 January 2017; accepted 6 February 2017
Available online 11 March 2017

Abstract

In finance and accounting, relative to quantitative methods traditionally used, textual analysis becomes popular recently despite of its substantially less precise manner. In an overview of the literature, we describe various methods used in textual analysis, especially machine learning. By comparing their classification performance, we find that neural network outperforms many other machine learning techniques in classifying news category. Moreover, we highlight that there are many challenges left for future development of textual analysis, such as identifying multiple objects within one single document.

JEL classification: C45; G00; G02; G12; G14; G17

Keywords: Machine learning; Textual analysis; Finance; Accounting; Media news; Sentiment; Information

1. Introduction

Text analytic, also referred to text mining or data mining, is the process of deriving valuable information from text. In accounting and finance, firm specific news inflows, conference calls and Securities and Exchange Commission (SEC) filings, provide ample resources for applying this technique. Since textual analysis is still an emerging area in accounting and finance, the classification algorithm is still not that well developed. In this paper, we review those popular methods, especially for machine learning techniques that are commonly used to generate the sentiment series. We also compare forecast performance of each algorithm to have a better understand their advantage and disadvantages. On top of that, we discuss research challenges and potential research direction in terms of application of this

☆ We are grateful to Wolfgang Karl Härdle, Weiping Li (the editor), Changyun Wang, seminar participants at China Securities Regulatory Commission, Humboldt-Universitt zu Berlin and participants at the Conference on (Statistical) Machine Learning and Its Applications for their very helpful comments. Financial support from a research grant of Sim Kee Boon Institute for Financial Economics for Tu is gratefully acknowledged.

* Corresponding author. Lee Kong Chian School of Business, Singapore Management University, Singapore 178899, Singapore.
E-mail addresses: liguo.2014@smu.edu.sg (L. Guo), shi01fg@163.com (F. Shi), tujun@smu.edu.sg (J. Tu).
Peer review under responsibility of China Science Publishing & Media Ltd.

http://dx.doi.org/10.1016/j.jfds.2017.02.001

technique in finance and accounting. The object of the paper is to understand the basic idea of textual analysis, the advantage and disadvantage of each classification method and how textual analysis contributes to the accounting and finance studies.

Frazier et al (1980) is the first paper to introduce content analysis for quantitatively evaluation of the narrative data based on individual words frequency. After splitting the text into words, tagging word, counting word frequency, factor analysis is applied to further analysis.[1] After that, accounting and finance scholars become actively to quantify the text information. Based on Harvard dictionary, Tetlock (2007) employs a principal components analysis to extract the most important semantic component from the (77*77) variance−covariance matrix of the categories.[2] He finds that high media pessimism can predict downward pressure on market prices. Antweiler et al (2004) study the influence of messages posted in Yahoo Finance Forum and Raging Bull across 45 companies in the Dow Jones and find that financial messages can predict index volatility.[3] Das and Chen (2007) extract small investor sentiment from web, classify the message with a "fuzzy" method and verify that sentiment is based on stock movement.[4] Li (2008) applies Fog index to measure the readability and finds companies with easy readable annual report tend to have high subsequent earnings.[5]

Indeed, we find a couple of survey papers about textual analysis application in finance studies. Li (2010) provides details about manual-based examples of textual analysis and also suggests some potential research topics in early days.[6] Kearney and Liu (2014) provide a survey paper about most recent literature on textual sentiment.[7] Das (2014) provides useful tips for researchers who are new to this topic.[8] It also provides some code and basic tools used for textual analysis. The latest survey paper, Loughran and McDonald (2016) underscore the methodological tripwires for those approaching this relatively new technique.[9] They emphasize the importance of transparency in transforming text into quantitative measures and also the importance of replicability in the less structured methods used in textual analysis.

In what follows we add value beyond simply offering an updated literature review. Instead, we provide a deeper understanding on the algorithm of textual analysis technique by comparing their classification performance. Among various techniques, we find neural network outperforms many other machine learning techniques in classifying news into categories. In addition, we also highlight the challenges to identify the different objects within the same document and the importance to understand the real meaning of trained news scores.

2. Sources of finance-related unstructured data

Unstructured Data refer to information which does not have a pre-defined data model or is not organized in a pre-defined manner. Although unstructured data is usually text heavy and difficult to analyze, many researchers use unstructured data to extract sentiment, construct sentiment index and predict return. Many studies in accounting and finance have analyzed various sources of finance-related unstructured data including mandatory filings and disclosures (e.g., 10-Ks, 10-Qs, 8-Ks annual reports, IPO prospectuses, RNS, etc), earning announcements and other press releases, conference calls (management presentation and Q&A sections) and investor road show presentations, financial media articles (e.g. WSJ, DJNS, FT, newswire service, etc.), analyst reports and research notes, regulatory announcements (e.g., SEC litigation releases), macro and sentiment news (e.g., Federal Open Market Committee minutes), internet message boards, social networks (e.g. Seeking Alpha http://seekingalpha.com).

3. General process of textual analysis

A general process of textual analysis includes three steps, namely, harvest text, clean and parse text, and analyze text.

In terms of harvesting text, some studies collect data from forum or web, such as Yahoo finance forum and twitter. Alternatively, researchers also collect data from financial database, such as Thomson Reuter News Database, news paper database, etc. These data mostly stored in a format file such as txt, xml and pdf which are easy for processing.

After collecting text information, we need to preprocess text to some specific data format. This is because most text is created and stored in a way that only humans can understand, and is not easy for computer to process. As a result, preprocessing text is required for unstructural data, including the regularity, removing all taggers and stop words and putting plain text into a word vector.

Finally, researchers are able to find text information based on different techniques. Some researchers count words based on word list, more positive word number representing positive sentiment. Other researchers train software to classify text or recognize pattern with machine learning method, such as nave byes, multiple linear regression, support vector machine and neural network. Fig. 1 shows the most popular classification methodologies employed by accounting and finance researchers.

4. Review of textual analysis in finance literature

4.1. Readability measure

Readability refers to degree that a given group of people find certain reading matter compelling and comprehensive (McLaughlin, 1969).[10] In finance and accounting literature, we find multiple readability measures (Table 1):

Among them, Fog Index is the most widely used one. Li (2008) is the first paper to examine the link between annual report readability and firm performance.[5] In that paper, the author defines Fog Index as a function of two variables: average sentence length and complex words (percentage of words with more than two syllables). Namely, they define Fog Index = 0.4 (average number of words per sentence + percent of complex words). According to Li (2008), Fog Index reflects the number of years of education needed to understand the text on a first reading.[5] FOG $\geq$ 18 means the text is unreadable, 14−18 means difficult, 12−14 means ideal, 10−12 means acceptable and 8−10 means childish can also understand.

Based on this readability measure, Li (2008) finds that firms with lower reported earnings tend to have annual reports that are harder to read (i.e., high Fog Index values or high word counts).[5] This finding is consistent with the point that firms with poor earnings tend to have more text and longer sentences to explain their situation to investors. Following Li (2008), Biddle (2009)[11] finds that firms which have high reporting quality often have greater capital investment efficiency. Miller (2010) finds that investors feel hard to process less readable annual reports so they trade fewer shares of firms whose annual reports has more words and higher Fog Index values.[12] Lawrence (2013) finds that number of words and Fog index in the annual report are related to retail investors stock holding.[13] Lundholm (2014) find that foreign firms need to encourage US investors to invest them, so their public document are easy to read.[14] Most recently, De Franco (2015) finds that companies with more readable analysts' reports often have more trade volume over three days after report date.[15]

However, the key issue in readability is how “readability” is defined. Loughran and McDonald (2014) empirically demonstrate that the Fog Index is a poorly specified readability measure when applied to business documents. For example, words like “financial”, “company”, “operations”, “management” and “customers” are usually well

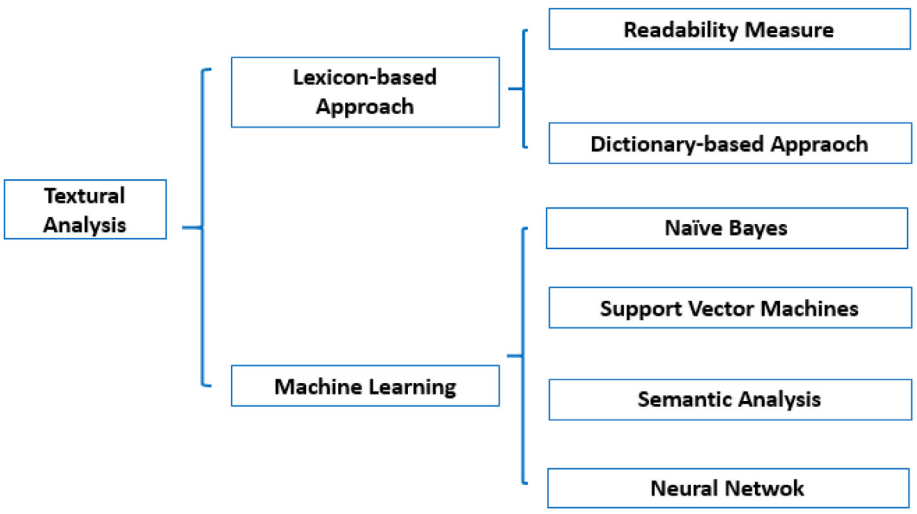

Fig. 1. General method for textual analysis.

Table 1
Readability measure.

Measurement	Formula
Flesch Kincaid	0.39 * (number of words/number of sentences + 11.8 * number of syllables/number of words−15.59
LIX	(number of words/number of sentences) + 100 * (number of words over 6 letters)/(number of sentences)
RIX	Number of words with 7 characters or more/number of sentences
Fog	0.4 * (number of words per sentence + percent complex words)
ARI	4.71 * (number of characters/number of sentences) + 0.5 * (number of words/number of sentences)−21.43
SMOG	1.043 * sqrt (30 * number of words more than two syllables/number of sentences) + 3.1291

understood by investors but they are defined as complicated words according to Fog Index construction.[16] As a result, content analysis is required for researchers to understand different types of documents.

4.1.1. Lexicon based approach

While readability focuses on the individuals ability to comprehend a message, the methods below is to computationally explore meaning from the message. Bag-of-words is the one of the most simple and popular approaches. It has been widely used by finance and accounting researchers. Technically, a dictionary is a tabulated collection of items, each with an associated attribute. Thus current discussion should be restricted to the term word lists where we are simply creating collections of words that attempt to identify a particular attribute of a document. Armed with such lists, a researcher can count words associated with each attribute and provide a comparative measure of sentiment. Since it is also easy to program and replicate, dictionary based approach is the most widely used approach by finance studies. In general, there are two dictionaries: Havard General Inquirer [a] and Loughran and Mcdonald dictionary (Loughran & McDonald, 2011).[17]

Based on Harvard General Inquirer, Tetlock et al (2007) find pessimistic tone of Abreast of the Market in Wall Street Journals can lead to low stock return and high stock volatility.[2] He also proves that when pessimism deviation increases one-standard, Dow index in the next day will decline 8.1 basis point. MacSkassy et al (2008) document that more negative words in S&P 500 firm news is related to lower earnings after controlling for accounting information and analyst forecast.[18] Based on more than 900,000 Thomson-Reuters news, Ranjan and Heston (2015) point out that negative (positive) news tone is related with low (high) short-term stock return.[19] Other related literature also includes Kothari (2009) who find positive (negative) tone of disclosure can lead to low (high) forecast dispersion and volatility and Hanley et al (2010) who use IPO reports from 1996 to 2005 and find positive net tone is associated with low first day return.[20,21] Similarly, Cahan et al (2013) measure sentiment with dataset of over 900,000 news stories from Thomson-Reuters News Analytics and argue that daily news can only predict stock return in short-term period, but weekly news predicts stock returns in long horizon.[22]

Recently, Loughran and Mcdonald find the word list of Harvard GI is not representative for financial terms. Hence, they construct a new word list based on interpretation of word in a business context. Feldman (2010) uses LM dictionary to measure market response to MD&A news and find that positive change of their tone measure can lead to high stock returns after controlling for accruals and earnings surprises.[23] Dougal et al (2012) use LM dictionary to measure tone of Abreast of the Market column in Wall Street Journals.[24] They find a pessimistic news tone is related to a negative return in next few trading days. Similar findings are also documented by Chen (2014) and Huang (2014).[25,26] Besides, after calculating news tone of financial columns in the New York Times from 1905 to 2005, Garcia (2013) finds news tone predicts stock return especially in recession periods.[27] Solomon et al (2014) find fund flow predicts occurrence of news and even news sentiment, which suggests institutional investors are informed.[28]

Although Loughran and MacDonald (2011) improve word representative in terms of Harvard GI, the challenge of this technique remains as a result of homographs (words with different meaning, but the same spelling).[17] For example, “Firm A grant call options to managers” which is a good news while “Firm A call back inferior products.” is a bad news. In these two sentences, the word “call” has different meanings but dictionary based approach is not able to differentiate them. Furthermore, Lexicon based approach may tend to misclassify many terms that are neutral financial meaning to positive category, such as “company”, “shares” and “outstanding”. The following is an example which suggests lexicon based approach may not be a good way for sentiment classification.

[a] The current version of the Harvard Psychosociological Dictionary, is available through the GI website (see http://www.wjh.harvard.edu/inquirer/).

"Fitch may raise J.C. Penney ratings, NEW YORK, June 17 (Reuters) - Fitch Ratings Services may raise J.C. Penney Co. Inc. [JCP.N] senior unsecured notes citing improved operating trends at its department stores and potential debt reduction from the sale of its Eckerd drugstore business. Fitch said it may raise J. C. Penneys BB senior unsecured notes, it BB-plus secured bank facility and its B-plus convertible subordinated notes. The action affects $5.2 billion of debt. Penney continues to make solid progress in turning around its department stores and catalog/internet business. The segments comparable store sales increased a strong 9.4 percent in the four months ended May 2004, and have been positive for three consecutive years, said Fitch."

According to Harvard GI, we can find 3 positive words: "progress", "positive" and "make" and 4 negative words: 3 instances of "raise" and also the word "make". As a result, Lexicon based approach will assign this news articles as negative while it should be positive rating in fact. The main issue is that the word "make" appears on both positive and negative lists, which confused classification. A better way is to assign those key words different weight with respect to different meaning. Equal weight has nothing contribution to a better classification performance. In the following, we randomly select 1000 Thomson Reuters News Archive as our testing data and use both Harvard GI and Loughran and Macdonal dictionary to calculate news tone for each news article. The classification performance is shown in the following table.

The results suggest that lexicon based approach achieves a reasonable good accuracy. Based on Harvard GI, we find 161 (451) over 500 negative (positive) news is correctly classified and overall accuracy is about 61.2%. Surprisingly, we find LM dictionary improves the results a lot. Overall accuracy using LM dictionary increased by 12.9%−74.1%. Consistent with LM, negative word training performs better than positive training with 393 exactly match over 500 testing sample. That is because in LM dictionary only covers 354 positive words but 2329 negative words, hence negative news seems to be better detected by LM dictionary.

4.2. Machine learning techniques

In terms of word training, machine learning techniques are widely used in textual analysis. In finance literature, we find Naive Bayes, Support Vector Machine and Neural Network are among the most popular machine learning techniques. On top of that, semantic analysis also plays an important role in analyzing information content of financial reporting or analysts' forecasts. In the following sections, we describe the property and algorithm of each machine learning technique and also provide examples for further illustrations.

In terms of training sample used in the following examples, we merge Thomson-Reuters News Archive database and News Analytics database. News Archive database provides original news body and news ID. The News Analytics database provides news sentiment score with positive, negative and neural score ranging from [−1,1]. The sum of these 3 scores equals 1. In the meanwhile, if the positive (negative) tone is the largest among 3 scores, the news is classified as positive (negative) news, labeled as 1 (−1). In addition, to have a clean measure, we only include positive and negative news to do the training. We also only use the news articles which are related to one specific company and use news articles with more than 50 words. Finally, to have a balanced sample, we randomly select 3000 positive news

Table 2a
Confusion matrix — Harvard GI Lexicon.

True	Predict	
	Negative	Positive
Negative	161	339
Positive	49	451

Table 2b
Confusion matrix — LM Lexicon.

True	Predict	
	Negative	Positive
Negative	393	107
Positive	152	348

and 3000 negative news as training sample. We select another 500 positive news and another 500 negative news as the testing sample.

4.2.1. Naive Bayes

Among alternative approaches for word classification, Naive Bayes is the most popular one. Firstly, it is one of the oldest, most established methodologies to analyze text. Secondly, since machines, instead of humans, are reading the text for content, a large amount of data can be easily handled. Thirdly, once the rules/filters of gauging the text are established, no additional researcher subjectivity affects the measuring of tone in the business communication document (Loughran and Macdonal, 2016).[9]

The earliest use of the Nave Bayes approach in finance is Antweiler and Frank (2004). In this section, we then introduce Naive Bayes Classification following Antweiler and Frank's (2004) framework.[3]

First, let W_i denote words stream, either in a message of type T or its antitype T′. Let m be the number of occurrences of this word in type T, and m′ be the number of occurrences in anti-type T′. Further let n and n′ denote the total number of words in classes T and T′ respectively. Hence, we got posterior belief as follows:

$$P(T|W_i) = \frac{P(T|W_{i-1})P(W_i|T)}{P(T|W_{i-1})P(W_i|T) + (1 - P(T|W_{i-1}))P(W_i|T')}$$

Correspondingly, the odds ratio can be expressed in the following way:

$$\Rightarrow \text{OddsRatio} : \frac{P(T|W_i)}{1 - P(T|W_i)} = \frac{P(T|W_{i-1})}{1 - P(T|W_{i-1})} * \frac{P(W_i|T)}{P(W_i|T')}$$

For a document with N number of words, we add up its logs of odds ratios as follows:

$$\Rightarrow P(T|W_N) = P(T)exp\left[\sum_{i=1}^{N} \log\left(\frac{P(W_i|T)}{P(W_i|T')}\right)\right] = P(T)exp\left[\sum_{i=1}^{N} \log\left(\frac{m_i}{m'_i} \Big/ \frac{n_i}{n'_i}\right)\right],$$

where we have $P(T|W_0) = P(T)$. After training is completed, the input (individual message k containing words W^k) will be returned to 3 probabilities $P(c|W_N^k)$ for each of the 3 categories, $c \in \{buy, hold, sell\}$. And we classify the highest probability according to:

$$\arg\max_c P\left(c \middle| W_N^k\right), \text{where c} \in \{buy,\ hold,\ sell\}.$$

Based on this classification method, Antweiler and Frank (2004) construct 3 types of Bullishness measure:

$$\textbf{a}.\ B_t = \frac{M_t^{buy} - M_t^{sell}}{M_t^{buy} + M_t^{sell}},$$

$$\textbf{b}.\ B_t^* \equiv \log\left[\frac{1 + M_t^{buy}}{1 + M_t^{sell}}\right] \approx B_t \log\left(1 + M_t^{buy} + M_t^{sell}\right),$$

$$\textbf{c}.\ B_t^{**} \equiv M_t^{buy} - M_t^{sell} = B_t\left(M_t^{buy} + M_t^{sell}\right),$$

where M_t^c stands for number of messages of type c in a given time interval.[3] Their empirical finding suggests that high disagreement among the postings are associated with higher subsequent trading volume.

Following Antweiler and Frank's (2004) framework, there are multiple literature classifying sentiment score using Naive Bayes method.[3] For example, Li (2010) calculate average tone of the forward-looking statements (FLS) in the MD&A section of the 10-K and find it positively linked to the subsequent earnings.[6] Huang, Zang, and Zheng (2014) use a similar approach to gauge the sentiment contained in 363,952 analyst reports.[29] Their trained Bayes algorithm categorize sentences in analyst reports to 3 categories: positive, negative, and neutral. And they show that an additional positive sentence is associated with a significant impact on a firms earnings growth rate even five years after the publication of the report.

Using the Nave Bayes algorithm, Buehlmaier and Whited (2014) model the probability of a firm being financially constrained on the basis of the MD&A section of the 10-K.[30] They find that more financially constrained firms are associated with higher stock returns. Surprisingly, they find that the largest, most liquid companies are the ones most affected by financial constraint risk. Other related literature includes Purda and Skillicorn (2015) and Buehlmaier and Zechner (2013).[31,32] All studies suggest Naive Bayes is a good algorithm to extract text information.

Indeed, compared to bag-of-word approach, Naive Bayes shows its own advantages. For example, Bag-of-words depends on dictionaries. If there is no readily available dictionary that is built for the setting of corporate filings, it will achieve a poor classification performance. For example, consider the sentence: "In addition, the Company has experienced attrition of its Medicare and commercial business in 1998 and 1999 and expects additional attrition." The bag-of-words model will count 2 positivewords "expect" and "experience" based on Harvard IV4 dictionary however this sentence should be classified as negative news. Instead, Nave Bayes gives weight for a list of key words. If similar sentence appeared in training sample, Nave Bayes is able to find the correct answer. Moreover, dictionary-based approach does not take into consideration the context of a sentence. E.g., if a sentence is about cost, then increase should be treated as a negative word. While Bag-of-words simply count increase as one positive word and misclassified it as good news. On the contrary, Naive Bayes will put both increase and other related words into a word vector and jointly determine the output. Most importantly, Bag-of-words ignores prior knowledge of researchers while Naive Bayes takes prior knowledge into account. For example, most of the sentences appearing in MD&A reports are neutral. Simply counting positive and negative words according to financial dictionary will overstate the meaning of each sentence.

In the following, we apply Naive Bayes algorithm to train Thomson Reuters News Archive data. Table 3a and b shows classification performance of this method.

Our in-sample accuracy is high, which is consistent with prior studies. Antweiler and Frank (2002) use 1000 messages to train the Naive Bayes and find the in-sample of accuracy for 2 categories (Buy and Sell) is 72.3% based on their Table 2a and b.[3] Das and Chen (2007) use a set of 374 messages taken from Yahoo to train their Naive Bayes model and find the in-sample accuracy is 60.7%.[4] For out-of-sample accuracy, however, it is usually low. For instance, Das and Chen (2007) report their out-of-sample results based on 913 news messages and it achieves a 31.54% accuracy. In our test, we use 6000 training sample and take 1000 news articles for out-of-sample tests. The overall out-of-sample accuracy of naive Bayes classification is about 58.7%, which is much lower than the in-sample fitting. That is because the naive Bayes goes through each feature and training sample so that it may take some noise features into account.

Comparing to Lexicon based approach, Naive Bayes may not necessarily out-performs the lexicon method due to noise features included in the training. Under our tests, it works reasonable well as Harvard dictionary but LM dictionary seems to out-perform Naive Bayes since LM dictionary include the most relevant financial terms. But we expect Naive Bayes will be more useful for the situation which requires prior knowledge. In current training, we simply set the prior belief as 0.5 for each key word being positive/negative category.

4.2.2. Support vector machine

Recently, Support Vector Machines (SVM) receives a great attention from the machine learning community. In this section, we review foundations and properties of SVM through the graphical analysis and also introduce some empirical applications in finance study.

Consider a linearly separable binary classification problem where the training set is defined as follows:

$$\mathbf{T} = \{(x_i, y_i); \mathrm{i} = 1, \ldots, \mathrm{P}\} \tag{2.2.1}$$

Table 3a
In-sample confusion matrix — Naive Bayes.

True	Predict	
	Negative	Positive
Negative	1520	1480
Positive	309	2691

Table 3b
Out-of-sample confusion matrix — Naive Bayes.

True	Predict	
	Negative	Positive
Negative	191	309
Positive	104	396

where $x_i \in \mathbf{R}^N$ stands for features of training sample and $y_i \in \{-1, 1\}$ stands for labels or target classifications for corresponding features. The target of SVM is to find a hyper-plane in $\mathbf{R}^N$ (input space) such that features in the same class will be assigned in the same region of the input space. Indeed, without any restriction, there are infinite ways to separate the input spaces. However, SVM suggests a good classification system should provide generalization ability. For example, without restrictions, the problem is equivalent to find parameters ($\mathbf{w}$, w_0) of the hyper-plane which satisfies the set of P constraints:

$$(w^T x_i + w_0) y_i \geq 0, \ \mathrm{i} = 1, \ldots, \mathrm{P} \tag{2.2.2}$$

SVM further suggests the hyper-plane should maximize the classification margin ρ, which is defined as the minimum distance between P and the closest pattern to P:

$$\rho = \min_i D(x_i, P) \tag{2.2.3}$$

where $D(x_i, P)$ is the Euclidean distance from feature x to P. This distance is given by:

$$D(x, P) = \frac{|\mathbf{w}^T x + w_0|}{||\mathbf{w}||} \tag{2.2.4}$$

Importantly, a small value of ρ means that P is close to one or more samples and hence a higher probability that samples not included in the training set may fall in the wrong side of the classification region. The learning criteria of SVM is to find the set of parameters that maximize the classification margin ρ under the constraints. For example, the

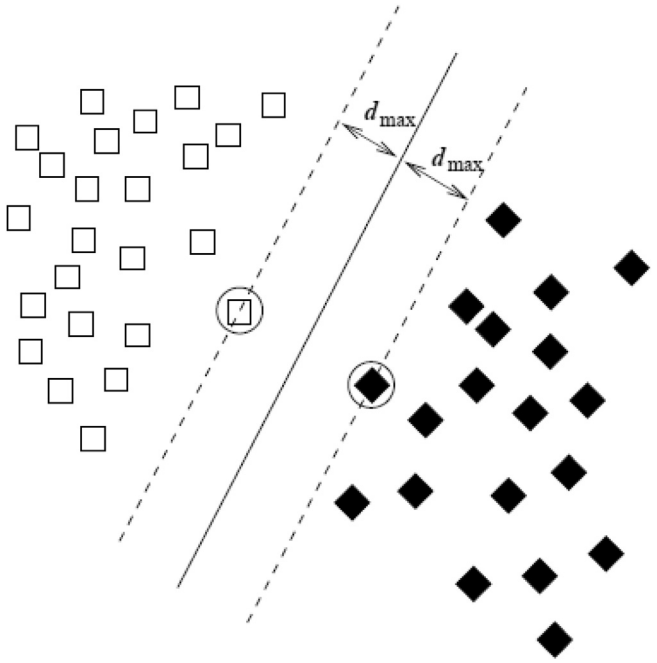

Fig. 2. Support vectors.

optimal classification hyper-plane according to this criteria for a simple bi-dimensional example is shown in Fig. 2. Note that, the exact location of P is only dependent on the support vectors closer to the hyper-plane (in Fig. 2, there are two support vectors which are located on the dashed lines).

As a result, to find optimal hyper-plane, we need to find classification margin ρ. According to (2.2.3) and (2.2.4), $||\mathbf{w}|| = \frac{1}{\rho}$ given $\min_i\{\mathbf{w}^T x_i + w_0\} = 1$. With these conditions, the optimization problem can be expressed as follows: maximize

$$f'(\mathbf{w}) = \frac{1}{||\mathbf{w}||}$$

$$\text{s.t.}(\mathbf{w}^T x_i + w_0)y_i - 1 \geq 0,\ \text{i} = 1, \ldots, \text{P}$$

Meanwhile, in empirical setting, to simplify the overall formulation, it is common to minimize the $f(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2$ instead of maximizing $f'(\mathbf{w})$. The following example shows how to find the hyper-plane in terms of textual analysis application.

Given two sentences, "Profits increased.", labeled as positive news, and "Profits decreased.", labeled as negative news, we can put them into word vector expression as shown in Table 4.

Namely, x_1, (1,0,1), stands for "Profits increased." classified as 1 while x_2, (1,1,0), stands for "Profits decreased." classified as -1. To find the hyper-plane, we first have a weight vector: $w = \{0, a, -a\}$. By substituting w into point x_1 and x_2 respectively, we will have the following equation hold:

$$\begin{cases} -a + w_0 = 1 & \text{using point } x_1, (1,0,1) \\ a + w_0 = -1 & \text{using point } x_2, (1,1,0) \end{cases}$$

$$\Rightarrow \begin{cases} a = -1 \\ w_0 = 0 \\ w = \{0, -1, 1\} \end{cases}$$

As a result, if x_1 and x_2 are regarded as two support vectors, we can write the hyper-plane for above training sample as $g(x) = -m_2 + m_3$.

Moreover, in the recent finance literature, SVM has drawn lots of attention from researchers. Manela and Moreira (2016) construct a text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal.[33] Based on this measure, they find News coverage related to wars and government policy explains most of the time variation in risk premia, consistent with time variation in rare disaster risk as a source of aggregate asset prices fluctuations. Other related work include Chen, Jeong and Wolfgang (2008) who propose a smooth support vector machines to predict default risk of firms, and Wolfgang, Moro and Hoffmann (2011) who find SVMs are capable of extracting the necessary information from financial balance sheets and then to predict the future solvency or insolvent of a company.[34,35]

Indeed, SVM has shown multiple advantages over other machine learning techniques. Here we only list out the two most important points due to space limitation. First, SVM avoids overfitting problems. For an unbalanced training sample, Naive Bayes tend to introduce noise during the learning process as it will cover all the documents. While for SVM, it only focus on the support vectors which is selected by different types of penalty functions. Second, For those linearly inseparable, SVM use kernels to map data into high dimensional space so that it becomes linearly separable

Table 4
Word vector.

Vocabulary	"profit"	"decrease"	"increase"
x_1	1	0	1
x_2	1	1	0

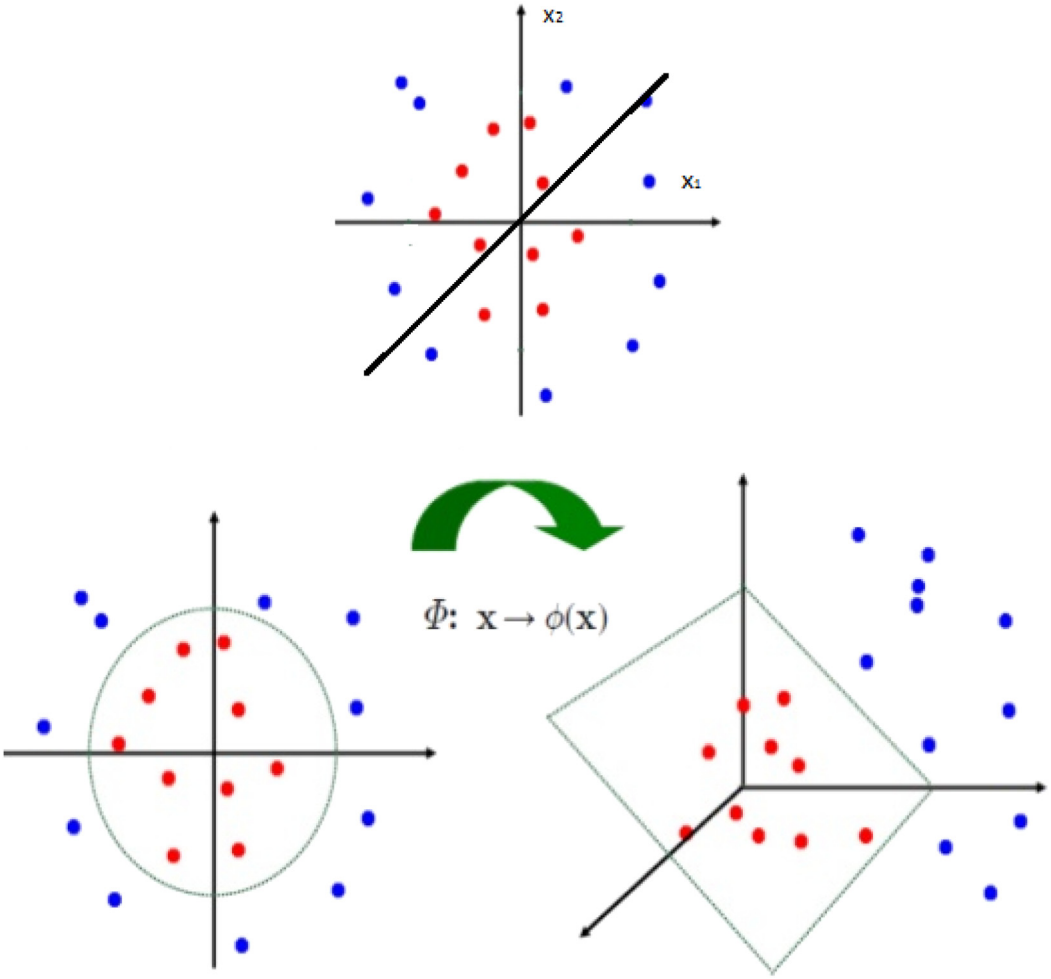

Fig. 3. Linear inseparable classification.

with hyper-plane. For example, in Fig. 3, we consider two features, x_1 and x_2, along with classification variable Y, {red, blue}. Linear regression cannot classify the points in a correct way while SVM can archive a better performance by mapping two dimension data to a three-dimension space and hence find a good hyper-plane to separate data points.

In comparison to Naive Bayes approach, we use the same news articles as training sample for SVM model. To avoid any data mining issue, we follow a general setting of SVM model. Namely, we set our loss function as $L\{y, s(X)\} = max\{0, 1 - s(X)y\}$ where s(X) = **w**'**x** and $y_i \in \{-1, 1\}$; and we set penalty function as $R(w) = 2^{-1}\sum_{i=1}^{p} w_i^2$. Above setting is consistent with Chen, Wolfgang and Elisbeth (2016). Based on this setting, we show the prediction results in Table 5a and b. Not surprisingly, SVM achieves a better out-of-sample forecast accuracy compared to Naive Bayes.

Table 5a
In-sample confusion matrix — SVM.

True	Predict	
	Negative	Positive
Negative	2091	909
Positive	172	2828

Table 5b
Out-of-sample confusion matrix — SVM.

True	Predict	
	Negative	Positive
Negative	319	181
Positive	37	463

The overall accuracy for testing sample is 78.2% which is much higher than both lexicon based approach and Naive Bayes classification. However, we find SVM seems to fit positive news better than negative news. The asymmetry results might be due to the support vectors selection — Although we have balanced positive and negative news in our training sample, only those candidates who are in the splitting edge will be considered as support vectors.

4.2.3. Neural network

There are many different types of neural networks and techniques for training them. In this section, due to the space limitation, we illustrate the key idea of neural networks via discussing a simple case: the classic back propagation neural network (BPN). For a general neural network, it consists of three types of layers (shown in Fig. 4): input layer, hidden layer and output layer. Each layer also consists of multiple units. The decision rule makes the units in the subsequent layers digest all information from the training sample by weighting all units of current layer. For example, the output of unit j in Fig. 7: $O_j = \frac{1}{1+e^{-I_j}}$, where $I_j = \sum_i w_{ij} y_i + \theta_j$. Importantly, w_{ij} and θ_j are the parameters we need to estimate. w_{ij} is the weight of information delivered from unit i in previous layer to unit j in current layer and θ_j is the bias term assigned to unit j.

Back propagation neural network (BPN) defines its own updating rules for both w_{ij} and θ_j. First, we need to find error of each unit. In output layer, the error is defined as $Err_j = O_j(1 - O_j)(T_j - O_j)$ where T_j is actual value for that category. In hidden layer: $Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk}$. Hence, we can update the weight $w_{ij} = w_{ij} + \delta w_{ij}$, where $\delta w_{ij} = (I) Err_j O_i$. I stands for learning rate. Bias term, θ_j can be updated as $\theta^* = \theta_j + \delta\theta_j$ where $\delta\theta_j = (I) Err_j$. Again, we take a simple example, "Profits increased", for illustration purpose. First, we transfer this sentence to a word vector

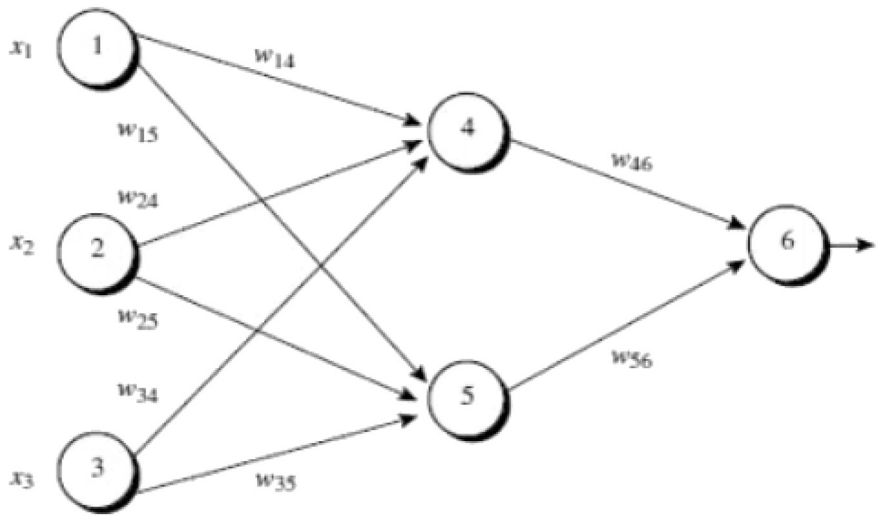

Fig. 4. Neural network structure.

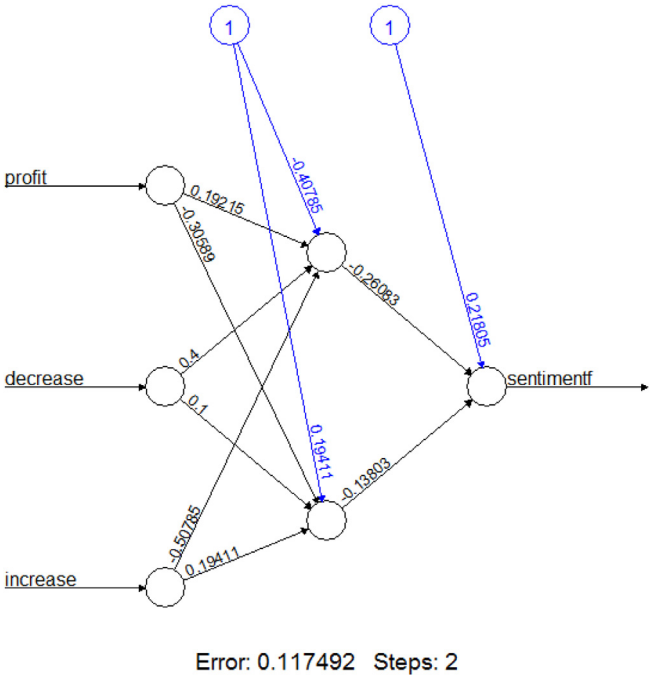

Fig. 5. Trained parameter of neural network.

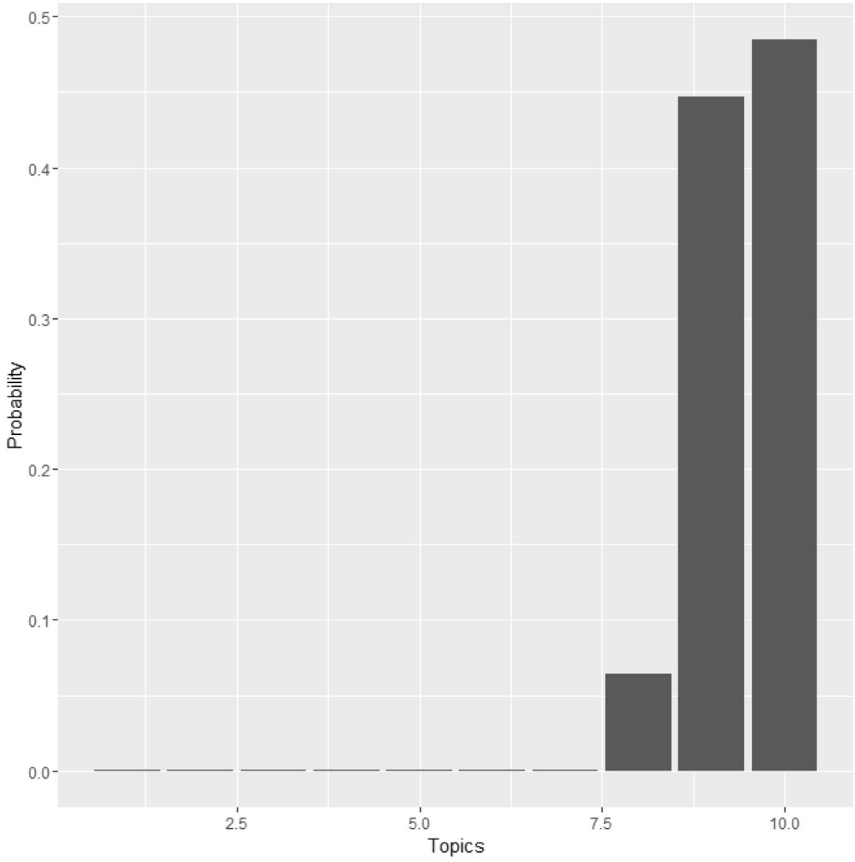

Fig. 6. Posterior topic distribution of document I.

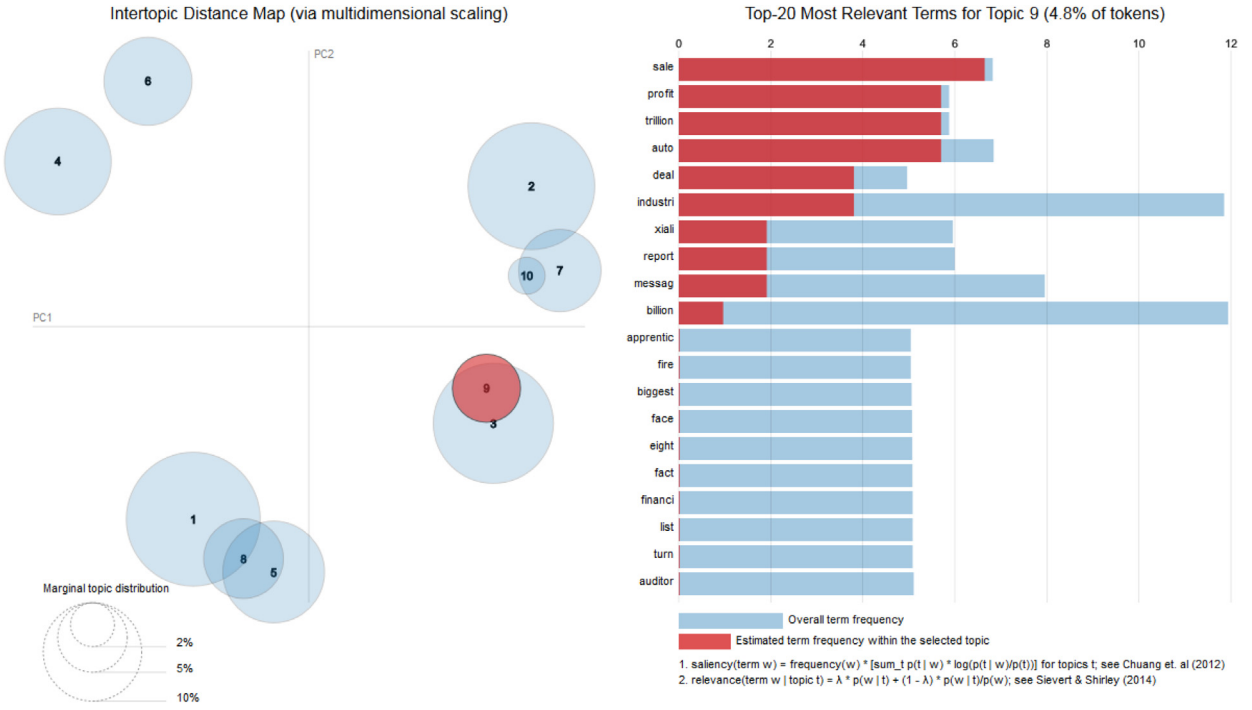

Fig. 7. Posterior word distribution of topic 9.

as what we did in SVM section: "Profits increased" $\Rightarrow [1, 0, 1]$. After that, we set the initial input, weight and bias values as shown in Table 6:

Table 6
Initial value for neural network.

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

We then calculate the net input and output in Table 7:

Table 7
Net input and output of neural network.

Unit j	Net input I_j	Output O_j
4	$0.2 + 0 - 0.5 - 0.4 = -0.7$	$1/(1 + e^{0.7}) = 0.332$
5	$-0.3 + 0 + 0.2 + 0.2 = 0.1$	$1/(1 + e^{0.1}) = 0.525$
6	$(-0.3)(0.332) - (0.2)(0.525) + 0.1 = -0.105$	$1/(1 + e^{0.105}) = 0.474$

And calculate the error for each node in Table 8:

Table 8
Back propagation error for each node.

Unit j	Err_j
6	$(0.474)(1 - 0.474)(1 - 0.474) = 0.1311$
5	$(0.525)(1 - 0.525)(0.1311)(-0.2) = -0.0065$
4	$(0.332)(1 - 0.332)(0.1311)(-0.3) = -0.0087$

Accordingly, we update the weight, $w_{46} = -0.3 + (0.9)(0.1311)(0.332) = -0.261$ with learning rate as 0.9. Under threshold value = 0.01, we will find the one step updated model as shown in Fig. 5. It reflects the structure of the trained neural network, i.e. the network topology. The plot includes the trained synaptic weights, all intercepts as well as basic information about the training process like the overall error (0.1311−0.0065 − 0.0087) and the number of steps needed to converge (step = 1 for illustration purpose).

In practice, neural network seems to be a powerful data-driven, self-adaptive, flexible computational tool that can capture nonlinear and complex underlying characteristics of any physical process with a high degree of accuracy: It can handle large amount of data sets, implicitly detect complex nonlinear relationships between dependent and independent variables and even detect all possible interactions between predictor variables. Due to high accuracy of neural network training, it is widely used in textual classifications, e.g., Thomson Reuters News Analytics. In finance study, especially in recent years, scholars start to use Thomson Reuters News Analytics data to study media effects on firm performance and return predictability. Heston and Ranjan (2014) directly compare the return predictability between Thomson Reuters News Analytics scores and other news tone measures based on different lexicon based method.[36] And they find that lexicon based approach predict short-term returns that are quickly reversed while neural network predicts larger and more persistent returns. In this case, neural network training makes news score more informative than others. On top of that, Terrence, Dmitry and Norman (2015) use institutional trading volume to predict both news inflows and sentiment score and find the institutional investors are informed.[37]

Neural network become widely accepted by finance studies due to its high accuracy of textual classification while the comparison between neural network and other machine learning techniques are not explored by financial studies. In the following, we use the same Thomson Reuters News Archive data and employ the basic neural network training model (BPN) to understand its training performance. To avoid data mining issue, we set one hidden layer with 30 units. The initial value of parameters are randomly generated from a normal distribution. Meanwhile, we set a threshold value equal to 0.01. Based on this setting, we present the prediction performance of both in sample and out of sample in Table 9a and b.

Table 9a and b indicates a surprising high accuracy of in-sample prediction — 99.85% accuracy. The out-of-sample results are also amazing — it achieves a 79.6% accuracy with balanced training for both categories. Moreover, the out-of-sample performance is much better than that of either lexicon based approach and Naive Bayes — Overall accuracy is 20% higher than Naive Bayes and 5% higher than LM dictionary. Since our training does not consider the grammar or semantic in each sentence, we expect Thomson Reuters News Analytics will perform better than general training. In this case, researches based on Thomson Reuters News Analytics data makes sense and should be paid more attention from finance scholars due to its high accuracy of information extraction.

4.2.4. Semantic analysis

Semantic analysis is a different method compared to other textual analysis. It is used to extract conceptual content and document relationships. Latent Dirichlet Allocation (LDA) is a popular mathematical approach to do the semantic analysis. It starts with bag-of-words modeling and then transform a text corpus into term-document frequency matrices. After that, it reduces the high dimensional term spaces of textual data to a lower dimensions according to authors' criteria. Within the new corpus, it generates weighted term lists for each concept or topic, produce concept or topic content weights for each document, and produce outputs that can be used to compute document relationship measures. Generally speaking, the target of LDA model is to estimate the word distribution of each topic and topic distribution for each document. In the following we introduce the process of LDA model.

First of all, LDA assumes a corpus consisting of a collection of D documents contains a fixed number of latent topics. Each document, d, is characterized by a discrete probability distribution over topics (θ_d), and each topic t, is characterized by a discrete probability distribution over words (ϕ_t). Under such framework, a document d can be generated by first sampling topics from distribution, θ_d and then sampling key words from each topic according to ϕ_t.

Table 9a
In-sample confusion matrix — neural network.

True	Predict	
	Negative	Positive
Negative	2994	6
Positive	3	2997

Table 9b
Out-of-sample confusion matrix — neural network.

True	Predict	
	Negative	Positive
Negative	403	97
Positive	107	393

Meanwhile, to estimate these two parameters, LDA takes into consideration of priors where θ_d and ϕ_t follows Dirichlet distribution. Mathematically, we have the following assumptions hold:

Choose at opic $z_{d,n} \sim \text{Multinomial}(\theta_d)$

Choose a word $w_{d,n} \sim$ from $p\left(w_{d,n}\middle|z_{d,n}, \phi_{z_{d,n}}\right)$

$p(\theta_d) \sim \text{Dirichlet}(\alpha)$

$p(\phi_t) \sim \text{Dirichlet}(\beta)$.

where θ_d is the document d probability vector of topics; $\phi_{z_{d,n}}$ is the word probability vector for topic $z_{d,n}$. Both topics $Z_{d,n}$ and words $w_{d,n}$ are discrete random variables and follow multinomial distribution. Moreover, α and β are given as prior beliefs.

Given above setting, a commonly used estimation algorithm for LDA is collapsed Gibbs sampling proposed by Griffiths and Steyvers (2006). The joint distribution can be written in the following way:

$$p(\mathbf{w},\ \mathbf{z},\ \theta,\ \Phi|(\alpha,\beta)) = \prod_{n=1}^{N_d} p\left(w_{d,n}\middle|\phi_{z_{d,n}}\right)p(z_{d,n}|\theta_d)p(\theta_d|\alpha)p(\Phi|\beta)$$

⇓ Rewrite the Joint Distribution :

$$p(\mathbf{w},\mathbf{z}|\alpha,\beta) = p(\mathbf{w}|\mathbf{z},\beta)p(\mathbf{z}|\alpha)$$

where the first factor $p(\mathbf{w}|\mathbf{z},\beta)$ generates the word, $w_{d,n}$ given topic $z_{d,n}$ according to prior word distribution β while the second factor generates topics according to prior topic distribution α. Once we have the joint distribution, we can easily calculate the posterior p($\mathbf{z}|\mathbf{w}$). Gibbs Sampling suggests to find $p(z_i = k|\mathbf{z}_{-i},\mathbf{w})$ to update parameters. According to Griffiths and Steyvers (2007):[38]

$$p(z_i = k|\mathbf{z}_{-i},\mathbf{w}) \propto \frac{n^t_{k,-i} + \beta_t}{\sum_{t=1}^{V} n^t_{k,-i} + \beta_t}\left(n^k_{d,-i} + \alpha_k\right)$$

where $n^k_{d,-i}$ stands for number of times topic k appears in document d except for current topic and $n^t_{k,-i}$ stands for number of times a word t appears in topic k except for current word. Finally the estimated parameters will be the following form:

$$\phi_{k,t} = \frac{n^t_k + \beta}{\sum_{t=1}^{V} n^t_k + V\beta}$$

$$\theta_{d,k} = \frac{n^t_d + \alpha}{\sum_{k=1}^{K} n^k_d + K\alpha}$$

For example, we apply LDA model to analyze Thomson Reuters news archives data. We randomly select 20 news articles and assume the corpus consists of 10 topics. By doing semantic analysis, we are able to find posterior word and topic distributions. For example, the following is the first news articles selected in our sample:

SEOUL, Jan 2 (Reuters) − POSCO <005490.KS>, the world's fourth-largest steel maker, posted 1.9 trillion won ($1.59 billion) in 2003 net profit, up from 1.1 trillion won profit in 2002, its CEO Lee Ku-taek said on Friday. The company's sales in 2003 also rose to 14.3 trillion won from 11.7 trillion won seen in 2002, a company spokeswoman quoted Lee as telling employees in his New Year speech. According to Reuters Research, analysts forecast POSCO would earn 2.04 trillion won in 2003 on sales of 14.01 trillion won. ($1 = 1197.3 Won).

The posterior topic distribution of this document is shown in Fig. 6. From the figure, it is obvious that topic 9 and 10 are the most relevant topics. We then further plot words distribution of topic 9 in Fig. 7 for further illustration. The key words of topic 9 suggested by LDA models are "sale", "profit", "trillion" and "deal" so on so forth. From this plot, we are able to know the first news article is about sales profit and usually the unit is trillion or billion which suggests large companies in the industry. Moreover, the left panel of our visualization presents a global view of the topic model, and answers questions how prevalent is each topic and how do the topics relate to each other. In this view, we plot the topics as circles in the two-dimensional plane whose centers are determined by computing the distance between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions, as discussed in (Sievert and Shirley, 2014).[39]

The above example suggests LDA model can be used to extracted most relevant information from each document. In terms of finance study, this technique can also be applied to analyze financial reporting or analyst reports. Huang et al (2014) provide one of the first applications of this method in accounting and finance, using the technique to examine the topical differences between conference call content and subsequent analyst reports.[26] One of advantages is that by comparing topical differences, it is clear to isolate the incremental information added by analysts while the traditional use of announcement returns cannot. Based on this setting, the author claim that analysts do provide valuable information which is beyond conference call to the market.

5. Challenges and future research

In this section, we discuss some challenges and potential research directions for future work. First of all, textual analysis extract text content from different types of documents but the nature of the content is not so clear. For example, news content is a mixed product of information and sentiment. We are not clear about whether the trained score reflects real information or just sentiment bias delivered by the journalists. Furthermore, if we are not clear about the real meaning of trained news score, then the tests of news effects on market reaction would be a joint tests. To some extend, multiple literature about news return predictability find stocks with positive (negative) news over one day have subsequent predictably high (low) returns for 1−2 days that are largely reversed, which suggests an overreaction on real information. This overreaction could be either driven by the sentiment bias of news articles itself or driven by the behavior bias from general investors. Without knowing the meaning of news score, we are not able to understand the real source that contributes to the return predictability.

Second, current machine learning model applied in finance and accounting literature is still at document level. We are not able to identify a specific object. Without extra information, the trained model or word vector cannot be applied to a general news item which has no predetermined objects. This limits the analysis to some specific resources, like 10-k files and conference calls or some specific news columns in Wall Street Journals. Moreover, for the same document, it may mention multiple objects with different purposes, however, we are not able to find an appropriate way to differentiate news score for these two objects. Recently, Thomson Reuters News Analytics data is able to handle this issue. Thomson Reuters claims that the machine assign different sentences to different objects (firms) during the preprocessing textual analysis so that each object within the same news article will receive their own news score. However, the training process for the preprocessing is not transparent which is not good for academy studies. In the future, we expect to see a clear standard to preprocess the text so that studies based on Thomson Reuters News Analytics will become more transparent to the public.

In addition, the endogeneity issues should also be considered when analyzing textual analysis based variables' effects on firm performance. Usually, those financial news or reporting coexists with many other effects which are not controlled in one regression model, especially for the information story. Meanwhile, causality issue exists when we

test the story under a contemporaneous framework. For example, the effect of disclosure readability on analyst following properties documented in Lehavy et al (2011)[40] is not established as a causality effect but association effect.

Moreover, when constructing readability measure, we should be careful about the firm-level complexity and document complexity. A complex firm with multiple business, corporate structure may produce financial reports more difficult to read solely due to the nature of their business. In this case, we need think of readability measure that tears out the firm complexity effect.

In the future, we expect textual analysis will play an important role in cracking unstructured data in finance and accounting. For example, it is desired to construct a more authoritative and extensive field-specific dictionaries. More qualitative information sources could be analyzed, such as business and political speeches, blogs, television news, videos and various social media. It is also interested to apply textual analysis to other markets like bonds, commodities and derivatives. Besides, we can also apply textual analysis to other languages.[41] For example, the German language is much more structured than English, but also suffers from syncretism, which is the case where a word form serves multiple grammatical purposes. It could also be applied to analyze management incentives and the features of the corporate textual disclosures. For example, how do ownership structure and board composition affect corporate disclosures? What are the roles of disclosures in corporate governance? It is also important to understand whether textual information contributes to the bankruptcy and fraud predictions. There is a large literature on bankruptcy and accounting fraud prediction using accounting variables and stock price data while the textual information is not well explored.

Overall, information plays a central role in how accountants document the operation of a firm and how financial markets assess value. Almost all quantitative data in this arena is contextualized by textual information which we are just now beginning to explore for deeper insights.

6. Conclusion

In finance and accounting, compared with the quantitative methods traditionally used, the textual analysis method or machine learning in general is substantially less precise. Nevertheless, textual analysis and machine learning technic has become more and more popular over time recently due to maybe the increasing need to handle tons of texts from firm specific news inflows, conference calls and Securities and Exchange Commission (SEC) filings, etc.

In this study, we review the growing literature on textual analysis in accounting and finance, discuss the most commonly applied methods, and compare their classification performance. Based on our analysis, neural network seems to outperform many other machine learning techniques in classifying news into categories. In addition, we highlight that there are many challenges left for future development of textual analysis, such as identifying multiple objects within one single document and separating out sentiment and information parts in the news tones.

References

1. Ingram RW, Frazier KB. Environmental performance and corporate disclosure. *J Account Res*. 1980;18(2):614−622.
2. Tetlock PC. Giving content to investor sentiment: the role of media in the stock market. *J Finance*. 2007;62(3):1139−1168.
3. Antweiler W, Frank M. Is all that talk just noise? The information content of internet stock message boards. *J Finance*. 2004;59(3):1259−1294.
4. Das SR, Chen MY. Yahoo! for amazon: sentiment extraction from small talk on the web. *Manag Sci*. 2007;53(9):1375−1388.
5. Li F. Annual report readability, current earnings, and earnings persistence. *J Account Econ*. 2008;45(2):221−247.
6. Li F. The information content of forward looking statements in corporate filings − a naive Bayesian machine learning approach. *J Account Res*. 2010;48(5):1049−1102.
7. Kearney C, Liu S. Textual sentiment in finance: a survey of methods and models. *Int Rev Financ Anal*. 2014;33(3):171−185.
8. Das SR. Text and context: language analytics in finance. *Found Trends Finance*. 2014;8(3):145−261.
9. Loughran T, McDonald B. Textual analysis in accounting and finance: a survey. *J Account Res*. 2016;54(4):1187−1230.
10. Laughlin G. Smog grading: a new readability formula. *J Read*. 1969;12(8):639−646.
11. Biddle GC, Hilary G, Verdi RS. How does financial reporting quality relate to investment efficiency? *J Account Econ*. 2009;48(2):112−131.
12. Miller BP. The effects of reporting complexity on small and large investor trading. *Account Rev*. 2010;85(6):2107−2143.
13. Lawrence A. Individual investors and financial disclosure. *J Account Econ*. 2013;56(1):130−147.
14. Lundholm RJ, Rogo R, Zhang JL. Restoring the tower of Babel: how foreign firms communicate with US investors. *Account Rev*. 2014;89(4):1453−1485.
15. Franco G, Hope OK, Vyas D, Zhou Y. Analyst report readability. *Contemp Account Res*. 2015;32(1):76−104.
16. Loughran T, Mcdonald B. Measuring readability in financial disclosures. *J Finance*. 2014;69(4):1643−1671.

17. Loughran T, Mcdonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10Ks. *J Finance*. 2011;41(1):57−59.
18. Tetlock PC, Saar-tsechansky M. More than words: quantifying language to measure firms fundamentals. *J Finance*. 2008;63(3):1437−1467.
19. Heston SL, Sinha NR. *News versus Sentiment: predicting Stock Returns from News Stories*; 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2311310.
20. Kothari SP, Li X, Short JE. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. *Account Rev*. 2009;84(5):1639−1670.
21. Hanley KW, Hoberg G. The information content of IPO prospectuses. *Rev Financ Stud*. 2010;23(7):2821−2864.
22. Cahan S, Chen C, Nguyen NH. *Media sentiment, Investor Sentiment, and Stock Price Sensitivity to Earnings*. Victoria University of Wellington; 2013.
23. Feldman R, Govindaraj S, Livnat J, Segal B. Managements tone change, post earnings announcement drift and accruals. *Rev Account Stud*. 2010;15(4):915−953.
24. Dougal C, Engelberg J, Garcia D, Parsons CA. Journalists and the stock market. *Rev Financ Stud*. 2012;25(3):639−679.
25. Chen H, De P, Hu YJ, Hwang BH. Wisdom of crowds: the value of stock opinions transmitted through social media. *Rev Financ Stud*. 2014;27(5):1367−1403.
26. Huang AH, Zang AY, Zheng R. Evidence on the information content of text in analyst reports. *Account Rev*. 2014;89(6):2151−2180.
27. Garcia D. Sentiment during recessions. *J Finance*. 2013;68(3):1267−1300.
28. Solomon DH, Soltes E, Sosyura D. Winners in the spotlight: media coverage of fund holdings as a driver of flows. *J Financ Econ*. 2014;113(1):53−72.
29. Huang A, Lehavy R, Zang A, Zheng R. *Analyst Information Discovery and Information Interpretation Roles: a Topic Modeling Approach*; 2014. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2409482.
30. Buehlmaier M, Whited T. *Looking for Risk in Words: a Narrative Approach to Measuring the Pricing Implications of Finance Constraints*; 2015. https://hub.hku.hk/handle/10722/213634.
31. Purda L, Skillicorn D. Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. *Contemp Account Res*. 2015;32(3):1193−1223.
32. Buehlmaier M, Zechner J. *Slow-moving Real Information in Merger Arbitrage*; 2014. http://hub.hku.hk/handle/10722/201724.
33. Manela A, Moreira A. News implied volatility and disaster concerns. *J Financ Econ*. 2017;123(1):137−162.
34. Härdle WK, Hoffmann L, Moro R. Learning machines supporting bankruptcy prediction. *Stat Tools Finance Insur*. 2011:225−250.
35. Härdle WK, Lee YJ, Schfer D, Yeh YR. *The Default Risk of Firms Examined with Smooth Support Vector Machines*; 2008. https://ssrn.com/abstract=2894311.
36. Heston SL, Sinha NR. *News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns*; 2013. http://finpko.faculty.ku.edu/myssi/FIN938/Heston%20%26%20Sinha_News%20vs%20Sentiment_WP_2014.pdf.
37. Hendershott T, Livdan D, Schrhoff N. Are institutions informed about news? *J Financ Econ*. 2015;117(2):249−287.
38. Griffiths TL, Steyvers M, Tenenbaum JB. Topics in semantic representation. *Psychol Rev*. 2007;114(2):211−244.
39. Sievert C, Shirley KE. LDAvis: a method for visualizing and interpreting topics. In: *Workshop on Interactive Language Learning*. 2014:63−70.
40. Lehavy R, Li F, Merkley K. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Account Rev*. 2011;86(3):1087−1115.
41. Tsarfaty R, Seddah D, Bler S, Nivre J. Parsing morphologically rich languages: introduction to the special issue. *Comput Linguist*. 2013;39(1):15−22.