Discovering Finance Keywords via Continuous-Space Language Models

MING-FENG TSAI, National Chengchi University
CHUAN-JU WANG, Academia Sinica
PO-CHUAN CHIEN, National Taiwan University

The growing amount of public financial data makes it increasingly important to learn how to discover valuable information for financial decision making. This article proposes an approach to discovering financial keywords from a large number of financial reports. In particular, we apply the continuous bag-of-words (CBOW) model, a well-known continuous-space language model, to the textual information in 10-K financial reports to discover new finance keywords. In order to capture word meanings to better locate financial terms, we also present a novel technique to incorporate syntactic information into the CBOW model. Experimental results on four prediction tasks using the discovered keywords demonstrate that our approach is effective for discovering predictability keywords for post-event volatility, stock volatility, abnormal trading volume, and excess return predictions. We also analyze the discovered keywords that attest to the ability of the proposed method to capture both syntactic and contextual information between words. This shows the success of this method when applied to the field of finance.

CCS Concepts: • **Information systems** → **Data analytics**; • **Computing methodologies** → **Natural language processing**

Additional Key Words and Phrases: Continuous-space language model, text mining, finance

ACM Reference Format:
Ming-Feng Tsai, Chuan-Ju Wang, and Po-Chuan Chien. 2016. Discovering finance keywords via continuous-space language models. ACM Trans. Manage. Inf. Syst. 7, 3, Article 7 (August 2016), 17 pages.
DOI: http://dx.doi.org/10.1145/2948072

1. INTRODUCTION

In much finance and accounting research, textual analysis has been used to examine the sentiment of corporate reports, financial news, investor message boards, and messages on social media [Antweiler and Frank 2004; Devitt and Ahmad 2007; Li 2008; Tetlock et al. 2008; Loughran and McDonald 2011; Uhl 2011; Wang et al. 2013; Malo et al. 2014; Nuij et al. 2014; Qiu et al. 2014]. The empirical results to date suggest that sentiment words are effective in measuring the sentiment of documents or have significant correlations with financial variables such as stock prices and their volatilities.

This research was partially supported by the Ministry of Science and Technology in Taiwan under the grants MOST 102-2420-H-004-052-MY2, 102-2221-E-004-006, and 102-2221-E-845-002-MY3.
Authors' addresses: M.-F. Tsai is with the Department of Computer Science & Program in Digital Content and Technology, National Chengchi University, No. 64, Sec. 2, Zhinan Road, Taipei 116, Taiwan; email: mftsai@nccu.edu.tw; C.-J. Wang is the corresponding author and is with the Research Center for Information Technology Innovation, Academia Sinica, No. 128 Academia Road, Sec. 2, Taipei 115, Taiwan. She was with Department of Computer Science, University of Taipei; email: cjwang@citi.sinica.edu.tw; P.-C. Chien is with the Department of Information Management, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City, 106, Taiwan. He was with the Department of Computer Science, National Chengchi University; email: r04725015@ntu.edu.tw.

DOI: http://dx.doi.org/10.1145/2948072

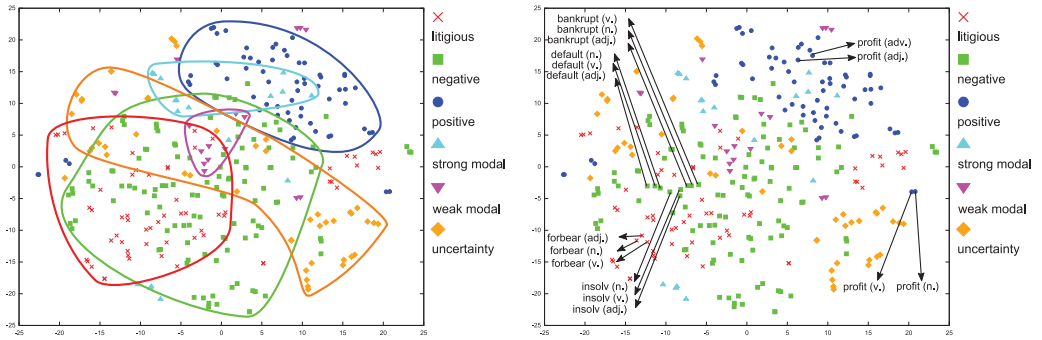

Fig. 1. 2-D visualization of the six financial sentiment word lists based on the learned-word representations.

For sentiment analysis, the lexicon is one of the most important and common resources, and usually has a great impact on results and the corresponding analyses [Feldman 2013]. Traditionally, tasks in mining financial texts utilize general-sentiment lexicons such as the Harvard Psychosociological Dictionary. However, such general-purpose sentiment lexicons often misclassify common words in financial texts [Loughran and McDonald 2011]; for example, words such as *board* and *vice* are in the Harvard Negative Dictionary, although these words often do no more than name a *board* of directors or a company's *vice*-presidents. Thus, Loughran and McDonald [2011] proposed a finance-specific sentiment lexicon that consists of six lists: negative, positive, litigious, uncertainty, weak modal, and strong modal. This finance-specific lexicon has been widely adopted and studied in the field of financial text analysis (e.g., Price et al. [2012], Garcia [2013], Li et al. [2013], and Wang et al. [2013]). However, the lexicon is constructed via a statistical approach based on a simple language model, and neither word syntax nor contextual information is considered during its construction. Considering the importance of the lexicon and the limitations stated earlier, in this article, we attempt to discover new keywords from the lexicon by the use of state-of-the-art continuous-space language models, which have recently yielded outstanding results across a variety of natural-language processing (NLP) tasks.

Continuous-space language models [Bengio et al. 2003; Schwenk 2007; Mikolov et al. 2010] are neural-network language models in which words are represented as high-dimensional real-valued vectors. These vector representations have recently demonstrated promising results across various tasks [Schwenk 2007; Collobert and Weston 2008; Glorot et al. 2011; Socher et al. 2011; Weston et al. 2011] because of their superiority in capturing syntactic and contextual regularities in language.

In this article, we use the words in the word list of Loughran and McDonald [2011] as seed words and apply the continuous bag-of-words (CBOW) model [Mikolov et al. 2013], a well-known continuous-space language model, on the textual information of 10-K financial reports for discovering new finance keywords. In particular, we use the continuous-vector representations of the words to find similar words based on the distance between their representations. We also present a novel technique to incorporate syntactic information into the CBOW model. To the best of our knowledge, this is the first work to incorporate syntactic information into continuous-space language models by adding part-of-speech (POS) tags to the words trained by the CBOW model.

Figure 1 visualizes the words of the six financial sentiment lists based on the learned word representations.[1] As shown in Figure 1(a), words in the same list generally

[1]Each word representation is a 200-dimensional real-valued vector generated by the CBOW model and transformed into two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE), a technique

aggregate into one group. There is little overlap between the two groups of positive words and negative words, whereas the groups of litigious and uncertainty words overlap substantially with the negative group. This means that, in finance, litigious and uncertainty words are usually associated with negative meanings. Furthermore, from Figure 1(b), we observe that the three words with similar meanings in finance, *default*, *insolvent*, and *bankruptcy*, are close to each other based on the learned-word representations. These locality phenomena attest to the CBOW model's ability to capture contextual regularities in finance reports; therefore, it seems applicable for discovering new financial keywords. Figure 1(b) also shows that, although word representations of the same word but with different POS tags are sometimes very close to each other (e.g., *profit* (n.) and *profit* (v.)), they are sometimes distant from each other (e.g., *profit* (n.) and *profit* (adj.)); this underscores the necessity of our proposed method of taking into account POS tags when expanding financial keywords.

For the experiments, we collected a corpus from the annual SEC[2]-mandated financial reports on Form 10-K, which contains 40,708 reports with 125,370 unique terms from the years 1996 to 2013. In addition, we also calculate four financial measures: post-event volatility, stock volatility, abnormal trading volume, and excess return for each report associated with a company. In our experiments, there are 3,911 financial sentiment seed words for keyword discovery. For comparison, we implement two baseline methods: random keyword expansion and expansion by latent Dirichlet allocation (LDA). Following keyword discovery, in order to verify the quality of the discovered keywords, we conduct the regression tasks of predicting the four financial measures just presented using only textual information in the reports. Experimental results show that for both of the tasks of post-event and stock volatility predictions, the regression models trained on keywords discovered by our methods are consistently better than those trained on the original seed words only and the two baselines. For the prediction tasks of abnormal trading volumes and excess returns, the results of using our keyword discovery methods are slightly better than the baselines. We also provide analyses of the discovered keywords that attest to the ability of the proposed method to capture both syntactic and contextual information between words; this shows the success of this method when applied to the field of finance.

The remainder of this article is organized as follows. We first briefly describe continuous-space language models and the CBOW model, in particular, in Section 2. We then describe how to discover new keywords via the CBOW model and how to incorporate syntactic information into the expansion process in Section 3. Section 4 presents the formulation of the financial risk prediction problem. In Section 5, we then detail our experimental settings and experimental results, after which we provide additional discussion and analysis. We present our conclusions in Section 6.

2. CONTINUOUS-SPACE LANGUAGE MODELS

Continuous-space language models are neural-network language models (NNLMs) in which words are represented as high-dimensional real-valued vectors. These representations have recently demonstrated promising results in NLP applications such as machine translation and speech recognition. These models have a long history of development [Rumelhart et al. 1986; Elman 1990]. Recently, Bengio et al. [2003] proposed a popular feed-forward NNLM to jointly learn word representations and a statistical language model. Following this, a recurrent NNLM was proposed that overcame some of the limitations of the feed-forward NNLM [Mikolov et al. 2010]. However, training these models is computationally expensive. To address this, Mikolov et al. [2013]

for dimensionality reduction that is particularly well suited for the visualization of high-dimensional data. The webpage for the visualization of the word representations is available at http://clip.csie.org/10K/FinDict.

[2]Securities and Exchange Commission.

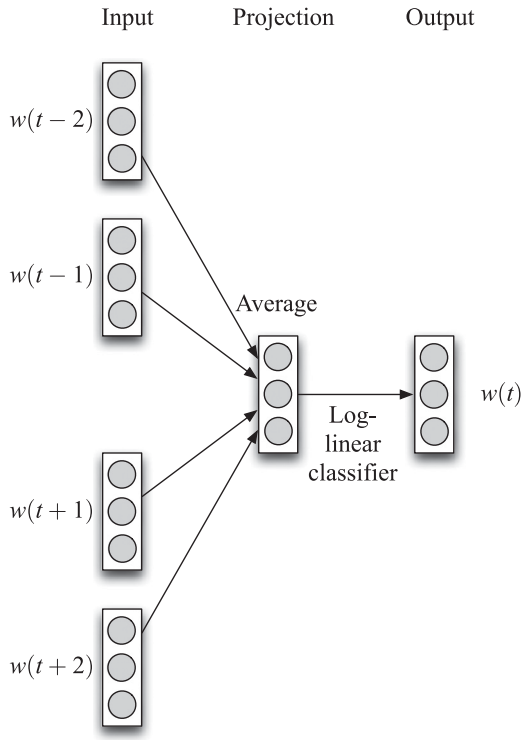

Fig. 2. CBOW model architecture.

proposed two models for computing continuous-space word representations: a CBOW mdoel and a continuous skip-gram model. Their experimental results show that these models yield large accuracy improvements at much lower computational costs. In particular, the skip-gram model takes much longer to train than the CBOW model, and usually works best with small amounts of training data. Since the dataset that we used for this study contains 40,708 reports and 125,370 unique words ranging from 1996 to 2013, which is quite a large amount of data, we mainly adopted the CBOW model[3] to compute the word representations.[4]

Unlike standard bag-of-words models, the CBOW model uses continuous context representations. Figure 2 depicts the CBOW model architecture, demonstrating four important concepts:

(1) The model predicts a word given the immediately preceding and following words.
(2) Each word is represented by a k-dimensional real-valued vector (in the figure, a 3-dimensional vector).
(3) The vectors of context words are averaged in the projection layer.
(4) A log-linear classifier is built on the averaged vector to obtain the resulting word.

[3]In these experiments, we used the word2vec toolkit (https://code.google.com/p/word2vec/).
[4]We also used the skip-gram model to find word representations, and found significant (over 70%) overlap between the words expanded by the skip-gram and CBOW models.

3. KEYWORD EXPANSION VIA FINANCIAL SENTIMENT LEXICON

3.1. Financial Sentiment Lexicon

The sentiment lexicon is one of the most important resources for sentiment analysis. Loughran and McDonald [2011] state that a general-purpose sentiment lexicon (e.g., the Harvard Psychosociological Dictionary) misclassifies common words in financial texts. Therefore, in this article, we use a finance-specific lexicon that consists of the six word lists provided by Loughran and McDonald [2011] as seed words to discover new keywords. The six lists are as follows:[5]

(1) Fin-Neg: negative business terminologies (e.g., *deficit*, *default*).
(2) Fin-Pos: positive business terminologies (e.g., *achieve*, *profit*).
(3) Fin-Unc: words denoting uncertainty, with emphasis on the general notion of imprecision rather than exclusively focusing on risk (e.g., *appear*, *doubt*).
(4) Fin-Lit: words reflecting a propensity for legal contest or, per our label, litigiousness (e.g., *amend*, *forbear*).
(5) MW-Strong (strong modal words): words expressing strong levels of confidence (e.g., *always*, *must*).
(6) MW-Weak (weak modal words): words expressing weak levels of confidence (e.g., *could*, *might*).

3.2. Simple Keyword Expansion

In this section, we introduce a simple keyword expansion method for discovering financial keywords from seed words. We first use a large collection of financial reports as training texts to learn continuous-vector representations of all the words. After training, each word is represented by a k-dimensional vector (called the representation of this word). Then, each word in the financial sentiment lexicon is used as a seed word to obtain those words with the highest N cosine distance values with respect to the seed word (called the top-N words for the seed word). The cosine distance values are calculated by the learned word vector representations. Given a pair of learned word vectors $\mathbf{x}$ and $\mathbf{y}$, the cosine distance $\cos(\theta)$ is represented by using a dot product and magnitude as follows:

$$\cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} x_i^2} \sqrt{\sum_{i=1}^{k} y_i^2}},$$

where x_i and y_i are the components of vector $\mathbf{x}$ and $\mathbf{y}$, respectively. Finally, we combine the top-N words for each seed word to construct an expanded keyword list for the financial sentiment lexicon.

3.3. Keyword Expansion with Syntactic Information

In addition to simple keyword expansion, we now incorporate syntactic information via POS tag attachments. We consider POS tags because, in general, the same word but with different POS tags usually yields different lists of top-N words. Table I shows the top five words for the word *default* with different POS tags (noun and adjective). As shown in Table I, only one of the words (i.e., *default* (v.)) overlaps in the two lists. In order to illustrate this phenomenon, we list some example sentences with the context of the word *default*, with noun and adjective POS tags, in Table II. Observed from the table, the context words of *default* with different POS tags are quite different from each other; since the CBOW model predicts a word given the immediately preceding

[5]The lists are available at http://www.nd.edu/~mcdonald/Word_Lists.html.

Table I. Top Five Words for the Word "Default"

default (NN)		default (JJ)	
Word	Cosine distance	Word	Cosine distance
trigger (adj.)	0.625872	delinqu (adj.)	0.611095
default (v.)	0.619979	securit (v.)	0.585107
insolv (n.)	0.600379	default (v.)	0.569897
uncur (adj.)	0.587429	delinqu (n.)	0.556338
nonpay (adj.)	0.583594	foreclos (n.)	0.549486

Table II. Example Sentences in the 10-K Reports for the Word Stem "Default" with Different POS Tags

default (NN)	. . . credit risk protection through credit **default** swap transactions . . .
	. . . Net premiums earned from credit **default** swaps, which are earned as written, . . .
	. . . credit spreads, and, for one credit **default** swap, erosion in the risk layers . . .
default (JJ)	. . . , the Company charges **defaulted** loans and related fees to bad debt expense . . .
	Accrued service charges related to **defaulted** loans are deducted from service charge revenue . . .
	. . . all manufacturers supplying a **defaulting** dealer are generally invoked regardless . . .

Table III. POS Tag Definitions and Tag Replacement Rules

After replacement	Before replacement
JJ	JJ (adjective)
	JJR (adjective, comparative)
	JJS (adjective, superlative)
NN	NN (noun, singular or mass)
	NNS (noun, plural)
	NNP (proper noun, singular)
	NNPS (proper noun, plural)
PRP	PRP (personal pronoun)
	PRP$ (possessive pronoun)
RB	RB (adverb)
	RBR (adverb, comparative)
	RBS (adverb, superlative)
VB	VB (verb, base form)
	VBD (verb, past tense)
	VBG (verb, gerund or present participle)
	VBN (verb, past participle)
	VBP (verb, non-3rd person singular present)
	VBZ (verb, 3rd person singular present)
WP	WP (wh-pronoun)
	WP$ (possessive wh-pronoun)

and following words, as shown in Figure 2, words with different surrounding contexts result in dissimilar word representations, thereby leading to different expanded words.

When considering syntactic information, we attach the POS tag to each word in the training texts first;[6] the POS tag to a word is attached using an underscore (e.g., default_VB). For simplicity, we represent some groups of POS tags with a single tag using tag replacement; for example, the tags JJR (adjective, comparative) and JJS (adjective, superlative) are replaced with JJ (adjective). The replacement rules are listed in Table III. Words from the sentiment lexicon with the four types of POS tags (i.e., JJ, NN, VB, RB) are taken as the seed words with which we discover new keywords.

[6]In this article, we use the most common POS tag scheme, the Penn Treebank POS tags.

Note that we only choose these four types of POS tags due to the rarity of the other two tags, PRP and WP, associated with the seed words in our corpus. The remaining steps are similar to those in simple keyword expansion.

4. FINANCIAL MEASURE PREDICTION

4.1. Four Financial Measures

In this article, we consider four financial measures for our prediction tasks: post-event volatility, stock price volatility, abnormal trading volume, and excess return. Note that, in the following experiments, only trading days are considered; for the calculation of the four measures, the starting point of the time period is the filing date of the corresponding financial report.

4.1.1. Post-Event Return Volatility. By following the definition in Loughran and McDonald [2011], the post-event return volatility is the root-mean square error from a Fama-French three-factor model for days [6, 252], with a minimum of 60 daily observations [Fama and French 1993].

4.1.2. Volatility. Volatility, a common measure for financial risk, is a measure of the variation of prices of a stock over a period of time. Let S_t be the price of a stock at time t. Holding the stock from time $t-1$ to time t leads to a simple return: $R_t = S_t/S_{t-1}-1$ [Tsay 2005]. The volatility of returns for a stock from time $t-n$ to t can thus be defined as

$$v_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t}(R_i - \bar{R})^2}{n}},$$

where $\bar{R} = \sum_{i=t-n}^{t} R_i/(n+1)$. In the experiments, the target value is the 12mo after the report's filing date volatility for each company.

4.1.3. Abnormal Trading Volume. By following the definition in Loughran and McDonald [2011], the abnormal trading volume is defined as the average volume of the 4d event window [0, 3],[7] in which volume is standardized based on its mean and standard deviation from days [−65, −6] of the so-called pre-event window. The pre-event window is chosen in accordance with previous evidence of information leakage [Aktas et al. 2007]; [−65,−6] is a common setting in empirical finance.

4.1.4. Excess Return. In this article, the excess return is defined as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the 4d event window [Loughran and McDonald 2011].

4.2. Regression Task

Given a collection of financial reports $D = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_n}\}$, in which each $\mathbf{d_i} \in \mathbb{R}^p$ and is associated with a company c_i, we aim to predict the four financial measures stated earlier of each company c_i (denoted by v_i). This prediction problem can be defined as

$$\hat{v}_i = f(\mathbf{d_i}; \mathbf{w}). \tag{1}$$

The goal is to learn a p-dimensional vector $\mathbf{w}$ from the training data $T = \{(\mathbf{d}_i, v_i) | \mathbf{d}_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$. In this article, we use support vector regression (SVR) [Drucker et al. 1997] to train the regression model. More details about SVR can be found in Schölkopf and Smola [2001].

[7]Recall that, here, the event is the report's filing date.

Table IV. Corpora Statistics

Year	# of documents	# of unique terms
1996	1,203	18,115
1997	1,705	22,262
1998	1,940	25,192
1999	1,971	26,118
2000	1,884	25,731
2001	1,825	26,290
2002	2,023	31,900
2003	2,866	42,561
2004	2,861	44,149
2005	2,698	45,570
2006	2,564	43,754
2007	2,495	40,905
2008	2,509	41,361
2009	2,567	42,369
2010	2,439	42,378
2011	2,416	42,835
2012	2,406	42,430
2013	2,336	42,928
Total	40,708	125,370

5. EXPERIMENTS

In this section, we first describe the details of our experimental settings. Then, we report and analyze the experimental results of the models trained via various keyword discovery techniques.

5.1. Experimental Settings

5.1.1. Dataset and Preprocessing. For our experiments, we built a corpus of the annual SEC-mandated financial reports on Form 10-K, containing 40,708 reports over 18 years (from 1996 to 2013), along with four financial measures: post-event volatility, stock volatility, abnormal trading volume, and excess return.[8] A Form 10-K is an annual report required by the SEC, which provides a comprehensive overview of a company's financial performance and includes audited financial statements. In this article, we use the corpus presented earlier, along with 3,911 financial sentiment keywords, to train the continuous-word representations and conduct our prediction experiments. Similar to Kogan et al. [2009], we use only Section 7 "management's discussion and analysis of financial conditions and results of operations" (MD&A) in our experiments because this section contains the most important forward-looking statements about the companies. Table IV lists the statistics of the documents and unique terms in the reports for each year.

All documents and the six financial-sentiment word lists were stemmed using the Porter stemmer [Porter 1980]; some stop words were also removed. Table V shows the statistics before and after stemming in each of the six financial sentiment lexicons. Note that as some words occurred in more than one word list, the number of unique stemmed sentiment words is 1,549 rather than 1,673.

For the four prediction tasks, the ground truth was defined in Section 4.1. The stock prices and trading volumes were obtained from the Center for Research in Security Prices (CRSP) US Stocks Database. Note that, for both of the post-event and stock

[8]The dataset is available at http://clip.csie.org/10K/data.

Table V. Financial Lexicon Statistics

Dictionary	# of words	# of stemmed words
Fin-Neg	2,329	901
Fin-Pos	354	151
Fin-Unc	297	136
Fin-Lit	886	449
MW-Strong	19	17
MW-Weak	26	19
Total	3,911	1,673

Table VI. An Example of the Top-10 Expanded Words with Different Expansion Methods

EXP-LDA on default		EXP-SIM on default		EXP-SYN on default (n.)	
Word	Cosine distance	Word	Cosine distance	Word	Cosine distance
pik	0.8665	nonpay	0.6224	trigger (adj.)	0.6259
lien	0.8588	trigger	0.5994	default (v.)	0.6200
lender	0.8435	represent	0.5688	insolv (n.)	0.6004
tranch	0.8389	uncur	0.5386	uncur (n.)	0.5874
represent	0.8300	unmatur	0.5291	nonpay (n.)	0.5836
subordin	0.8248	insolv	0.5247	unmatur (adj.)	0.5716
libor	0.7869	cure	0.5058	trigger (n.)	0.5554
repay	0.7841	waiv	0.4879	trigger (v.)	0.5511
princip	0.7796	indentur	0.4793	acceler (n.)	0.5502
refinanc	0.7786	obligor	0.4665	nonpay (n.)	0.5487

volatilities, we work in the logarithm domain for the predicted variables. This is a common practice in finance [Kogan et al. 2009].

5.1.2. Keyword Expansion. In our experiments, we used Section 7 (MD&A) of the corpus (which contains 125,370 unique terms in total) as training texts for the word2vec tool to learn the continuous-vector representations of words.[9] The context (window) size for the CBOW model was set to 5, and the dimensionality of the word vectors was set to 200.

For simple expansion (denoted as EXP-SIM hereafter), we used all 1,673 stemmed sentiment words (Table V) as seed words to discover new keywords via the learned-word vector representations. For syntactic information (EXP-SYN), we used NLTK[10] to attach the POS tag to each word in the training texts (recall that the POS tag is attached to a word with an underscore). For both EXP-SIM and EXP-SYN, we used the top-20 expanded words for each seed word (see Table VI) to construct the discovered keyword lists. Note that a robustness analysis for different numbers of words, N, was conducted in later experiments to justify choosing the top-20 words for keyword expansion. In total, for EXP-SIM, the discovered list contained 10,258 unique words, and for EXP-SYN, the list had 16,467 unique words. (Recall that the original sentiment dictionary contained only 1,549 unique words.)

For comparison purposes, we also constructed keyword lists using two other baseline methods: keyword expansion via LDA (EXP-LDA) and random keyword expansion (EXP-RAN). With LDA keyword expansion, each seed word in the financial sentiment lexicon was used to obtain those words with the highest 20 cosine distance values with respect to the seed word; this was calculated using their LDA-learned topic distributions with $k = 200$ topics (see Table VI). With random keyword expansion, for each seed word,

[9]The pretrained word vectors are available at http://clip.csie.org/10K/data.
[10]http://www.nltk.org/.

Table VII. Statistics of Discovered Keyword Lists

List	# of unique words	# of overlapped words	% of overlapped words
SEN	1,549	416	11.28%
EXP-LDA	16,654	1,268	33.39%
EXP-SIM	10,258	1,497	40.60%
EXP-SYN	16,467	1,929	52.32%

we randomly selected 20 words from the vocabulary and then combined these words into a keyword list. Note that all expanded keyword lists (EXP-SIM, EXP-SYN, EXP-LDA, and EXP-RAN) also included the 1,549 original sentiment words from the six financial sentiment word lists.

5.1.3. Word Features. In the experiments, we used the bag-of-words model and followed Kogan et al. [2009] in adopting the logP feature to represent the 10-K reports. Given a document $\mathbf{d}$, the word feature LOG1P was calculated as

$$\texttt{LOG1P} = \log(1 + \texttt{TC}(t, \mathbf{d})),$$

where $\texttt{TC}(t, \mathbf{d})$ denotes the term count of t in $\mathbf{d}$.

5.1.4. Evaluation Metrics. We measured regression performance using the mean squared error (MSE) between the predicted values and the true values, which is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(v_i - \hat{v}_i)^2,$$

where n is the number of tested companies.

5.2. Experimental Results

Table VII tabulates the statistics of the expanded keyword lists. In the table, SEN denotes the original financial sentiment lexicon. The second column denotes the number of unique words in each list. We observe that EXP-LDA had the largest number of unique words among the three EXP-LDA, EXP-SIM, and EXP-SYN expansion methods; this shows that the words found by LDA are far more varied than the other two methods. In addition, we collected three online financial dictionaries[11] to validate the expanded keyword lists. Note that we considered only unigrams in the three dictionaries; there were a total of 3,687 unique words after Porter stemming. The third and fourth columns, respectively, denote the number and the percentage of the words that also appeared in each of the keyword lists. From Table VII, it is worth mentioning that, although only 10,258 words were found via simple keyword expansion, this coverage rate was still higher than that of LDA keyword expansion (which contained 16,654 unique words). Moreover, by adding syntactic information, the coverage rate increased from 40.60% to 52.32%. In addition, Table VI lists the top-10 found words for the word *default* with the three different expansion methods. As shown in the table, night out of the top-10 words expanded via LDA (EXP-LDA) are different from those via our methods (EXP-SIM) and (EXP-SYN).[12] These interesting statistics and the example suggest that

[11]http://www.investopedia.com/terms/a/?page=1, http://www.investinganswers.com/financial-dictionary, http://www.investorwords.com.

[12]Note that *pik* stands for the abbreviation for Payment-In-Kind (PIK), the use of a good or service as payment instead of cash.

Table VIII. Regression Performance for Log Post-Event Volatility Prediction

[LOGP]			Baseline expansion			
Year	SEN	RAN	EXP-RAN	EXP-LDA	EXP-SIM	EXP-SYN
	Mean squared error					
2001	1.5439	1.8787	1.3171	1.2386	1.2493	**1.2095**
2002	1.7492	1.9666	1.6175	1.5957	1.6146	**1.5416**
2003	1.5897	1.7617	1.4632	1.3321	1.3698	**1.3278**
2004	1.3746	1.7121	1.2009	1.1089	1.1105	**1.0742**
2005	1.2411	1.3008	1.0630	0.9948	0.9597	**0.9266**
2006	1.1932	1.2407	1.0383	0.9815	0.9453	**0.9019**
2007	1.2116	1.3075	1.1071	1.0696	1.0390	**1.0026**
2008	1.6537	1.7944	1.5075	1.4516	1.4351	**1.3694**
2009	1.4152	1.6439	1.2882	1.2338	1.2185	**1.1788**
2010	1.4353	1.6410	1.3130	1.2641	1.2714	**1.2284**
2011	1.3835	1.6741	1.2729	1.2501	1.2586	**1.2077**
2012	1.3337	2.5685	1.2077	1.1771	1.1815	**1.1369**
2013	1.0179	1.7914	0.9205	0.8925	0.8847	**0.8617**
Average	1.3956	1.7140	1.2551	1.1993	1.1952	**1.1513**

Note: Boldface numbers denote the best performance among the six word lists.

Table IX. Regression Performance for Log Volatility Prediction

[LOGP]			Baseline expansion			
Year	SEN	RAN	EXP-RAN	EXP-LDA	EXP-SIM	EXP-SYN
	Mean squared error					
2001	0.1871	0.2615	0.1655	0.1518	0.1548	**0.1516**
2002	0.2194	0.2816	0.1923	0.1839	0.1802	**0.1736**
2003	0.2181	0.3583	0.1858	0.1716	0.1700	**0.1643**
2004	0.1791	0.3167	0.1473	0.1425	0.1335	**0.1321**
2005	0.1570	0.2462	0.1274	0.1150	0.1180	**0.1107**
2006	0.1431	0.1973	0.1163	0.1091	0.1084	**0.1043**
2007	0.2393	0.2315	0.2264	0.2219	**0.2204**	0.2229
2008	0.7465	0.7713	0.6973	0.6760	0.6781	**0.6537**
2009	0.2519	0.3402	0.2443	**0.2365**	0.2424	0.2387
2010	0.1798	0.1903	0.1659	**0.1500**	0.1555	0.1514
2011	0.1400	0.1665	0.1250	**0.1216**	0.1232	0.1217
2012	0.2885	0.3580	0.2397	**0.2208**	0.2227	0.2290
2013	0.2832	0.3618	0.2190	0.1979	0.1912	**0.1861**
Average	0.2487	0.3139	0.2194	0.2076	0.2076	**0.2031**

Note: Boldface numbers denote the best performance among the six word lists.

continuous-space language models outperform LDA in capturing syntactic and contextual regularities in language, thus should be more suitable for discovering new keywords.

Tables VIII, IX, X, and XI tabulate the experimental results of the four regression tasks, in which the training data was composed of the financial reports in a five-year period, and the following year was the test data. For example, the reports from year 1996 to 2000 constituted a training corpus; the learned model was tested on the reports of year 2001. Recall that in addition to the experiments with EXP-SIM and EXP-SYN training, we also conducted experiments with random keyword expansion (EXP-RAN) and LDA keyword expansion (EXP-LDA), treating them as the two baselines. In addition, RAN denotes the word list containing 1,549 (i.e., the number of words in SEN) randomly selected words from the vocabulary. Both the RAN and EXP-RAN columns denote the

Table X. Regression Performance for Abnormal Trading Volume Prediction

[LOGP]			Baseline expansion			
Year	SEN	RAN	EXP-RAN	EXP-LDA	EXP-SIM	EXP-SYN
	Mean squared error					
2001	1.7500	1.7498	**1.7427**	1.7492	1.7553	1.7543
2002	1.9625	**1.9467**	1.9698	1.9866	1.9752	1.9932
2003	7.5203	**7.5081**	7.5145	7.5353	7.5305	7.5286
2004	83.9186	**83.8545**	83.8883	83.8794	83.9250	83.8819
2005	9.6592	9.6567	9.6474	9.6672	9.6465	**9.6421**
2006	5.9210	5.9562	5.8992	5.9013	5.9199	**5.8887**
2007	3.3442	3.3759	3.3386	3.3482	3.3461	**3.3380**
2008	2.0489	2.0943	2.0333	2.0378	2.0347	**2.0308**
2009	10.2552	10.3573	**10.2315**	10.2546	10.2317	10.2332
2010	12.0894	12.0936	12.0363	12.0365	12.0277	**12.0137**
2011	8.1203	8.1307	8.0798	**8.0707**	8.0866	8.0798
2012	4.1952	4.2770	4.1437	4.1287	4.1439	**4.1073**
2013	3.8222	3.8736	3.7559	3.7517	3.7454	**3.7211**
Average	11.8928	11.9134	11.8678	11.8729	11.8745	**11.8625**

Note: Boldface numbers denote the best performance among the six word lists.

Table XI. Regression Performance for Excess Return Prediction

[LOGP]			Baseline expansion			
Year	SEN	RAN	EXP-RAN	EXP-LDA	EXP-SIM	EXP-SYN
	Mean squared error					
2001	93.0883	93.6504	92.7968	92.7896	**92.6400**	92.6935
2002	61.2859	61.7805	61.2565	61.1809	61.2247	**61.1700**
2003	55.9239	**55.2580**	55.7454	55.5613	55.6952	55.6053
2004	141.7900	**141.1752**	141.9637	142.0180	142.0220	141.9710
2005	23.8518	23.6263	23.7293	**23.5003**	23.5586	23.6023
2006	22.8875	**22.6962**	22.9008	22.8541	22.7620	22.7799
2007	77.3309	**77.1594**	77.3304	77.2375	77.1635	77.2836
2008	59.0151	59.0353	59.0017	59.0996	**58.9928**	59.0579
2009	**120.9810**	121.1584	121.0707	121.2110	121.0950	121.1180
2010	130.3510	**130.1582**	130.3318	130.3830	130.2690	130.2290
2011	29.9918	30.8834	29.7853	29.6308	**29.6116**	29.6408
2012	30.7088	31.2066	30.8899	30.8484	**30.8429**	30.8591
2013	**1257.7100**	1258.2305	1257.9345	1257.8200	1257.8900	1258.1400
Average	161.9166	162.0014	161.9028	161.8565	**161.8283**	161.8577

Note: Boldface numbers denote the best performance among the six word lists.

results averaged from 20 randomly (expanded) word lists. The boldface numbers in the four tables denote the best performance among the six word lists.

Strikingly, for the task of post-event volatility prediction, as shown in Table VIII, the results completely match our expectation. The models trained on the four expanded keyword lists (EXP-∗) are consistently better than those trained on the original sentiment keywords only (SEN) and the random word list (RAN). Furthermore, the results of EXP-SYN in all years are significantly better than the EXP-RAN and EXP-LDA baselines with a p-value of less than 0.05; additionally, incorporating syntactic information (EXP-SYN) leads to better results than our simple expansion method (EXP-SIM), as expected. For the results of the task of log volatility prediction, Table IX shows that similar patterns hold. Note that, for EXP-SIM, the number of words used for training the regression and ranking models is even less than that of EXP-RAN.

Table XII. Stock Prices and Trading Volumes for Air T, Inc.

	Date	Stock prices	Trading volumes
	20040617	5.29000	1400
	20040618	5.21000	650
	20040621	5.25000	5650
	20040622	5.38000	3782
	20040623	5.37000	5307
Report filing date	20040624	5.27000	600
	20040625	7.55000	254562
	20040628	11.30000	1124845
	20040629	13.12000	2763145

Table XIII. Stock Prices for Superconductor Technologies, Inc.

	Date	Stock prices
	20130306	0.19050
	20130307	0.21000
Report filing date	20130308	0.21490
	20130311	0.18900
	20130312	2.26500
	20130313	3.86000
	20130314	3.20000
	20130315	3.13000

When predicting abnormal trading volumes and excess returns (see Tables X and XI), on average, the proposed expansion methods (EXP-SYN and EXP-SIM) yield the best performance among the six word lists, although their performance is not significantly better than that for the risk prediction tasks reported in Tables VIII and IX. In addition, for both tasks, we note that the resulting MSEs vary significantly because of the wide range of abnormal trading volumes and excess returns. Take, for example, the company Air T, Inc.: Table XII shows an abrupt rise in trading volumes after the report filing date on June 24, 2004. Also, Table XIII demonstrates that the stock price of Superconductor Technologies Inc. on March 12, 2013 became nearly 12 times the price on March 11, 2013, resulting in a considerable amount of excess return for the company in that year.[13] Such dramatic changes complicate our solely text-based prediction of abnormal trading volumes and excess returns, especially as compared with the prediction of log post-event and stock volatilities.

Furthermore, to examine the effect of the numbers of expanded words (top-20 words used in our experiments) on prediction performance, we conducted a robustness analysis for the log post-event volatility prediction task of EXP-SIM and EXP-SYN with different numbers of top-N expanded words, in which N varies from 2 to 100, as shown in Figure 3. As shown in the figure, increasing the number (top N) of expanded words indeed decreases MSEs, thus improves performance. However, observe that there is a diminishing return between prediction performance and the number (top N) of expanded words. Also. note that increasing the number of expanded words entails additional computational costs. For example, the sizes of the simple expansion lists (EXP-SIM) for $N = 15$, $N = 20$, and $N = 40$ are 8,611, 10,258, and 15,651, respectively; those with syntactic information (EXP-SYN) are 13,927, 16,467, and 24,072, respectively. Furthermore,

[13]On March 11, 2013, Superconductor Technologies Inc. announced a one-for-twelve (1:12) reverse split of its common stock, effective at the close of the business day.

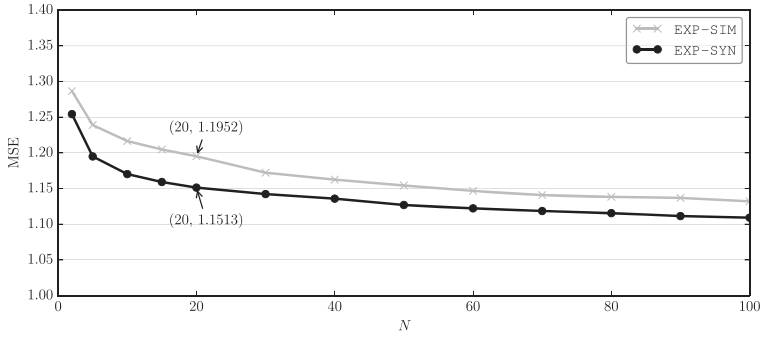

Fig. 3. Post-event volatility prediction of EXP-SIM and EXP-SYN using different top-N expanded words.

observe that the pattern of the performance on two lists holds as N increases. Considering the trade-off between performance and computational cost, we chose the top-20 expanded words in our experiments, which is also the default setting in the word2vec tool.

5.3. Discussion

In this section, we provide some insight into the discovered words from the three main aspects of a financial statement (from an accounting point of view): assets, liabilities, and equity. In the following description, we take several related words from each aspect and their expanded words as examples, then discuss the roles that these words usually play in financial reports. Furthermore, we develop an information retrieval system for 10-K reports,[14] with which searches can be based on metadata or on full-text (or other content-based) indexing; the system is therefore of great help in extracting relevant texts and further analyzing the relationships among multiple words.

5.3.1. Assets. Given the seed word **finance**, the proposed expansion method produces words such as *raise*, *fund*, *obtain*, *security*, and *cash*. In financial reports, **finance** often refers to financial assets, which can be *cash*, cash equivalents, or contractual rights to receive cash or another financial asset from another entity. In addition, a company may *raise funds* or *obtain* other assets; for example, a *security* is a tradable financial asset of any kind.

Given the seed word **inventory**, our method produces words such as *obsolete*, *lifo*, and *excess*. Although **inventory** (the physical goods you sell or use to make your product) is an asset for a company, *excess* amount of **inventory** usually leads to low turnover rates, which may further lead to *obsolete* **inventory** (especially common in 3C industries). Additionally, the word *LIFO*, also known as last in first out, is one of the most commonly used **inventory** accounting methods regulated by GAAP.[15] Later, we show parts of the original texts from 10-K reports that contain the expanded words *excess* and *obsolete* for *inventory* (extracted from the report of Digital Lightwave Inc. in 2005):

> The decrease in cost of goods sold was due to a reduction in charges related to *excess* and *obsolete* **inventory** and other inventory claims.

5.3.2. Liabilities. Seed word **interest** yields the expanded words *fix*, *rate*, *noninterest*, and *debt*. In corporate finance, a *debenture* is a medium- to long-term *debt* instrument used by large companies to borrow money at a *fixed rate* of **interest**. On the other hand,

[14]Available at http://clip.csie.org/10K/.
[15]Generally Accepted Accounting Principles.

a *noninterest* bearing note is a *debt* for which there is no documented requirement for the borrower to pay the lender any *rate* of **interest**.

Given the seed word **allowance**, we have words such as *doubt*, *uncollectible*, *provision*, and *reserve*. Regulated by IFRS[16] and GAAP, companies are required to recognize a portion of income as a loss once they earn uncollected revenue, which is called **allowance** for a *doubtful* (or *uncollectible*) account. Moreover, a *provision* can be a liability of uncertain timing or amount; in IFRS, this is referred to as *reserve*. Quoted here are parts of the original texts from 10-K reports that contain the expanded words *provision* and *uncollectible* for the word *allowance* (extracted from the report of Concorde Career Colleges Inc. in 2006):

> The *provision* for *uncollectible* accounts as a percentage of revenue was 5.2% in 2005 compared to 4.3% in 2004. Depending on the effectiveness of the Company's internal and external collection efforts, the *provision* for *uncollectible* accounts may vary as a percentage of revenue in future periods.

5.3.3. Equity. Given the seed word **expense**,[17] we obtain *administration*, *depreciation*, *amortization*, *salary*, and *cost*. For a company, *salary*, *depreciation*, and *amortization* **expenses** are common; *depreciation* **expenses** are usually related to fixed assets, while *amortization* costs are related to intangible assets such as patents or goodwill.

Given the seed word **dividend**, we have *declared*, *undeclared*, *paid*, and *preferential*. A **dividend** is a *payment* made by a company to its stock holders, usually as a distribution of profits. There is one special kind of stock, called *preferenced* stock shares, whose holders have a higher priority to receive **dividends** than ordinary shareholders, but the **dividend** amount is usually predetermined. If a company fails to *declare dividends* on that year, it becomes an *undeclared* **dividend**. Take, for example, a piece of the original texts from 10-K reports of the company, Trailer Bridge Inc. in 2005:

> The *undeclared* **dividends** on the *preferred* stock series "B" increased to $1,115,796 in 2004 from $846,385 in 2003, primarily due to increases in the contractual **dividend** rate from 2003. These **dividends** will never be *paid* and are recorded because they were contractual obligations... The Company has not *declared* or *paid* **dividends** on its common stock during the past five years.

5.3.4. Summary. As demonstrated in these analyses, expanded words usually have similar meanings, high co-occurrences, or high correlation with the seed words (from the original financial dictionary). This discussion attests to the ability of the proposed method to capture both syntactic and contextual regularities in language, and shows the success of the method when applied to the field of finance.

6. CONCLUSIONS

In this article, we applied the CBOW model to textual information in 10-K financial reports to discover new finance keywords. Specifically, we adopted high-dimensional word representations to expand keywords from the well-known financial sentiment lexicon proposed by Loughran and McDonald [2011]. Additionally, we proposed a novel approach to incorporate syntactic information into the CBOW model to capture more similarly meaningful keywords. The experimental results on the four prediction tasks using the discovered keywords demonstrate that our approach is effective for discovering predictability keywords. Finally, the discussions from an accounting point of view

[16]International Financial Reporting Standards.

[17]In accounting, **expense** has a very specific meaning. It is an outflow of cash or other valuable assets from a person or company to another person or company. In terms of the accounting equation, expenses reduce owners' equity.

demonstrate the ability of the proposed method to capture syntactic and contextual regularities between words, and shows the success of this method when applied to the field of finance.

ACKNOWLEDGMENTS

We thank Yu-Wen Liu for assistance.

REFERENCES

Nihat Aktas, Eric De Bodt, Fany Declerck, and Herve Van Oppens. 2007. The PIN anomaly around M&A announcements. *Journal of Financial Markets* 10, 2, 169–191.

Werner Antweiler and Murray Z. Frank. 2004. Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance* 59, 3, 1259–1294.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3, 1137–1155.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (IMCL'08)*. 160–167.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*. 984–991.

Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. *Advances in Neural Information Processing Systems* 9, 155–161.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science* 14, 2, 179–211.

Eugene F. Fama and Kenneth R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 1, 3–56.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM* 56, 4, 82–89.

Diego Garcia. 2013. Sentiment during recessions. *The Journal of Finance* 68, 3, 1267–1300.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 513–520.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*. 272–280.

Feng Li. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45, 2, 221–247.

Feng Li, Russell Lundholm, and Michael Minnis. 2013. A measure of competition based on 10-k filings. *Journal of Accounting Research* 51, 2, 399–436.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66, 1, 35–65.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4, 782–796.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*. 1045–1048.

Wijnand Nuij, Viorel Milea, Frederik Hogenboom, Flavius Frasincar, and Uzay Kaymak. 2014. An automated framework for incorporating news into stock trading strategies. *IEEE Transactions on Knowledge and Data Engineering* 26, 4, 823–835.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program: Electronic Library and Information Systems* 14, 3, 130–137.

S. McKay Price, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance* 36, 4, 992–1011.

Xin Ying Qiu, Padmini Srinivasan, and Yong Hu. 2014. Supervised learning models to predict firm performance with annual reports: An empirical study. *Journal of the Association for Information Science and Technology* 65, 2, 400–413.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and PDP Research Group (Eds.). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1*. MIT Press, Cambridge, MA. 318–362.

Bernhard Schölkopf and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech & Language* 21, 3, 492–518.

Richard Socher, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 129–136.

Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance* 63, 3, 1437–1467.

Ruey S. Tsay. 2005. *Analysis of Financial Time Series*. Wiley, Hoboken, NJ.

Matthias W. Uhl. 2011. Explaining U.S. consumer behavior with news sentiment. *ACM Transactions on Management Information Systems* 2, 2, Article 9, 18 pages.

Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP'13)*. 802–808.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Vol. 3*. 2764–2770.

Received September 2015; revised April 2016; accepted May 2016