# FIN10K: A Web-based Information System for Financial Report Analysis and Visualization

Yu-Wen Liu*, Liang-Chih Liu†, Chuan-Ju Wang‡, and Ming-Feng Tsai*
*Department of Computer Science, National Chengchi University, Taipei, Taiwan
†Institute of Finance, National Chiao Tung University, Hsinchu, Taiwan
‡Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
g10435@cs.nccu.edu.tw, tony919kimo@gmail.com, cjwang@citi.sinica.edu.tw, mftsai@nccu.edu.tw

## ABSTRACT

In this demonstration, we present FIN10K, a web-based information system that facilitates the analysis of textual information in financial reports. The proposed system has three main components: (1) a 10-K Corpus, including an inverted index of financial reports on Form 10-K, several numerical finance measures, and pre-trained word embeddings; (2) an information retrieval system; and (3) two data visualizations of the analyzed results. The system can be of great help in revealing valuable insights within large amounts of textual information. The system is now online available at http://clip.csie.org/10K/.

## Keywords

web-based system; text mining; data visualization

## 1. INTRODUCTION

In the past few decades, a great deal of textual information in financial markets has accumulated, including financial reports and investor message boards. This textual information contains copious signals that significantly affect observable economic numbers such as stock prices and the corresponding trading volumes. Due to this close relationship between textual information and financial numerical measures, a growing body of accounting and finance research has adopted techniques from natural language processing such as sentiment analysis to deal with the textual information in finance [6, 10, 3, 4, 9]. For example, academics have investigated how financial sentiment keywords in annual SEC-mandated financial reports influence investor expectations about a company's future stock prices and potential risk [5, 7].[1] Practitioners such as fund managers utilize these sentiment keywords to offer prospects for companies and design their own investment strategies. However, when analyzing financial textual information, academics and practitioners both encounter a major hurdle — the findings are usually insignificant and biased because most finance keywords are context-sensitive. Therefore, it is essential to construct an information system to assist people to discover meaningful insights within large amounts of textual information in finance.

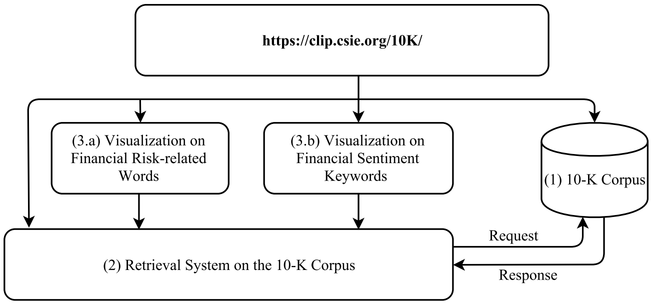

https://clip.csie.org/10K/
(3.a) Visualization on Financial Risk-related Words
(3.b) Visualization on Financial Sentiment Keywords
(1) 10-K Corpus
Request
Response
(2) Retrieval System on the 10-K Corpus

**Figure 1: System architecture.**

In this demonstration, we present FIN10K, a web-based information system for financial report analysis and visualization which aims to bridge the gap between technical results and useful interpretations. The system has three main parts: (1) a 10-K Corpus, including an inverted index of financial reports on Form 10-K, some numerical finance measures, and pre-trained word embedding vectors,[2] (2) an information retrieval system, and (3) two data visualizations of the analyzed results. First, the 10-K corpus contains 40,708 financial reports from years 1996 to 2013 along with three financial measures and word embedding vectors trained using a continuous bag-of-words (CBOW) model [8].[3] Second, a retrieval system is provided that quickly locates and highlights target keywords in the original texts of specified financial reports for companies. In addition, in the retrieval system, we also provide the annualized stock return volatility and the relative risk level for corresponding companies, both of which can be treated as proxies for financial risk. Third, in our system, two data visualizations are provided to showcase interesting results. The first part visualizes high-risk and low-risk words learned via the ranking models proposed in [11]. The visualization shows the correlation between words and financial risk and provides quantitative values for the words, which are plotted as bouncy balls of varying sizes, representing the intensity of their correlation with financial risk. Moreover, the second part visualizes the words in the six financial sentiment lexicons from [7] based on our learned word representations via the CBOW models. Each word representation is a 200-dimensional real-valued vector and is transformed

[1] SEC indicates Securities and Exchange Commission. One of the SEC-mandated financial reports is the 10-K report.

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

[3] We used the word2vec toolkit: https://code.google.com/p/word2vec/.

*CIKM'16 October 24-28, 2016, Indianapolis, IN, USA*

ACM ISBN 978-1-4503-4073-1/16/10.
DOI: http://dx.doi.org/10.1145/2983323.2983328

into two-dimensional space using $t$-distributed stochastic neighbor embedding (t-SNE), a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional data [13]. The visualization results drive our other work related to financial keyword expansion published in [12]. Note that for both visualizations, the selected words can be fed back to the retrieval system, which is of great help in extracting relevant texts and further analyzing the relationships among multiple words.

The remainder of this paper is arranged as follows. In Section 2 we describe our platform in detail. We then provide two case studies in Section 3. Section 4 concludes the paper.

## 2. SYSTEM DESCRIPTION

The proposed FIN10K system consists of three main components (see Figure 1). In the following subsections, we provide detailed descriptions for each component of the system.

### 2.1 The 10-K Corpus

Our collected corpus is composed of three types of data: financial reports on Form 10-K, three financial measures, and pre-trained word embedding vectors. There are in total 40,708 SEC financial reports from publicly traded U.S. companies with release dates in the period from 1996 to 2013; for each report, we provide the original form 10-K, the Management Discussion and Analysis (MD&A), and the tokenized MD&A section. In addition to this textual information, we also calculate three financial measures: abnormal trading volumes, and stock return volatilities measured in the twelve-month period before and after each report. Finally, we publish pre-trained vectors via word2vec trained on the collected financial reports. The models (with and without syntactic information) contain 200-dimensional vectors for each word. For details on incorporating syntactic information, please see [12].

### 2.2 The Retrieval System for 10-K Corpus

Figure A.1 illustrates the interface of the retrieval system and search results for the example query word "forbear*" and metadata "data:[2002 to 2003]." First, a user can either enter a keyword in the search bar (1) or select the pre-defined negative words (4) and positive words (5) from the financial sentiment lexicons. In addition, the user can conduct a meta search for the reports from a specific year or a specified period via the dropdown selector (2) and the text field (3), respectively. Then, the user activates the retrieval by clicking the GO! button (6). After the retrieval, the system returns the relevant snippets with the corresponding release dates, company names, stock return volatilities measured in the twelve-month period after each report (i.e., $\sigma^+(+12)$), and their relative risk levels (7). We split the volatilities into five relative risk levels; see [11] for detailed calculations. In addition, the user can obtain the original MD&A section via button (8).

The retrieval system was developed using Bottle, a Python Web Framework, and the front-end framework Bootstrap. In addition, the 10-K corpus is indexed using an open source pure-Python text indexing library.

### 2.3 Visualization

The FIN10K system also presents interactive SVG graphs with D3.js to help users quickly understand not only the relation between stock volatility [11] and textual information but also the syntactic and contextual information among financial sentiment words [7].

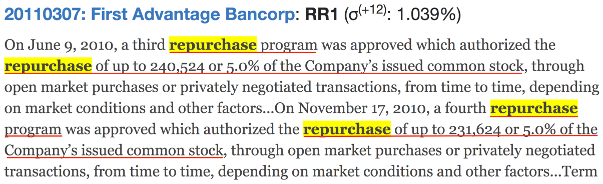
Figure A.2 is a visualization of high-risk and low-risk words trained via the ranking models proposed in [11]. First, a user can specify the words learned from 5-year period reports via the dropdown selector (1). We categorize risk-associated words into two types: low-risk words (i.e., words with negative weights) and high-risk words (those with positive weights). In the figure, the left panel (2) exhibits the low-risk words, and the right one (3) plots the high-risk words; note that the circle area is proportional to the absolute value of the learned weight for each word.[4] Low-risk words such as "unsecur" (4) and "merger" are associated with low future stock volatilities; in contrast, high-risk words such as "forbear" (5) and "lender" are related to high future stock volatilities. For example, as observed in Figure 1, the future risks (measured by stock volatilities) of the top-3 retrieved companies are rather high (i.e., RR4 or RR5) for the query "forbear*." Additionally, by clicking on the bubble, the corresponding word is fed back to the retrieval system, which retrieves the relevant financial reports; thus the user can further analyze the contextual information of each learned word.

Figure A.3 shows a screenshot of the scatter plot for the words in the six financial sentiment lexicons based on the learned word representations. The six lexicons are *litigious*, *negative*, *positive*, *strong*, *uncertainty*, and *weak*, each of which is assigned a specific color (1). When a user hovers the mouse over a particular node (2), the word is shown with its POS tag and the corresponding sentiment lexicon. Also, each node is clickable; the word on it can be fed back to the retrieval system. Since our proposed approach in [12] also incorporates syntactic information into the CBOW model to learn the continuous word vector representations, each word in the figure is associated with a POS tag; the tag mappings are listed in the right panel (3).

## 3. CASE STUDY

### 3.1 Case 1: Ranking risk levels of companies via keywords in financial reports

Via the ranking models learned in [11], we draw attention to some strong and interesting correlations between texts and financial risk. With the presented visualization and retrieval system, users can more easily understand the results and thus gain greater insight and understanding into the usefulness of textual information in financial reports.



**20110307: First Advantage Bancorp**: **RR1** ($\sigma^{(+12)}$: 1.039%)

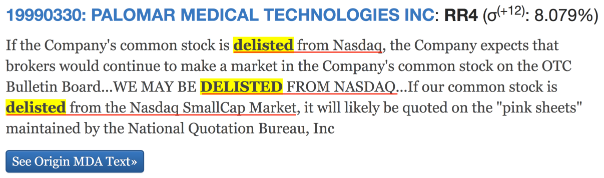
On June 9, 2010, a third **repurchase** program was approved which authorized the **repurchase** of up to 240,524 or 5.0% of the Company's issued common stock, through open market purchases or privately negotiated transactions, from time to time, depending on market conditions and other factors...On November 17, 2010, a fourth **repurchase** program was approved which authorized the **repurchase** of up to 231,624 or 5.0% of the Company's issued common stock, through open market purchases or privately negotiated transactions, from time to time, depending on market conditions and other factors...Term

**Figure 2: Low-risk word "repurchase."**

Among the retrieved results, stemmed low-risk keywords such as "repurchas" and "unsecur" have negative weights and thus negative relations to the magnitude of future stock volatility. Figure 2 shows that the stemmed word "repurchas" is mostly in the context of "share repurchases" or "repurchase program." Companies announcing share repurchase programs signal undervaluation of their prior stock return and further reveal information about their positive perspective on future profitability [2]. Thus, "repurchas" is in general negatively correlated with financial risk and hence lends a negative weight to volatility. On the other, "unsecur" is mostly in the context of "unsecured term loan" and "unsecured line of credit." Companies allowed to provide unsecured loans by banks are usually regarded as

[4] The Porter stemmer was used to generate the stemmed forms of the risk-associated words in the system.

financially healthier borrowers than those that are required to pledge part of their assets as collateral [1]. Therefore, "unsecur" is also negatively correlated with financial risk.


19990330: PALOMAR MEDICAL TECHNOLOGIES INC: RR4 ($\sigma^{(+12)}$: 8.079%)

If the Company's common stock is delisted from Nasdaq, the Company expects that brokers would continue to make a market in the Company's common stock on the OTC Bulletin Board...WE MAY BE DELISTED FROM NASDAQ...If our common stock is delisted from the Nasdaq SmallCap Market, it will likely be quoted on the "pink sheets" maintained by the National Quotation Bureau, Inc

**Figure 3: High-risk word "delist."**

In contrast to low-risk words, the retrieved high-risk keyword "deficit" implies a negative perspective on future profitability and is thus positively correlated with financial risk and volatility of stock returns. However, it could be difficult to explain why the word "nasdaq" is associated with high financial risk. With our retrieval system, though, it can be easily found that the high-risk words "delist," "nasdaq," and "smallcap" are found mostly in the context of the sentence "delist from the Nasdaq smallcap market," as shown in Figure 3.

### 3.2 Case 2: Visualization of six financial sentiment lexicons

As illustrated in Figure A.3, words in the same financial sentiment lexicon generally belong to one or two groups. In addition, Figure A.4 shows that the positive and negative words are for the most part distanced from each other; Figure A.5 indicates that the group of litigious words overlaps substantially with the negative group, which could be due to the fact that in finance litigious words usually carry negative implications. From the 2-D visualization, we observe that words with similar meanings in finance are usually close to each other, based on the learned word representations; for instance, achiev v.s. accomplish and default v.s. bankruptcy (see Figures A.4 and A.5). The visualization facilitates the discovery of highly-correlated words and makes it possible to expand sentiment keywords in financial reports.

## 4. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a new web-based information system that can be used to retrieve relevant financial reports for a query and to visualize the analyzed results. In its current form, it conducts both full-text and meta search and displays the relevant snippets of financial reports with their corresponding release dates, company names, stock return volatilities measured in the twelve-month period after each report, and their relative risk levels. Additionally, it visualizes high- and low-risk words learned via the ranking models, as well as the syntactic and contextual information among financial sentiment words. The web demo effectively bridges the gap between technical analyzed results and useful interpretations and give a first impression for academics and practitioners interested in understanding and inspecting valuable insights into large amounts of financial textual information.

We continue to develop and extend the system, with emphasis on incorporating several type of statistics into the retrieval system. The goal of this extension is equip users with a better understanding of the corresponding financial measures of the retrieved financial reports, for instance by aggregating the risk levels and displaying their statistics for a given query. On the other hand, enhancing the interactiveness of our system will be one of the future work; for this enhancement, we may consider the techniques adopted in *Tweet Sentiment Visualization App* (https://www.csc.ncsu.edu/faculty/healey/tweet_viz/).

## Acknowledgments

This work has been partially supported by the Ministry of Science and Technology in Taiwan under the grants MOST 102-2221-E-845-002-MY3 and MOST 104-2221-E-004-010.

## 5. REFERENCES

[1] M. T. Billett, M. J. Flannery, and J. A. Garfinkel. The effect of lender identity on a borrowing firm's equity return. *The Journal of Finance*, 50(2):699–718, 1995.

[2] A. K. Dittmar. Why do firms repurchase stock*. *The Journal of Business*, 73(3):331–355, 2000.

[3] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[4] D. Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

[5] S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith. 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, pages 272–280, 2009.

[6] M.-C. Lin, A. J. T. Lee, R.-T. Kao, and K.-T. Chen. Stock price movement prediction using representative prototypes of financial reports. *ACM Transactions on Management Information Systems*, 2(3):19:1–19:18, 2008.

[7] T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[8] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 746–751, 2013.

[9] T. H. Nguyen and K. Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL '15, pages 1354–1364, 2015.

[10] S. M. Price, J. S. Doran, D. R. Peterson, and B. A. Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011, 2012.

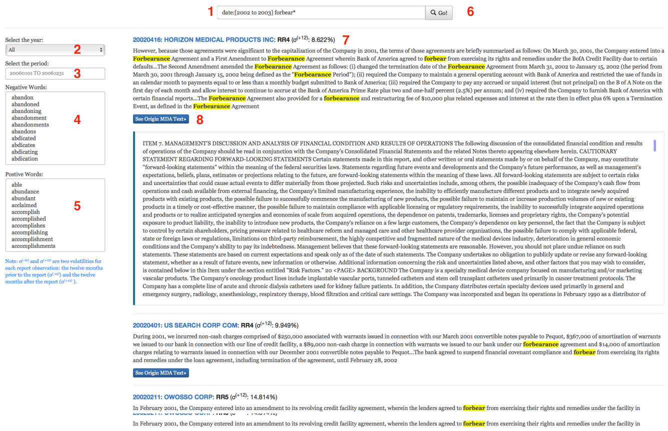
[11] M.-F. Tsai and C.-J. Wang. *Advances in Information Retrieval: 35th European Conference on IR Research (ECIR 2013)*, chapter Risk Ranking from Financial Reports, pages 804–807. 2013.

[12] M.-F. Tsai and C.-J. Wang. Financial keyword expansion via continuous word vector representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1453–1458, 2014.

[13] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

# APPENDIX

## A. THE SCREENSHOTS OF THE SYSTEM



date:[2002 to 2003] forbear*

Select the year: All

Select the period: 20060101 TO 20061231

Negative Words:

abandon
abandoned
abandoning
abandonment
abandonments
abandons
abdicated
abdicates
abdicating
abdication

Postive Words:

able
abundance
abundant
acclaimed
accomplish
accomplished
accomplishes
accomplishing
accomplishment
accomplishments

Note: $\sigma^{(-12)}$ and $\sigma^{(+12)}$ are two volatilities for each report observation: the twelve months prior to the report ($\sigma^{(-12)}$) and the twelve months after the report ($\sigma^{(+12)}$ ).

**20020416: HORIZON MEDICAL PRODUCTS INC**: **RR4** ($\sigma^{(+12)}$: 8.622%)

However, because those agreements were significant to the capitalization of the Company in 2001, the terms of those agreements are briefly summarized as follows: On March 30, 2001, the Company entered into a **Forbearance** Agreement and a First Amendment to **Forbearance** Agreement wherein Bank of America agreed to **forbear** from exercising its rights and remedies under the BofA Credit Facility due to certain defaults...The Second Amendment amended the **Forbearance** Agreement as follows: (i) changed the termination date of the **Forbearance** Agreement from March 31, 2002 to January 15, 2002 (the period from March 30, 2001 through January 15, 2002 being defined as the "**Forbearance** Period"); (ii) required the Company to maintain a general operating account with Bank of America and restricted the use of funds in an calendar month to payments equal to or less than a monthly budget submitted to Bank of America; (iii) required the Company to pay any accrued or unpaid interest (but not principal) on the B of A Note on the first day of each month and allow interest to continue to accrue at the Bank of America Prime Rate plus two and one-half percent (2.5%) per annum; and (iv) required the Company to furnish Bank of America with certain financial reports...The **Forbearance** Agreement also provided for a **forbearance** and restructuring fee of $10,000 plus related expenses and interest at the rate then in effect plus 6% upon a Termination Event, as defined in the **Forbearance** Agreement

ITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS The following discussion of the consolidated financial condition and results of operations of the Company should be read in conjunction with the Company's Consolidated Financial Statements and the related Notes thereto appearing elsewhere herein. CAUTIONARY STATEMENT REGARDING FORWARD-LOOKING STATEMENTS Certain statements made in this report, and other written or oral statements made by or on behalf of the Company, may constitute "forward-looking statements" within the meaning of the federal securities laws. Statements regarding future events and developments and the Company's future performance, as well as management's expectations, beliefs, plans, estimates or projections relating to the future, are forward-looking statements within the meaning of these laws. All forward-looking statements are subject to certain risks and uncertainties that could cause actual events to differ materially from those projected. Such risks and uncertainties include, among others, the possible inadequacy of the Company's cash flow from operations and cash available from external financing, the Company's limited manufacturing experience, the inability to efficiently manufacture different products and to integrate newly acquired products with existing products, the possible failure to successfully commence the manufacturing of new products, the possible failure to maintain or increase production volumes of new or existing products in a timely or cost-effective manner, the possible failure to maintain compliance with applicable licensing or regulatory requirements, the inability to successfully integrate acquired operations and products or to realize anticipated synergies and economies of scale from acquired operations, the dependence on patents, trademarks, licenses and proprietary rights, the Company's potential exposure to product liability, the inability to introduce new products, the Company's reliance on a few large customers, the Company's dependence on key personnel, the fact that the Company is subject to control by certain shareholders, pricing pressure related to healthcare reform and managed care and other healthcare provider organizations, the possible failure to comply with applicable federal, state or foreign laws or regulations, limitations on third-party reimbursement, the highly competitive and fragmented nature of the medical devices industry, deterioration in general economic conditions and the Company's ability to pay its indebtedness. Management believes that these forward-looking statements are reasonable. However, you should not place undue reliance on such statements. These statements are based on current expectations and speak only as of the date of such statements. The Company undertakes no obligation to publicly update or revise any forward-looking statement, whether as a result of future events, new information or otherwise. Additional information concerning the risk and uncertainties listed above, and other factors that you may wish to consider, is contained below in this Item under the section entitled "Risk Factors." 20 <PAGE> BACKGROUND The Company is a specialty medical device company focused on manufacturing and/or marketing vascular products. The Company's oncology product lines include implantable vascular ports, tunneled catheters and stem cell transplant catheters used primarily in cancer treatment protocols. The Company has a complete line of acute and chronic dialysis catheters used for kidney failure patients. In addition, the Company distributes certain specialty devices used primarily in general and emergency surgery, radiology, anesthesiology, respiratory therapy, blood filtration and critical care settings. The Company was incorporated and began its operations in February 1990 as a distributor of

**20020401: US SEARCH CORP COM**: **RR4** ($\sigma^{(+12)}$: 9.949%)

During 2001, we incurred non-cash charges comprised of $250,000 associated with warrants issued in connection with our March 2001 convertible notes payable to Pequot, $367,000 of amortization of warrants we issued to our bank in connection with our line of credit facility, a $89,000 non-cash charge in connection with warrants we issued to our bank under our **forbearance** agreement and $14,000 of amortization charges relating to warrants issued in connection with our December 2001 convertible notes payable to Pequot...The bank agreed to suspend financial covenant compliance and **forbear** from exercising its rights and remedies under the loan agreement, including termination of the agreement, until February 28, 2002

**20020211: OWOSSO CORP**: **RR5** ($\sigma^{(+12)}$: 14.814%)

In February 2001, the Company entered into an amendment to its revolving credit facility agreement, wherein the lenders agreed to **forbear** from exercising their rights and remedies under the facility in

In February 2001, the Company entered into an amendment to its revolving credit facility agreement, wherein the lenders agreed to **forbear** from exercising their rights and remedies under the facility in

**Figure A.1: The Retrieval System for 10-K Corpus.**

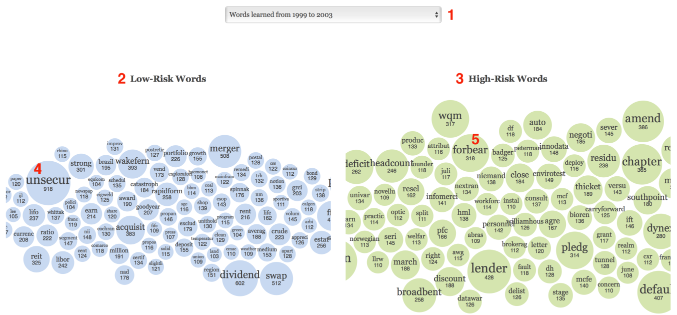
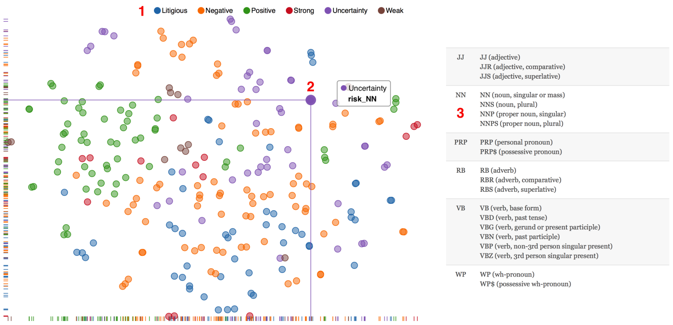


**Figure A.2: Visualization of high-risk and low-risk words.**



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

**Figure A.3: Visualization of the six financial sentiment lexicons based on the learned word representations.**

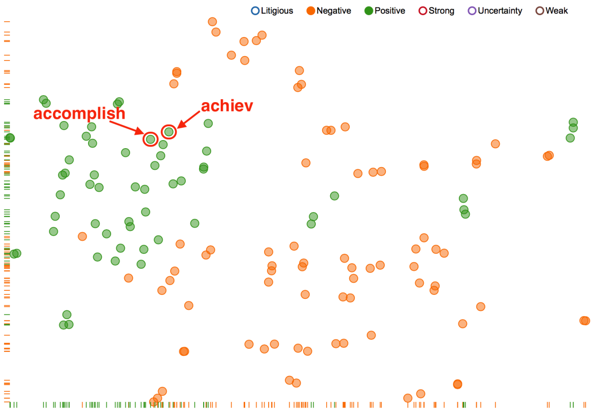


**Figure A.4: Positive and negative words.**

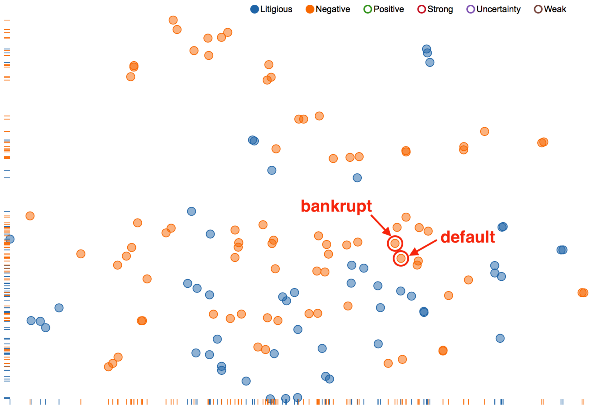


**Figure A.5: Negative and litigious words.**