Predicting Firm Financial Performance from SEC Filing Changes using Automatically Generated Dictionary

Aparna Gupta*[1], Vipula Rawte[2], and Mohammed J. Zaki[2]

[1]Lally School of Management, Rensselaer Polytechnic Institute, Troy, NY, 12180, US

[2]Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY, 12180, US

April 24, 2023

Abstract

Textual data are increasingly used to predict firm performance, however extracting useful signals towards serving this goal with a continuously growing repository of financial reports and documents is challenging, even by the state-of-the-art machine learning and Natural Language Processing (NLP) techniques. We propose a novel approach to automatically create a word list from SEC filings (10-K and 8-K reports) using advanced deep learning and NLP techniques and compare their performance against the widely used Loughran-McDonald sentiment dictionaries. We additionally analyze a corpus of 8-K and 10-K documents to evaluate their relative informativeness for firm performance prediction. Since 8-K filings provide corporate updates along a fiscal year, we compare their content against changes in 10-Ks between consecutive years to assess the incremental value of information provided in these regulatory filings. Information effectiveness is examined by predicting six key financial indicators for a set of US banks using ridge regression. Our results positively support sentiment dictionaries expansion by automatically extracting meaning from text and highlight the benefits obtainable from utilizing update filings.

Keywords: regulatory filings; text analytics; bank risk; performance; prediction; attention score.

JEL Code: C18, C45, C53.

*Corresponding author; Email: guptaa@rpi.edu.

1 Introduction and Motivation

The volume of unstructured text data has been growing rapidly in various domains such as in finance, economics, biomedicine, agriculture, and law. In finance, text data are available in various forms like financial news, earnings call transcripts, company regulatory filings, analyst report and micro-blogs. In addition to the traditionally used quantitative variables for various prediction tasks, text data has shown promising results in financial sentiment analysis, prediction of bank failures, and stock market volatility (Das et al., 2014; Loughran and McDonald, 2016; Khalil and Pipa, 2022; Duarte et al., 2021). Examining this rapidly growing, large volumes of text data manually is impractical. Thus, innovative Machine Learning (ML) and Natural Language Processing (NLP) techniques are being developed to analyze these large volumes of text data.

ML techniques typically require text data to be represented as a set of textual features. A naive textual feature definition is to compute scores (e.g., word count) for each word in a dictionary. The most commonly used textual feature is the Term Frequency-Inverse Document Frequency (TF-IDF) score, used in Kogan et al. (2009). Research in finance and accounting often uses the Harvard Psycho-sociological Dictionary, particularly, Harvard-IV-4 TagNeg[1] to build the textual features based on the "tone" of the text. This dictionary, however, was not specifically developed for the finance domain. Loughran and McDonald (2011) noted that some words in the Harvard Dictionary get inappropriately misclassified as "negative" in the financial context. Thus, for more meaningful and relevant financial textual features, a *hand-curated* Loughran-McDonald (L&M) sentiment word list was created using 10-K filings from 1994-2008 (Loughran and McDonald, 2011). Many important works have utilized this word list, such as (Tsai and Wang, 2014; Tsai et al., 2016; Tsai and Wang, 2017), and continues to be widely used for extracting financial textual features. However, an increasing volume of financial documents, as well as an evolving vocabulary in the financial domain, necessitates any meaningful dictionary to support NLP-based financial prediction to be periodically updated. L&M's *hand-curated* word list is difficult to manually update.

In order to address this challenge, we *automatically* create a word list using a novel method based on a neural network-based technique for NLP pre-training, namely BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), which requires no manual effort. We utilize a BERT-based automatic word list building framework, which we refer to as *attention-score based word list*, for the financial performance prediction and compare it against L&M word list based prediction. We fo-

[1] http://www.wjh.harvard.edu/inquirer

cus on the US banking sector for this word innovations evaluation, and utilize the banks' 10-K and 8-K filings made to the SEC (Securities and Exchange Commission) as the textual corpus. The US SEC mandates each public company to annually submit a comprehensive financial report, called 10-K. Besides annual 10-K reports, the US SEC also requires public companies to file updates for significant unscheduled events such as acquisitions, bankruptcy, or other changes during a fiscal year in reports called 8-Ks.

In this paper, we study and compare the informativeness of 8-Ks and 10-Ks in terms of the automatic word dictionary in predicting key financial indicators for US banks. Given that 8-Ks are update filings and 10-K are annual report filings required of all public companies, we compare the prediction performance of these two kinds of reports. Textual features are defined by TF-IDF scores for L&M sentiment words and our attention-score based words extracted from the 8-Ks and 10-Ks. Our automatic word lists are domain-specific, *automatically* created from the 10-K reports using the BERT model. Besides studying the two types of filings separately, we also study the informativeness of combined textual features extracted from 8-K and 10-K filings. After creating the automatic word list using attention scores, we investigate the following three research questions in this paper:

1. **RQ1: Does the textual content of 8-K and 10-K filings uncover any useful information about a bank?**

    (a) RQ1.1: Are sentiment words helpful for constructing textual features in prediction tasks?

2. **RQ2: How do bank's 8-K filings compare with 10-K documents in terms of their informativeness for prediction tasks?**

3. **RQ3: Are updates reported in bank's 8-K filings more informative than the changes in 10-Ks in two consecutive years?**

The annual 10-K reports contain detailed information about company's activities, risk factors, plans and performance, and other relevant data. These reports provide valuable information to investors to help them make a range of investment and risk management decisions. Typically, a 10-K filing consists of 15 items. In particular, *Item 7: Management's Discussion and Analysis (MD&A) of Financial Condition and Results of Operations* and *Item 7A: Quantitative and Qualitative Disclosures about Market Risk* discuss the financial results of the firm for the previous fiscal year. Prior research has demonstrated usefulness of Item 7/7A (combined) in predicting various quantitative financial indicators for firms. Kogan et al. (2009) predict quantitative variable such as stock volatility using features extracted from firms' 10-Ks. 10-K reports are also used to predict other variables such as post-event return volatility, abnormal

trading volume and excess return (Tsai et al., 2016). Likewise, researchers (Huang, 2010; Rawte et al., 2018) have demonstrated the usage of 10-K *Item 1A: Risk factors* in predicting risk measures, annual stock returns, and key financial ratios such as ROA and Tobin's Q Ratio. Therefore, using existing word dictionaries for textual analysis of 10-Ks has been found to be informative.

Unlike 10-Ks, which are mandatory annual filing for all public companies, 8-Ks can be filed on a need basis. These additional reports are important to the investors as they are timely notifications of significant changes. Brown and Tucker (2011) introduced a measure to determine how much of the text in 10-K reports for two consecutive years was different, and find that the changes in 10-Ks have a positive association with stock prices. Since the very purpose of 8-Ks is to report changes in a firm, it is important to examine (a) how 8-Ks compare with the changes in 10-Ks for two consecutive years in their degree of informativeness and (b) how much more are the changes in 10-Ks effective in predicting quantitative variables than a single year's entire 10-K. In the paper, we examine these questions from the perspective of attention-score based word list. While 10-Ks have received much attention in the literature, there is only limited exploration of NLP based learning using 8-K documents. An 8-K-powered end-to-end sequence-to-sequence neural network using a Gated Recurrent Units (GRU) plus attention mechanism was proposed to predict future corporate event sequences (Zhai and Zhang, 2019). In another study, Naive Bayes and multi-class Support Vector Machine (SVM) classification were implemented to categorize 8-K contents as financial, operational, legal, administrative, or human resources related (Lee and Lee, 2008).

Before we present the methodology adopted to build the attention scores based word lists to address the above research questions, we provide a discussion of the related literature, followed by a summary of our main findings and an overview of the structure of the rest of this paper.

1.1 Related Work

Financial research utilizes quantitative data for a plethora of important decision and risk analytics questions regarding investment decisions, portfolio optimization, securities risk analysis, and risk assessment and management in a variety of context and risk types (Emerson et al., 2019). In support of these problems, different ML models, such as, Support Vector Machines (SVM), Single Hidden Layer Feed-forward Neural Networks (SLFN) and Multi-layer Perceptrons (MLP) have been used. For instance, for the prediction of future price movements, Nousi et al. (2019) used two sets of features for ML classifiers: (1) handcrafted features formed on the raw order book data and (2) features extracted using ML algorithms. Convolutional Neural Networks (CNN) have also been used to detect price movement patterns in high-

frequency limit order book data (Tsantekidis et al., 2017).

Other models such as Random Forest (Khaidem et al., 2016), XGBoost (Wang and Ni, 2019), Bidirectional Long Short-Term Memory (BiLSTM) and stacked LSTMs (Sardelich and Manandhar, 2018) have been implemented to predict business risk and stock volatility. Specifically, predicting daily stock volatility using news and price data was developed using a neural network-based Bidirectional Long Short-Term Memory (BiLSTM) and stacked LSTMs (Sardelich and Manandhar, 2018). Therefore, in recent decades quantitative data has been richly complemented with qualitative and textual data to enhance the approach for problem solving. Prediction of various quantitative variables is explored using a range of financial textual data extracted from news articles, earning call transcripts, regulatory company filings, analyst reports, and social media text. In this line of research, Tetlock et al. (2008), among the earliest work, created textual features by using negative words in the Harvard-IV-4 TagNeg dictionary and constructed document-term matrices from news stories. These features were used to predict firms' accounting earnings and stock returns. From the early works, the field has come a long way. An ensemble of deep learning models based on CNN, LSTM, and GRU, and supervised models based on Support Vector Regression (SVR) was used to evaluate a benchmark dataset consisting of microblogs and news headline for financial sentiment analysis (Akhtar et al., 2017).

Risk mining identifies a set of risks related to a business area or entity. Two text-based methods were proposed in Nopp and Hanbury (2015) to find bank risks using risk sentiment analysis of CEO letters. The first method is a dictionary-based approach using negative, positive and uncertainty tonality relevant for the banks. The second method predicted the evolution of quantitative risk indicators using classification techniques. Risk Mining involves identifying a set of risks relevant to a business unit or a firm by combining Web mining and Information Extraction (IE) techniques to automatically detect risks even before they materialize. This can be very powerful for various business contexts by making valuable business intelligence available (Leidner and Schilder, 2010). Therefore, besides stock performance and investment analysis, textual analytics can be employed to identify, detect, and measure the impact of risk exposures of a firm. In this paper, we use several risk indicators to judge informativeness of 10-K versus 8-K regulatory filings.

Financial news are important for driving investment decisions process, where financial sentiment analysis faces the challenge of lack of labeled data specific to newly emerging investment domains, such as alternative investment opportunities, crypto-assets, etc. The general-purpose pre-trained language models fail to capture these emerging financial contexts. Chang et al. (2016) proposed a novel tree-structured LSTM to automatically measure the usefulness of financial news using both news and cu-

mulative abnormal returns (CAR). A dual-layer attention-based neural network model was developed to predict stock price movement using the text in financial news (Yang et al., 2018). FinBERT has been proposed to show how BERT can be fine-tuned on the financial sentiment analysis dataset (FiQA) to outperform the general BERT model (Araci, 2019). DeSola et al. (2019) introduced a domain-specific pretrained language model (FinBERT) for financial NLP applications, where the model was trained using 10-K filings from 2017 to 2019, and applied to a variety of financial NLP downstream tasks.

In this work, we use textual data from 8-K and 10-K reports and validate our experimental results by predicting bank variables using Kernel Ridge Regression (Zaki and Meira Jr, 2020). Our work is primarily motivated by a few earlier works (Kogan et al., 2009; Tsai and Wang, 2017; Tsai et al., 2016), where text from 10-K reports is used to predict different financial quantitative variables, or additional textual features are extracted by expanding the L&M sentiment word list semantically and syntactically using word2vec (Mikolov et al., 2013). With a similar motivation, the *uncertainty* word list in L&M dictionary was expanded using word2vec to predict stock volatility (Theil et al., 2018). Theil et al. (2020) further expanded the L&M dictionary by training industry-specific word embedding models using word2vec to predict volatility, analyst forecast error and analyst dispersion. Sedinkina et al. (2019) showed how automatic domain adaption of the L&M sentiment list using word2vec improved the prediction of excess return and volatility.

The aforementioned dictionary expansion approaches use the word2vec model to select the top-k closest words to the words existing in the L&M dictionary. Since word2vec is a static continuous bag-of-words (CBOW) based model, it fails to capture the dynamic context of the words. Thus, in our work, we automatically create a word list based on the dynamic attention scores of the words using the BERT model. However, long length of 10-K and 8-K documents presents a challenge; in previous work, Dereli and Saraçlar (2019) used a CNN based approach to predict volatility using 10-K text clipped at length 20,000 tokens. Since 8-K and 10-K reports in our sample are much longer than 20,000 tokens, in order to avoid losing some trailing text, we use the entire text of 8-K and 10-K filings.

1.2 Contribution and Main Findings

We contribute to the financial analytics literature using NLP to support decision processes based on an automatic word list creation framework. Use of these enhanced word lists is demonstrated in some key performance and risk prediction task for banks. To the best of our knowledge, this is the first exhaustive study on the comparative analysis of 8-Ks and 10-Ks by using automatically created word lists. This study also focuses on comparing 8-Ks with the textual changes in 10-K filings from two consecutive years. The

breadth of the experiments conducted and results obtained indicate that the latest dynamic pre-trained language models like BERT can be useful in uncovering more textual meaning than the traditional static ML methods. Additionally, our empirical results outline the direction for follow-on work where they can be used as baselines for further research.

The rest of the paper is organized as follows. We present the methodology for attention-scores based word list construction in the next section, Section 2. This is followed by describing the data, its pre-processing and the variables used in this study. We investigate the research questions developed earlier in the current section by constructing a responsive experimental study and discuss the results in Section 4. We conclude the paper with a summary and discussion of future work in Section 5.

2 Word List Building and Prediction Methodology

Our primary goal is to analyze the 8-K filings, 10-K filings and changes in 10-K filings in their degree of informativeness towards the selected bank financial indicators. The key background to the research questions is how can the textual features be expanded with increasing volumes of textual data available for the firms. We propose a novel method to automatically create a word list from public companies' 10-K filings. This word list is used in our experiments to predict banks' financial indicators and compare the performance against a word2vec approach used for expanding the L&M word list (Tsai and Wang, 2017). The methodology of the analysis is described next.

A word can be represented via a numeric word embedding vector in several ways. The most naive way is one-hot encoding, where the categorical variables are converted into binary variables. However, one-hot encoding is computationally expensive because of high dimensionality of the corpus vocabulary. Two different language model architectures (word2vec) were proposed by Mikolov et al. (2013) to create word embeddings: (1) CBOW or Continuous Bag Of Words - predicts a word given a sequence of words and (2) skip-gram - predicts a sequence of words given a word. Although these models capture the static semantic context of words, they fail to incorporate the dynamic contextual information, where the same word could have different meanings in different contexts, or polysemy. The L&M dictionary expansion approach described in Tsai et al. (2016) showed an improvement over the original L&M word list, however, it failed in the case of *polysemous* words since the expanded word list created using word2vec is static. We address this issue by creating a word list that is based on the dynamic contexts of words.

2.1 Automatic Word List using Attention Score

The BERT model (Devlin et al., 2018) based on the Transformer architecture (Vaswani et al., 2017) has shown breakthrough results on almost all the NLP benchmark tasks, such as question answering, named entity recognition, and so on. In our work, we propose a novel method to create a word list automatically using the word-level attention scores in a sentence based on the BERT contextual language representation model.

2.1.1 Bidirectional Encoder Representations from Transformers (BERT)

Static word embedding models cannot capture word *polysemy* since they generate the same embedding for the same word in all contexts. Contextualized word embeddings address this issue by dynamically capturing word semantics in different contexts. The most common contextual language models are Embeddings from Language Models (ELMo) (Peters et al., 2018), OpenAI Generative Pre-trained Transformer (GPT) (Radford et al., 2018) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). There are two contexts for a given word in a sentence, namely its right and left contexts. ELMo is shallow and bidirectional, OpenAI GPT follows just one direction (left to right), while BERT is deeply bidirectional.

Neural machine translation (NMT) uses an artificial neural network to predict the likelihood of a sequence of words by modeling entire sentences. The traditional models used are Recurrent Neural Networks (RNN), LSTM, and GRU. These models are based on sequential processing over time. In such Encoder-Decoder models, the meaning of the entire source sequence is condensed into the final encoder state, which typically becomes burdensome if the source sequence is too long. This problem is solved by using the attention mechanism which allows the decoder to pay attention to different parts of the source sequence at different decoding steps.

In a self-attention mechanism, the source and target sequences are the same, and thus, it is also known as intra-attention. This means that each word attends to the rest of the words in a sequence to determine its importance in the entire sequence. Vaswani et al. (2017) introduced the Transformer model based on a self-attention mechanism. Unlike RNNs, a Transformer does not require the sequence to be processed in a sequential order, instead it can be processed in parallel. Since it does not have recurrent networks that can remember how sequences are fed into a model, it uses positional encoding of different words. BERT combines Transformers with a masked language model (to predict masked words) and a next sequence prediction task; it uses context in both directions and is deep, containing many layers,

from a base of 12 layers to a high of 24 layers.

We discuss the self-attention characteristics of the BERT model. Consider there are n tokens in an input sentence, then the first step is to create three vectors from each of the encoder's input vectors. Each input vector, h_i, is transformed into query, key, and value vectors, q_i, k_i, v_i, respectively, through separate linear transformations. These abstractions are useful for calculating attention. The attention head computes attention weights, α, as given in Eq. (1), between all pairs of words as softmax-normalized dot products between the query and key vectors. The output vector o_i of the attention head is a weighted sum of the value vectors, as shown in Eq. (2) below.

$$\alpha_{ij} = \frac{\exp\left(q_i^T k_j\right)}{\sum_{l=1}^{n} \exp\left(q_i^T k_l\right)}. \tag{1}$$

$$o_i = \sum_{j=1}^{n} \alpha_{ij} v_j. \tag{2}$$

The above self-attention function can be simplified as follows:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

where d_k is the dimension of the key vector and Q, K, V are Query, Key, and Value matrices, respectively.

In this work, we use the BERT-Base, Uncased model[2] with 12 hidden layers and 12 attention heads on the 10-K filings corpus for years 2006 through 2016. In the BERT model, the last hidden layer is very close to the target during the training process, and therefore, may be biased to those target values[3]. Therefore, we use second-to-the-last hidden layer and attention head to compute the attention scores of words in a sentence. Attention scores are computed for all the words in all the sentences in a document.

> the following discussion is intended to provide information to facilitate the understanding and assessment of the consolidated financial condition and results of operations of the banco ##rp and its subsidiaries . it should be read in conjunction with the audit ##ed consolidated financial statements and notes appearing elsewhere in this annual report on form 10 - k .

Figure 1: A sentence taken from a 10-K report Item 7/7A (combined). The figure shows the word "understanding" highlighted in red to have a higher attention score of 0.28 in the sentence.

Fig. 1 shows an instance of a sentence from a 10-K report to demonstrate words getting higher attention score in a sentence. We add all the attention scores for common words in all the sentences of a document. These words and their attention scores are stored in a dictionary. We then average all the

[2]https://github.com/google-research/bert

[3]https://github.com/hanxiao/bert-as-service

common words across the entire corpus, along with removing English stop-words. After being stemmed using *nltk PorterStemmer*[4], averages are once again computed for all the words with a common stem. Finally, words with attention score greater than a threshold of 1.0 are extracted. BERT uses the *wordpiece* tokenizer (Sennrich et al., 2015) to create sub-words to handle the issue of out-of-vocabulary words. This poses the challenge of reconstructing these sub-words back into words, also known as word alignment. We create a new sub-word list consisting of a total of 2329 sub-words, without any external domain expertise modification or using any text from the 8-Ks. We evaluate the effectiveness of our sub-word list on previously unseen data, such as 8-Ks, for different prediction tasks.

2.2 Feature Representation, Prediction And Evaluation

The textual features for our prediction experiments are computed using TF-IDF values shown in Eq. (4) using *scikit-learn*[5].

$$tf\text{-}idf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \tag{4}$$

where $tf_{i,j}$ is the number of occurrences of word i in document j, df_i is the number of documents containing the word i, and N is the total number of documents. The TF-IDF scores illustrate the importance of words from the word list in the documents. Our objective for defining the textual features this way is to determine whether the TF-IDF based numeric value or raw word scores are better features for predicting bank financial indicators.

For quantitatively analyzing the relative informativeness of banks' 8-K and 10-K reports, we perform text regression analysis to predict banks' financial indicator variables. For a set of financial reports $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, where each $\mathbf{s}_i \in \mathbb{R}^p$, our goal is to predict several bank financial indicator variables, denoted by $y_i, i \in \{1, \ldots, K\}$. Thus, the prediction problem is stated as follows:

$$\hat{y}_i = f(\mathbf{s_i}; \mathbf{w}), \tag{5}$$

where the goal is to learn a p-dimensional vector $\mathbf{w}$ from the training data $D = \left\{\left(\mathbf{s}_i, y_i\right) | \mathbf{s}_i \in \mathbb{R}^p, y_i \in \mathbb{R}\right\}$ to predict the response variables, $y_i, i \in \{1, \ldots, K\}$. The accuracy of prediction is judged using mean squared error (MSE) given as follows:

$$\text{mean squared error (MSE)} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2. \tag{6}$$

[4]https://www.nltk.org/

[5]https://scikit-learn.org/

We next describe the data and the bank financial indicator variables chosen to address the research questions discussed in Section 1, where all the comparisons for prediction accuracy performance are conducted based on the above (Eq. (6)) MSE performance metric.

3 Data Description and Experimental Setup

In this section, we describe the dataset, data sources and pre-processing required to support this study. The financial indicator variables chosen for the analysis of the research questions are discussed, along with providing a justification for their selection.

3.1 Dataset and Pre-Processing

We downloaded a corpus of 8-K and 10-K reports used by Chen et al. (2017), which contains 8-K and 10-K reports of 578 bank holding companies (BHCs) for the period of 2006 through 2012 and 2006 through 2016, respectively. As described earlier, 8-K filings report changes that occur in a firm, and therefore are not mandatory if there no reportable changes, whereas a 10-K filing is a comprehensive report, and public firms are required to file them annually.

10-Ks are fairly long documents with a comprehensive report on a firm and can be complicated due to numerous displays and tables. A 10-K filing has 15 sections, where for the purposes of this study, we focus on Item 1: "Risk factors," Item 7: "Management's Discussion and Analysis" (MD&A)" and Item 7A: "Quantitative and Qualitative Disclosures about Market Risks." We chose these sections for their important content pertaining to a firm's risk-related information and future performance discussion.

We selected six key quantitative bank financial indicators to study the research questions. These variables are: (1) Return on Assets (ROA), (2) Earnings per Share (EPS), (3) Tobin's Q Ratio, (4) Tier 1 Capital Ratio, (5) Leverage Ratio, and (6) Z-score. Some of these variables are generic financial indicators relevant for all firms, but others, such as Tier 1 Capital Ratio and Z-score are particularly relevant for bank holding companies. The data for these variables are obtained from CompuStat and we describe these variables briefly later in this section.

The 10-Ks are typically filed at the end of each fiscal year whereas 8-Ks are filed at any time of the year, including multiple filings per year, whenever there are reportable changes in the firm. Since the changes reported in 8-Ks accumulate through the year, we group the 8-Ks from the first half-year [Quarter 1 (Q1) and Quarter 2 (Q2)] to compare learning from them against 8-Ks grouped for the entire year. We merged all 8-Ks month-wise and chronologically into a single 8-K document. For section-wise analysis of 10-Ks,

	# documents
total 10-K	5321
after extracting Items 1A, 7/7A 10-K	2167
total 8-K	925
8-K [Q1+Q2, full]	864
common 8-K and 10-K	712
8-K vs 10-K changes	710

Table 1: Number of documents of each type.

as stated earlier, we chose Items 1A, 7, and 7A, and combined Items 7 and 7A into one document. We filter out documents less than 250 bytes in size to eliminate empty and non-informative documents.

Statistics for the filtered dataset are reported in Table 1. Each 8-K and 10-K file is tokenized using the *nltk* word tokenizer. HTML tags are removed using a Python library called *beautifulsoup4*[6] and punctuations are also removed except for '.', '%', and '$' signs. We replaced all the numerals with a '#' and removed all the English stop-words, followed by stemming the tokens using the *nltk PorterStemmer* library.

Since 8-Ks represent changes in a company in the previous year for a corresponding 10-K filed in the current fiscal year, we denote 8-Ks with a timestamp t, for the previous year, and corresponding 10-K with the timestamp $t+1$. We thus represent the 8-K and 10-K corpus as 8-K^t[Q1 + Q2], 8-K^t, 10K^t and 10K^{t+1}. We extract the following from each document: **(A) words**, **(B) sentiment words** from the L&M dictionary[7], which is a collection of six sentiment lists: (i) negative (ii) positive (iii) uncertainty (iv) litigious (v) strong modal (vi) weak modal, **(C) syntactically expanded sentiment words** (Tsai and Wang, 2017), and finally, **(D) our attention score based words**.

In this work, we reproduce the dictionary expansion approach of Tsai et al. (2016) to extract the syntactically expanded sentiment words. Instead of semantically expanded word list, we choose syntactically expanded word list since the latter has shown improvement over both, sentiment words and semantically expanded sentiment words (Tsai and Wang, 2014). Based on the documents summarized in Table 1, we define two datasets for our study: Dataset-1 consisting of all 8-Ks and 10-Ks, and Dataset-

[6] https://pypi.org/project/beautifulsoup4/

[7] https://sraf.nd.edu/textual-analysis/resources/

2 consisting of 8-Ks and 10-Ks paired by bank CIK number for each year. This dataset identification is done to answer RQ1, which requires all 8-Ks and 10-Ks for which Dataset-1 is constructed. For RQ2 and RQ3, we require banks for which both 8-Ks and 10-Ks are available for each year, for which Dataset-2 is utilized. We next describe the bank financial indicator variables chosen for this study.

3.2 Bank Variables

We discuss the six key bank financial indicators chosen to conduct our experiments in support of addressing the three research questions, RQ1-RQ3, described in Section 1. Some of these variables are generic financial indicators relevant to all firms, such as ROA, EPS, Tobin's Q Ratio and Leverage Ratio; others are specifically relevant for banks, such as Tier 1 Capital Ratio and Z-score. Additionally, ROA, EPS, Tobin's Q Ratio are performance indicators, while Tier 1 Capital Ratio, Leverage Ratio and Z-Score are measures of risk for a bank. Data for these indicators is extracted from the Compustat database (Gupta and Owusu, 2019). To facilitate the prediction task, as the values of the selected bank indicators are not in a uniform range, we scale all the bank indicator variables using *sklearn MinMaxScaler* by the following formula.

$$X_{\text{new},i} = \frac{X_i - \min(X)}{\max(X) - \min(X)}, \tag{7}$$

where the range for all variables after the above transformation is between 0 and 1. The distribution for all the six bank variables are shown in box plots in Fig. 2 and the summary statistics are reported in Table 2.

	ROA	**EPS**	**Tobin's Q Ratio**	**Tier 1 Capital Ratio**	**Leverage Ratio**	**Z-score**
mean	4.41E-01	5.97E-01	7.32E-02	2.32E-01	4.86E-01	9.40E-02
std. dev.	4.77E-02	5.23E-02	5.91E-02	7.92E-02	3.07E-02	1.04E-01

Table 2: Summary statistics for the six bank financial indicator variables.

We describe each variable briefly, along with a reason for the choice for the purpose of the study.

1. **Return on Assets (RoA):** combines information from a firm's income and balance sheet statements, therefore is a combination of a stock and flow variables. It measures profitability as a ratio of the firm's annual net income to the firm's total assets, given in Eq. (8). Clearly, this variable is

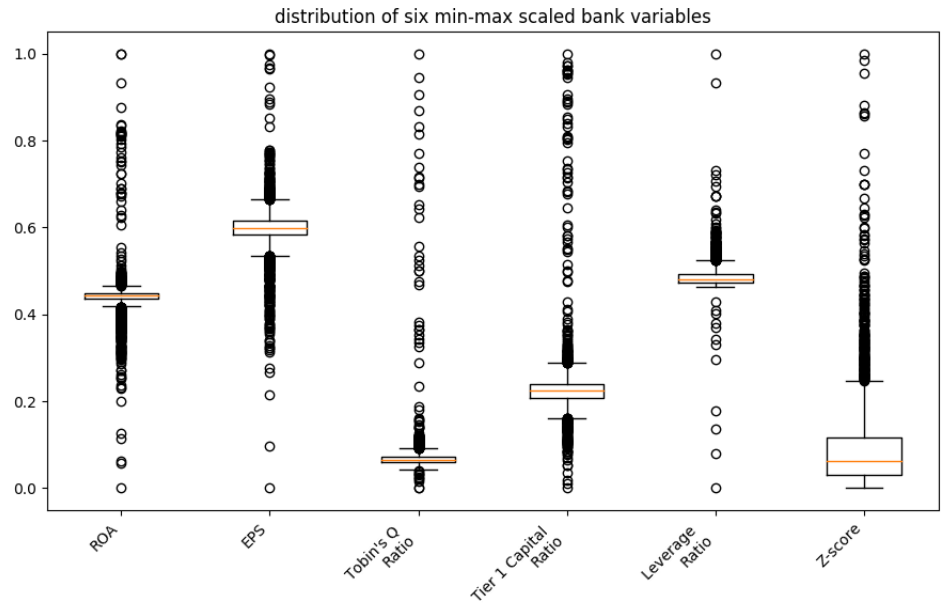


Figure 2: Box plots for all the six bank financial indicator variables that show the min-max scaled values for each variable.

relevant for all firms in indicating how beneficially the firm's assets are being used to generate income, irrespective of how the firm is financed. Return on Equity (RoE) is the related profitability measure specific to how profitable the firm is for its shareholders. We chose this variable to capture overall profitability of a bank, where the higher the ROA, the more efficiently the bank is generating income from its assets. RoA varies across firms and by industry, and while the total assets may not change for a firm that dramatically year to year, the 10-Ks and 8-Ks can give information regarding anticipated net income to produce a prediction of a future year's RoA.

$$\text{ROA} = \frac{\text{Net Income}}{\text{Total Assets}} \tag{8}$$

2. **Earning per share (EPS):** also measures a firm's profitability, but with a focus on common stockholders of the firm. It measures what share of the firm's income is due to each share of the stock of the firm once the payout to preferred stockholders is made, as calculated in Eq. (9). Once again, this profitability indicator is relevant to all firms and is particularly informative when it is used in the Price-to-EPS (PE) ratio. Therefore, predicting EPS using information from 8-Ks and 10-Ks is highly desirable.

$$\text{Earnings per Share} = \frac{\text{Net Income - Preferred Dividends}}{\text{End-of-Period Common Shares Outstanding}} \tag{9}$$

3. **Tobin's Q Ratio:** is a performance indicator for any firm and bears a qualitative theme in its definition. It is the ratio of market value of a firm to its book value. However, since market value of a firm's liability is not easy to compute, due to lack of price observability from illiquid liabilities, it is computed often by book value of liabilities as given in Eq. (10). It is entirely constructed using stock variables obtained from the balance sheet of a firm, and basically measures how much value the management of a firm has generated beyond the book value of the firm. The book value terms in the ratio are obviously not the interesting part in predictive sense, therefore the most interesting input in this ratio is the market value of the firm's equity, which includes price per share and number of shares outstanding for the firm in a future year. We hope to be able to predict this information using banks' 8-Ks and 10-K documents.

$$\text{Tobin's Q Ratio} = \frac{\text{Equity Market Value } + \text{ Liabilities Book Value}}{\text{Equity Book Value} + \text{Liabilities Book Value}} \tag{10}$$

4. **Leverage Ratio:** How a firm finances its assets is very crucial, given borrowing is both costly and risky. Debt financing is costly since it is an obligation a firm is required to payout interest and repayment of principal for. It is risky since as income fluctuates, the firm can find itself unable to meet the debt obligations. Leverage ratio is computed as given in Eq. (11), and is decisively a risk indicator relevant for all types of firms. Leverage ratio varies by industries and for banks it is a highly important indicator since a bank's balance sheet and the entire business model relies on liabilities of a variety of kinds, ranging from retail checking accounts, savings accounts, certificates of deposits, to the bonds issued by the bank. A high leverage ratio indicates high risk for a firm using a large amount of debt to finance its assets. Banks are the most leveraged institutions in the United States. Leverage ratio is also a stock variable with information coming from a firm's balance sheet. Changes in a firm's borrowing behavior, and asset decisions reflected in the market value of its equity are features that can be learned from the 8-Ks and 10-K of the bank to predict the bank's evolving leverage.

$$\text{Leverage Ratio} = \frac{\text{Average Total Assets}}{\text{Average Equity}} \tag{11}$$

5. **Tier 1 Capital Ratio:** This is a bank specific variable, given that banks function under stringent regulatory framework in any economy. The indicator measures a bank's equity capital and disclosed reserves, called the Tier 1 capital, relative to the bank's risk-weighted assets given in Eq. (12). This risk indicator measures how well a bank is able to cover its risk-weighted assets, which are all assets systematically weighted for their credit risk, by its equity capital and disclosed reserves. Should the

value of the assets drop due to increased or realized risk, the Tier 1 Capital should be able to cushion the loss, thus protecting the bank's liabilities. Change in a bank's quality of assets is the biggest risk, which we hope to infer from the bank's 8-K and 10-K filings.

$$\text{Tier 1 Capital Ratio} = \frac{\text{Tier 1 Capital}}{\text{Total risk-weighted assets}} \tag{12}$$

6. **Z-score:** Z-score, the final indicator which is also specific to banks, links a bank's capitalization with its return (ROA) and risk (volatility of ROA). Z-score [8] is an indicator of bank's risk proposed in (Roy, 1952). Z-score relates a bank's capital to variability in its assets' returns in order to measure the amount of variability the bank will be able to absorb without becoming insolvent. The Z-score combines a bank's Return on Assets, ROA, with its capital-to-assets ratio, CAR=Equity/Assets, and compares it against the bank's volatility in ROA as given in Eq. (13).

$$Z\text{-score} = \frac{ROA + CAR}{\sigma(ROA)} \tag{13}$$

where, $\sigma(ROA)$ is the standard deviation of ROA for a specific time period. Given that ROA is the ratio of net income to total assets and CAR is the ratio of equity to assets, the numerator of Eq. (13) is a ratio of net income plus equity to assets. Therefore, the numerator of the Z-score measures the resources available to the firm as a fraction of the total firm value, which is then compared to the riskiness of the ROA. Learning about the Z-score from bank's 8-K and 10-Ks would be very instructive to the banking regulators.

At the federal level, three different regulatory bodies, the Federal Deposit Insurance Corporation (FDIC), the Federal Reserve, and the Comptroller of the Currency, have the regulatory responsibility and oversight over banks in the United States. There may be additional state level entities enforcing their additional oversight on banking and savings & loans (S&L) enterprises. The regulatory guideline restricts bank's lending relative to how much capital the bank assigns to its assets. This is important because banks can "write down" the capital part of their assets if there is a drop in total asset value. As such, assets financed by debt cannot be written down since these funds are owed to the bank's bondholders and depositors. Additionally, during the 2007-2009 global financial crisis, many banks were found to have insufficient capital to withstand their losses and remain solvent. Internationally coordinated Basel III standards were enforced so as to increase bank's capital buffers and make sure that they are able to withstand financial distress before becoming insolvent in the future. The above variables specifically picked

[8] Not to be confused with Altman's Z-score (Altman, 1968) built by a regression analysis to measure creditworthiness of any firm in terms of a set of financial ratios.

for bank holding companies are hence important selection for not just investor guidance, but also for regulatory consumption.

4 Research Questions Results and Discussion

For determining the degree of informativeness of 8-K and 10-K filings and evaluating effectiveness of the novel attention-based word list, a text regression approach similar to Kogan et al. (2009) and Tsai et al. (2016) is used. Textual features are separately constructed as described in Section 2.2 using (i) words, (ii) L&M sentiment words, (iii) syntactically expanded L&M sentiment words, and (iv) attention words to compare their effectiveness. We also compare the predictive power of textual features against quantitative variables, which is the numeric baseline consisting of linear regression and historical bank financial indicators data used for predicting their future values. All experiments are performed on a 2.3GHz Intel Xeon E5-2670 v3 Processor machine, with 251GB RAM and a Tesla K40m GPU, and implemented in Python 3.6.

regression model	words	sentiment words	syntactically expanded sentiment words
Linear	1.05E-01	7.43E-01	1.27E-01
SVR (kernel='rbf')	1.57E-01	1.49E-01	1.49E-01
Random Forest	8.01E-02	8.45E-02	8.20E-02
MLP (100)	9.48E-02	8.75E-02	**7.58E-02**
MLP (1000, 500, 100)	8.11E-02	7.96E-02	7.72E-02
MLP (10000, 5000, 500, 100)	9.50E-02	8.87E-02	7.85E-02
Ridge (kernel='poly')	**7.55E-02**	**7.76E-02**	8.12E-02
Lasso	8.12E-02	8.12E-02	8.12E-02
ElasticNet	8.20E-02	8.20E-02	8.20E-02
Gradient Boost	8.09E-02	8.42E-02	8.10E-02

Table 3: Comparison of different regression models using MSE (on 10-Ks only: Dataset-1). Using 10-fold cross-validation shows that ridge regression gives an overall better performance while MLP does well just in one case for the syntactically expanded sentiment words.

Although Kogan et al. (2009) and Tsai et al. (2016) used Support Vector Regression (SVR), we tested several other regression techniques using different textual features extracted from all the 10-Ks of Dataset-1 and using a 10-fold cross-validation. The results are tabulated in Table 3. Instead of SVR, we find Ridge Regression to be an overall better performing regression technique. Therefore, for conducting the experiments for investigating the research questions, we use *sklearn KernelRidge* implementation of Ridge Regression (RR) with parameters set at *kernel = 'poly', alpha=0.1, gamma=0.1, degree=3*. We develop our experimental study and discuss the results for addressing the three main research questions stated in Section 1, starting with the first research question.

RQ1: Does text content of 8-K and 10-K filings uncover any useful information about a bank?

We analyzed textual features defined on different collections and segments of 8-Ks and 10-Ks to study their textual predictive power for the set of six bank financial indicators. The baseline to compare the enhanced role of textually guided prediction is defined by predicting the financial indicator in terms of their own historical data. These baseline results are given in the first column of Table 4 and Table 5. Thereafter, incrementally different set of textual features are added to the prediction model to evaluate their efficacy in improving the quality of prediction as judged by the Mean Square Error (MSE) measure. These results are presented in the remaining columns of these tables.

In Table 4, the focus is on textual features defined based on first half of a year's 8-K filings and full year's 8-K filings. Among textual features, we consider the choices of textual features based on (i) words, (ii) sentiment words, (iii) syntactically expanded sentiment words, and (iv) attention words. This variety of textual features is considered to compare the best information extraction capability to achieve the goals of the prediction task, as well as address the research question of usefulness of 8-K and 10-K textual content. The MSE values in bold show the least values among the numeric versus textual features. We find that the textual content of 8-Ks is best summarized by extracting all words to define features. Specific nuances based word extraction to define textual features doesn't provide any advantage over the MSE obtained from the baseline. Among the bank variables, textual features from 8-K filings end up adding value only in some cases.

The prediction accuracy improves by adding textual features for prediction of *ROA*, *EPS*, and *Z-score*, but infuse noise in the prediction of *Tobin's Q Ratio*, *Tier 1 Capital Ratio*, and *Leverage Ratio*. As discussed earlier, *ROA* and *EPS* are profitability indicators and *Z-score* is a bank specific risk indicator, but all three of these variables are strongly inter-related and depend on information from balance sheet and income statement in terms of net income of the bank. Among groups of 8-Ks, considering just the first half or full

		8-K [Q1 + Q2]				8-K [full]			
	baseline [numeric]	**W [8848]**	**SW [996]**	**SESW [3138]**	**AW [1781]**	**W [11697]**	**SW [1106]**	**SESW [3514]**	**AW [1862]**
ROA	9.73E-04	**9.17E-04**	1.03E-03	9.52E-04	*9.23E-04*	**7.59E-04**	8.87E-04	7.93E-04	*7.60E-04*
EPS	2.61E-03	**2.53E-03**	2.98E-03	2.64E-03	*2.59E-03*	**2.18E-03**	2.48E-03	2.29E-03	*2.21E-03*
Tobin's Q Ratio	**3.81E-05**	6.48E-05	8.02E-05	6.60E-05	*6.91E-05*	5.89E-05	7.02E-05	6.11E-05	*6.15E-05*
Tier 1 Capital Ratio	**1.86E-04**	6.02E-04	6.97E-04	5.94E-04	*6.34E-04*	5.45E-04	6.25E-04	5.45E-04	*5.79E-04*
Leverage Ratio	**1.07E-03**	1.33E-03	1.61E-03	1.39E-03	*1.34E-03*	1.22E-03	1.40E-03	1.30E-03	*1.23E-03*
Z-score	6.59E-03	**6.36E-03**	6.69E-03	6.48E-03	*6.70E-03*	**5.77E-03**	6.25E-03	6.10E-03	*6.04E-03*

W: words; **SW:** sentiment words; **SESW:** syntactically expanded sentiment words; **AW:** attention words

*number inside square brackets represents feature dimension

Table 4: MSE values for 8-K. Values in bold show the least MSE values whereas values in italics show the least MSE values between sentiment words and attention words.

year's 8-K both improve the prediction compared to the numeric baseline, although having access to all year's 8-Ks improves the prediction a little further. The *Tobin's Q Ratio*, *Tier 1 Capital Ratio*, and *Leverage Ratio* variables are strongly balance sheet variables and incremental information of updates provided in 8-Ks proves to not be very informative for their prediction.

In Table 5, similar results are presented for baseline for all the bank financial indicators against different types of textual features extracted from bank's 10-K filings. For comparison of relative informativeness of different sections of the 10-K document, we consider textual features based only on Item 1A, only on Item 7/7A and complete 10-K. The baselines in Table 4 and Table 5 differ since the sample of banks and years for which documents of each type are available are different.

Once again textual features extracted from 10-K filings perform better than the baseline for *ROA*, *EPS*, and *Z-score*. Among *Tobin's Q Ratio*, *Tier 1 Capital Ratio*, and *Leverage Ratio*, for which 8-Ks didn't add value, 10-K textual features are able to slightly improve accuracy of prediction for *Tobin's Q Ratio*. This is promising as *Tobin's Q Ratio* is the most qualitative flavored variable in the set of financial indicators considered in this study. In terms of sections of 10-K filings, Item 7/7A is uniformly more informative than Item 1A, however considering the entire 10-K proves to be most helpful in improving prediction

		10-K [Item 1A]				10-K [Item 7/7A]				10-K [full]			
	baseline [numeric]	**W [7072]**	**SW [929]**	**SESW [3125]**	**AW [1852]**	**W [11367]**	**SW [965]**	**SESW [3752]**	**AW [1949]**	**W [23624]**	**SW [1173]**	**SESW [4908]**	**AW [2042]**
ROA	**1.38E-03**	**1.38E-03**	1.33E-03	1.42E-03	*1.32E-03*	**1.07E-03**	1.25E-03	1.12E-03	*1.13E-03*	**1.05E-03**	1.21E-03	1.20E-03	*1.15E-03*
EPS	**1.85E-03**	**1.79E-03**	1.97E-03	1.88E-03	*1.87E-03*	**1.66E-03**	2.02E-03	1.80E-03	*1.76E-03*	**1.65E-03**	1.93E-03	1.73E-03	*1.75E-03*
Tobin's Q Ratio	**5.04E-04**	5.79E-04	8.76E-04	7.08E-04	*7.09E-04*	5.82E-04	9.12E-04	7.24E-04	*6.33E-04*	**5.01E-04**	9.14E-04	7.34E-04	*6.55E-04*
Tier 1 Capital Ratio	**3.88E-04**	8.19E-04	1.45E-03	9.69E-04	*9.91E-04*	9.18E-04	1.90E-03	1.31E-03	*1.13E-03*	8.53E-04	1.73E-03	1.33E-03	*1.22E-03*
Leverage Ratio	**7.75E-04**	8.29E-04	8.94E-04	8.38E-04	*8.41E-04*	8.18E-04	9.03E-04	8.40E-04	*8.53E-04*	8.10E-04	8.89E-04	8.29E-04	*8.22E-04*
Z-score	9.22E-03	**8.47E-03**	8.59E-03	8.61E-03	*8.48E-03*	**8.35E-03**	9.20E-03	8.59E-03	*8.41E-03*	**8.24E-03**	8.50E-03	8.24E-03	*8.33E-03*

W: words; **SW:** sentiment words; **SESW:** syntactically expanded sentiment words; **AW:** attention words

*number inside square brackets represents feature dimension

Table 5: MSE values for Item 1A, Item7/7A combined and full 10-K. Values in bold show the least MSE values whereas values in italic show the least MSE values between sentiment words and attention words.

accuracy. Among types of textual features, once again considering all words is most effective. This begs the question we posed in **RQ1.1**, namely when can sentiment words be useful in prediction tasks.

RQ1.1: Are sentiment words helpful for constructing textual features for prediction tasks?

To determine how helpful sentiment words are for predicting key financial indicators for banks, in Table 4 and Table 5 we focus on the columns for textual features defined using (i) sentiment words (SW) and (ii) syntactically expanded sentiment words (SESW), and compare the results with the remaining columns. While SW and SESW based textual features never win in improving performance of the prediction tasks, we see that SESW textual features in all cases perform better than the SW textual features.

Besides comparing the two sentiment words based textual features, we also compare the performance of textual features made from the automatically created attention based words (AW), and specifically compare their performance against the L&M sentiment word (SW) textual features. Table 4 and Table 5 have the MSE values in italics for the case of lower values among these two word lists. The AW textual features perform better than the SW textual features in all 8-K and 10-K prediction experiments. In fact, AW textual features perform better than SESW textual features in almost half the cases among all the bank variables and textual content from 10-K Item 1A, 10-K Item7/7A, 10-K [full], 8-K [Q1 + Q2], and 8-K [full], and when it trails, it does so with a very small margin. Therefore, features extracted using our automatic word list are better than L&M sentiment word list and comparable to the syntactically expanded sentiment word list.

In summary, while not for all bank financial indicators prediction tasks, text data is definitely useful for predicting some important profitability and risk indicator bank indicators. In all experimental comparisons, the case with all words based textual features taken from full 10-K document and full year of 8-K documents provide the greatest reduction in the MSE in the prediction tasks. Therefore, we conclude on the first research question in the affirmative – text content of these regulatory filings does provide useful information. The answer, however, for the sub-question is mixed. The original L&M sentiment words textual features are the weakest among the word options considered. While different sentiment and attention words textual features don't outperform the full word textual features, they provide indications for improvements possible in information extraction from textual content for prediction tasks.

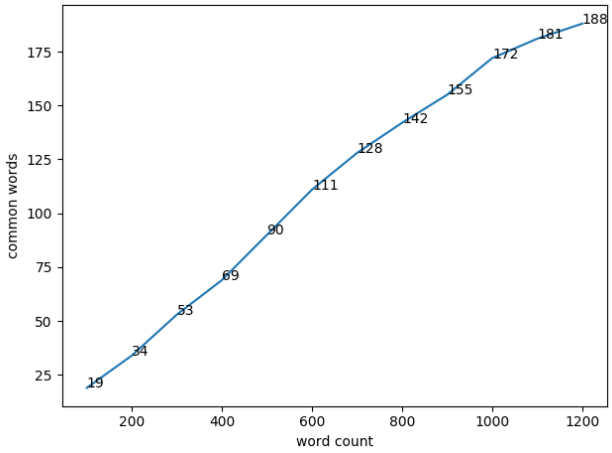
Comparison of Words in Different Word Lists

In the light of the relative performance of the different word lists obtained above, we present a comparison of actual words included in the L&M sentiment word list and the attention score based word list. The results in Table 4 and Table 5 showed that the attention score words based textual features outperformed in all experiments over L&M sentiment words based textual features. We identify the top 1200 words by word count in each document type from both lists and plot the common words by the position they hold in the word count. The plots in Fig. 3 and Fig. 4 show the matched words between attention words and L&M words for 10-Ks Items 1A and 7/7A, and 8-Ks [full] and 10-Ks [full], respectively. Among the 1200 top words by word count, there are only 162 to 188 words in common, which span the entire word count range in the x-axis of these plots. 10-K Item 7/7A has the least number of common words and 10-K Item 1A has the highest number of common words. Beyond the common words, the differences in these two word lists explain the differences seen in results in Table 4 and Table 5.

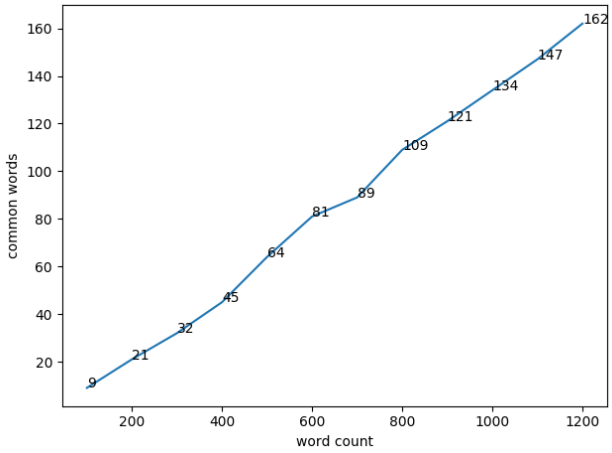
Having established informativeness of both 8-K and 10-K filings of banks, we now turn to the second research question of this study, namely, the relative importance of 8-K versus 10-K filings. The results presented in Table 4 and Table 5 showed that 10-K textual features better predicted more bank financial indicators, hence providing some insight on this research question. We next investigate the question more thoroughly.

RQ2: How do bank's 8-K filings compare with 10-K documents in terms of their informativeness for prediction tasks?

Since both 8-Ks and 10-Ks report a company's updates and performance, we study in exactly what way 8-Ks content differs in informativeness from that in 10-Ks in predicting the bank financial indicators.

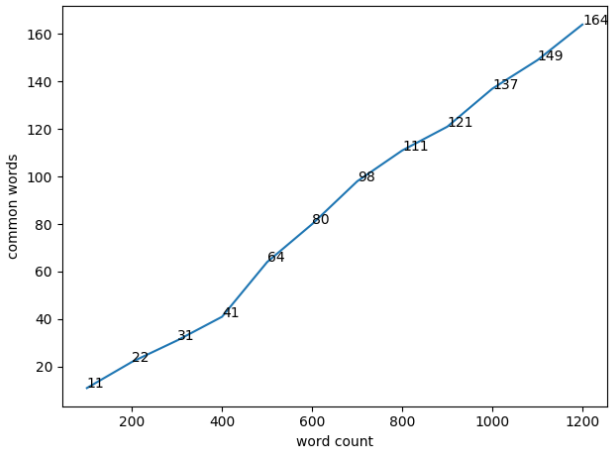


(a) 10-K Item 1A

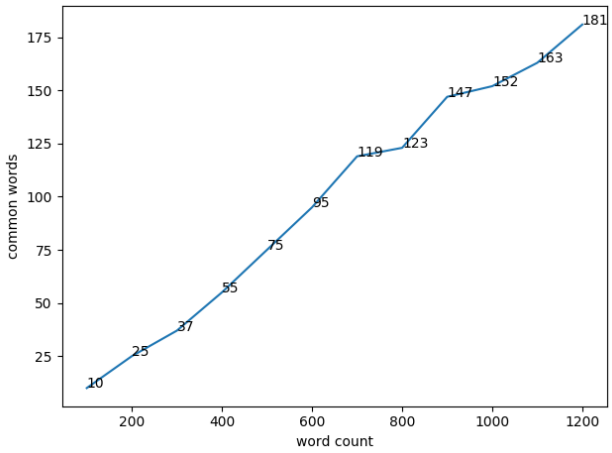


(b) 10-K Item 7/7A

Figure 3: Common words for 10-K Item 1 and 10-K Item 7/7A sentiment and attention word lists for top 1200 words.



(a) 10-K [full]



(b) 8-K [full]

Figure 4: Common words for 10-K [full] and 8-K 8-K [full] sentiment and attention word lists for top 1200 words.

Notation on 10-K is important for this discussion, where 10-K^t denotes the filing made in year t with information on the firm from year $t-1$. 8-Ks filed in a year are all updates from that year, and 10-K^{t+1} filed in year $t+1$ should be expected to include all the information in 8-Ks of year t. Informativeness comparison of 8-Ks versus 10-Ks effectively becomes one of examining if 8-Ks have some unique information that neither 10-K^t nor 10-K^{t+1} offer. We construct several experiments to address this question, summarized in Table 6. For the same six bank financial indicators, we conduct the prediction using Ridge Regression

under baseline and compare it after including textual features using all the words extracted from six different textual content: (i) 8-K^t (ii) 10-K^{t+1} Item 1A (iii) 10-K^{t+1} Item 7/7A (iv) 10-K^{t+1} [full] (v) 8-K^t + 10-K^t [full] (vi) 8-K^t + 10-K^{t+1} [full]. Experiments here after for addressing **RQ2** and **RQ3** are conducted using all words since performance of this category was found to be the best in Table 4 and Table 5.

For the datasets containing 8-K reports, we evaluate two scenarios, namely, using the 8-K from only the first two quarters (denoted 8-K[Q1+Q2] in the second to last column), and using the full 8-K (denoted 8-K[full] in the last column). When content of 8-Ks and 10-Ks are combined, this is done using two operations: (1) `concatenation` and (2) `sum` of their document vectors. We report results from the `concatenation` operation since these were found to be consistently better than the `sum` operation.

If 10-Ks truly incorporate all information relevant for stakeholders, investors and regulators of a bank, intuition would suggest that the case (iv) 10-K^{t+1} [full] among the above 6 content cases should perform the best for all bank variables, offering the most up-to-date comprehensive textual information. Table 6 shows the relative prediction performance for all the bank financial indicators and all the 6 textual content cases. The best performing experiment is marked in bold text. Compared to the baseline, as seen in the analysis for **RQ1**, textual features improve the prediction performance for the same 4 of the 6 bank variables, namely *ROA*, *EPS*, *Z-score*, and *Tobin's Q Ratio*. We can safely ignore the column where half year of 8-K content is used, since this restricted use of 8-K information never generates better outcomes, even though in cases where 8-K information is valued, both half and full year of 8-K information is helpful.

Quite interestingly, instead of 10-K^{t+1} [full], *ROA*, *EPS*, *Z-score*, and *Tobin's Q Ratio* prediction is most helped by the 8-K^t [full] + 10-K^t [full] content. What is surprising is that 8-K^t + 10-K^t [full] content even beats 8-K^t + 10-K^{t+1} [full] content in predicting these variables. In Table 5, 10-K [full] showed a marginal reduction in the MSE from baseline for predicting *Tobin's Q Ratio*, however with 8-K content combined, the reduction in MSE is significantly higher. For *Tier 1 Capital Ratio*, even though the textual features from any version of document content don't improve prediction performance, the best case of MSE is still obtained from 8-K^t + 10-K^t [full] content. The only bank variable for which 10-K^{t+1} [full] content does well compared to other textual content options, even though it doesn't beat the baseline, is for *Leverage ratio*.

In summary, 8-Ks in any configuration, full or half-year worth of updates, don't perform well in prediction accuracy improvement. This should be expected given that 8-Ks are only filings for updates. As seen in this study, When combined with the previous year's status report in 10-K^t, they produce the best textually enhanced prediction for most bank financial indicators. Therefore, 8-K content is informative beyond 10-K content from t and $t+1$.

	filing type	baseline [numeric]	8-K [Q1 + Q2]	8-K [full]
ROA	**8-K^{t}**	1.04E-03	1.21E-03	1.14E-03
	10-K^{t+1} [Item 1]		1.07E-03	
	10-K^{t+1} [Item7]		1.07E-03	
	10-K^{t+1} [full]		*1.02E-03*	
	8-K^{t} + 10-K^{t} [full]		**9.14E-04**	**7.76E-04**
	8-K^{t} + 10-K^{t+1} [full]		1.07E-03	1.04E-03
EPS	**8-K^{t}**	3.01E-03	3.31E-03	3.09E-03
	10-K^{t+1} [Item 1]		3.04E-03	
	10-K^{t+1} [Item 7]		3.03E-03	
	10-K^{t+1} [full]		*2.90E-03*	
	8-K^{t} + 10-K^{t} [full]		**2.49E-03**	**2.24E-03**
	8-K^{t} + 10-K^{t+1} [full]		2.97E-03	2.89E-03
Tobin's Q Ratio	**8-K^{t}**	5.24E-05	8.97E-05	8.39E-05
	10-K^{t+1} [Item 1]		8.18E-05	
	10-K^{t+1} [Item 7]		7.43E-05	
	10-K^{t+1} [full]		*7.32E-05*	
	8-K^{t} + 10-K^{t} [full]		**5.22E-05**	**4.89E-05**
	8-K^{t} + 10-K^{t+1} [full]		7.62E-05	7.24E-05
Tier 1 Capital Ratio	**8-K^{t}**	**3.75E-04**	9.20E-04	8.91E-04
	10-K^{t+1} [Item 1]		6.05E-04	
	10-K^{t+1} [Item 7]		7.57E-04	
	10-K^{t+1} [full]		*7.14E-04*	
	8-K^{t} + 10-K^{t} [full]		**4.69E-04**	**4.32E-04**
	8-K^{t} + 10-K^{t+1} [full]		7.02E-04	6.88E-04
Leverage Ratio	**8-K^{t}**	**1.00E-03**	1.21E-03	1.17E-03
	10-K^{t+1} [Item 1]		1.17E-03	
	10-K^{t+1} [Item 7]		1.15E-03	
	10-K^{t+1} [full]		***1.08E-03***	
	8-K^{t} + 10-K^{t} [full]		1.51E-03	1.42E-03
	8-K^{t} + 10-K^{t+1} [full]		1.16E-03	1.11E-03
Z-score	**8-K^{t}**	6.27E-03	7.51E-03	7.24E-03
	10-K^{t+1} [Item 1]		6.12E-03	
	10-K^{t+1} [Item 7]		6.06E-03	
	10-K^{t+1} [full]		***6.00E-03***	
	8-K^{t} + 10-K^{t} [full]		6.10E-03	**5.96E-03**
	8-K^{t} + 10-K^{t+1} [full]		6.51E-03	6.33E-03

Table 6: MSE values for common 8-Ks and 10-Ks. Values in bold show the least MSE values whereas values in italics show the least values between 8-K at time t and 10-K at time $t+1$ [full].

In summary, we find 8-Ks to be informative, and especially so when combined with 10-K from t. 10-K annual filings are expected to provide a comprehensive and current status of a firm in interest of the investors and shareholders of a firm. Therefore, it is fair to expect that the changes in 10-K filings from year to year should have all the information reported in the year's 8-Ks and should be at least as informative as the year's 8-Ks. The final research question examines this issue of relative informativeness of changes in 10-Ks versus a year's 8-K filings.

	filing type	**8-K [Q1 + Q2]**	**8-K [full]**
ROA	**8-K^t**	**8.09E-04**	**6.89E-04**
	10-K [full] changes	1.03E-03	
EPS	**8-K^t**	**2.45E-03**	**2.14E-03**
	10-K [full] changes	2.95E-03	
Tobin's Q Ratio	**8-K^t**	6.89E-05	**6.44E-05**
	10-K [full] changes	**6.75E-05**	
Tier 1 Capital	**8-K^t**	**5.63E-04**	**5.36E-04**
	10-K [full] changes	6.38E-04	
Leverage Ratio	**8-K^t**	**8.44E-04**	**7.91E-04**
	10-K [full] changes	1.10E-03	
Z-score	**8-K^t**	**5.72E-03**	**5.22E-03**
	10-K [full] changes	6.11E-03	

Table 7: MSE values for 8-Ks vs changes in 10-Ks. Values in bold show the least MSE.

RQ3: Are updates reported in bank's 8-K filings more informative than the changes in 10-Ks in two consecutive years?

For addressing this research question, we define '10-K [full] change' features by taking a difference of the TF-IDF textual features matrices (using all words): 10-K TF-IDF^{t+1} − 10-K TF-IDF^t. We then compare the 10-K change features against the 8-K^t features for the same prediction task of the six bank financial indicators. We just focus on the changes between the TF-IDF scores of the same words appearing in 10-Ks for two consecutive years, and in doing so, retain both positive and negative values of changes. We show the results using 8-K reports from only the first two quarters of a year (denoted 8-K[Q1+Q2] in the second to last column) and using all 8-K reports for the year (denoted 8-K[full] in the last column).

Compared to the results in Table 6, here we report results for changes in 10-K for two consecutive years. This results in some 10-Ks (and corresponding 8-Ks) getting dropped from the analysis due to unavailability of the reports in consecutive years. Table 7 shows that, except for the case of *Tobin's Q Ratio*, we get lower MSE values for 8-K^t features when compared to the 10-K change features. This validates the findings of Table 6, where neither 10-K^{t+1} nor 10-K^t textual features were by themselves able to improve prediction, but when 8-K^t was combined with 10-K^t, it resulted in the best performance for most bank financial indicators. It is therefore fair to conclude that the content of 8-Ks are more indicative of the changes than the changes in the text in 10-Ks, and information contained in 8-K [full] is better than 8-K [Q1 + Q2].

5 Conclusion and Future Work

Natural language processing is increasingly utilized in economics, finance and accounting research to gain valuable insights from text data, where carefully curated domain-specific dictionaries are widely used. An increasing number of financial documents combined with evolution of vocabulary reflecting changes in the relevant domains poses a challenge to hand-curated word lists, such as the L&M's dictionary. In order to address this challenge, we apply a novel large-scale BERT language model based methodology to *automatically* create word list using attention scores. We demonstrated the use of this automatically created word list to define textual features for some important prediction research questions.

Informativeness of regulatory filings, such as 10-Ks and 8-Ks, is important to investors, shareholders and the regulators. Using the automatic word list generator capability, beyond the extensively used L&M sentiment word lists, we studied and compared the informativeness of 8-Ks and 10-Ks in predicting some key financial indicators for banks. Given 8-Ks are update filings and 10-K are annual report filings required of all public companies in the US, we evaluated the relative performance of the two kinds of reports through a series of research questions.

We found that text content of both 8-K and 10-K filings are valuable in predicting some of the performance and risk indicators of the banks. Given banks are highly regulated firms across the globe, being able to improve accuracy of prediction beyond what is supported by numeric variables is very valuable. The L&M word list is outperformed by various word list enhancements considered in this paper, including the automatic attention-score based word list. We additionally find that the optional 8-K filings made by the banks within a year for reporting updates prove to be incrementally valuable beyond all in-

formation extracted from the relevant 10-K documents. And finally, the year-to-year changes in the 10-K annual reports are not as informative as the updates reported in 8-K filings.

This study sets the stage for many threads of future investigation. A brute force extensive word list can be utilized for textual analytics tasks, however as seen in this study, innovative methods for extracting information beyond those considered in this study must be pursued. We remained focused on banking sector and a specific set of financial indicators in this study to address the posed research questions. While at least one variable had a qualitative meaning associated with it, natural language processing can be most valuable for predicting hard to quantify concepts, such as culture, governance, sustainability, climate change risk, and so on. These concepts would also be relevant for firms beyond the banking sector, which should be investigated in future studies.

6 Acknowledgments

This work was supported in part by NSF Award III-1738895.

References

Akhtar, M. S., Kumar, A., Ghosal, D., Ekbal, A., and Bhattacharyya, P. (2017). A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 540–546.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Brown, S. V. and Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year md&a modifications. *Journal of Accounting Research*, 49(2):309–346.

Chang, C. Y., Zhang, Y., Teng, Z., Bozanic, Z., and Ke, B. (2016). Measuring the information content of financial news. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3216–3225.

Chen, Y., Rabbani, R. M., Gupta, A., and Zaki, M. J. (2017). Comparative text analytics via topic modeling in banking. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE.

Das, S. R. et al. (2014). Text and context: Language analytics in finance. *Foundations and Trends® in Finance*, 8(3):145–261.

Dereli, N. and Saraçlar, M. (2019). Convolutional neural networks for financial text regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 331–337.

DeSola, V., Hanna, K., and Nonis, P. (2019). Finbert: pre-trained model on sec filings for financial natural language tasks. *University of California*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duarte, J. J., Montenegro González, S., and Cruz, J. C. (2021). Predicting stock price falls using news data: Evidence from the brazilian market. *Computational Economics*, 57(1):311–340.

Emerson, S., Kennedy, R., O'Shea, L., and O'Brien, J. (2019). Trends and applications of machine learning in quantitative finance. In *8th International Conference on Economics and Finance Research (ICEFR 2019)*.

Gupta, A. and Owusu, A. (2019). Identifying the risk culture of banks using machine learning. *Available at SSRN 3441861*.

Huang, K.-W. (2010). Exploring the information contents of risk factors in sec form 10-k: A multi-label text classification application. *Available at SSRN 1784527*.

Khaidem, L., Saha, S., and Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

Khalil, F. and Pipa, G. (2022). Is deep-learning and natural language processing transcending the financial forecasting? investigation through lens of news analytic process. *Computational Economics*, 60(1):147–171.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (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*, pages 272–280.

Lee, J. and Lee, H. (2008). Predicting corporate 8-k content using machine learning techniques. *Graduate School of Business Stanford University*.

Leidner, J. L. and Schilder, F. (2010). Hunting for the black swan: risk mining from text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 54–59. Association for Computational Linguistics.

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

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

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

Nopp, C. and Hanbury, A. (2015). Detecting risks in the banking system by sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 591–600.

Nousi, P., Tsantekidis, A., Passalis, N., Ntakaris, A., Kanniainen, J., Tefas, A., Gabbouj, M., and Iosifidis, A. (2019). Machine learning for forecasting mid-price movements using limit order book data. *Ieee Access*, 7:64722–64736.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Rawte, V., Gupta, A., and Zaki, M. J. (2018). Analysis of year-over-year changes in risk factors disclosure in 10-k filings. In *Proceedings of the Fourth International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets*, pages 1–4.

Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica: Journal of the econometric society*, pages 431–449.

Sardelich, M. and Manandhar, S. (2018). Multimodal deep learning for short-term stock volatility prediction. *arXiv preprint arXiv:1812.10479*.

Sedinkina, M., Breitkopf, N., and Schütze, H. (2019). Automatic domain adaptation outperforms manual domain adaptation for predicting financial outcomes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 346–359.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Theil, C. K., Štajner, S., and Stuckenschmidt, H. (2018). Word embeddings-based uncertainty detection in financial disclosures. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 32–37.

Theil, C. K., Štajner, S., and Stuckenschmidt, H. (2020). Explaining financial uncertainty through specialized word embeddings. *ACM Transactions on Data Science*, 1(1):1–19.

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

Tsai, M.-F. and Wang, C.-J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250.

Tsai, M.-F., Wang, C.-J., and Chien, P.-C. (2016). Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems (TMIS)*, 7(3):1–17.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A. (2017). Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 1, pages 7–12. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, Y. and Ni, X. S. (2019). A xgboost risk model via feature selection and bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.

Yang, L., Zhang, Z., Xiong, S., Wei, L., Ng, J., Xu, L., and Dong, R. (2018). Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 441–445. IEEE.

Zaki, M. J. and Meira Jr, W. (2020). *Data mining and machine learning: Fundamental concepts and algorithms*. Cambridge University Press.

Zhai, S. S. and Zhang, Z. D. (2019). Forecasting firm material events from 8-k reports. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 22–30.