MAPPING SENTIMENT: A TEXTUAL ANALYSIS ON 10-K DOCUMENTS USING AN ARTIFICIAL NEURAL NETWORK

A THESIS

Presented to

The Faculty of the Department of Economics and Business

The Colorado College

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Arts

By

Ryan Sin

May 2017

MAPPING SENTIMENT: TEXTUAL ANALYSIS ON 10-K DOCUMENTS USING AN ARTIFICIAL NEURAL NETWORK

Ryan Sin

May 2017

Mathematical Economics

Abstract

Using nearly 8000 10-K documents published in 2016 and 2017, we generate contextual vectors through artificial neural networks and test whether the language of 10-K documents, without any detailed numeric indicators of financial performance, correlate with earnings per share and other financials of the S&P 500. We find significant correlation between earnings per share and contextual vectors, concluding that semantic analysis is a valuable tool that has great potential in financial analysis.

KEYWORDS: textual analysis, 10-K, EDGAR, semantic, embedding vector, financial performance, language, context
JEL CODES: C45, G17

ON MY HONOR, I HAVE NEITHER GIVEN NOR RECEIVED
UNAUTHORIZED AID ON THIS THESIS

Signature

TABLE OF CONTENTS

Introduction

Publicly traded companies are required to file a Form 10-K annually, typically containing company history, equity, holdings, risks, etc. (U.S. Securities and Exchange Commission, 2009). This is a way of properly informing current shareholders of a company's financial well-being, ensuring vertical transparency. Investors often use this document to gauge the financial health of companies prior to purchasing stocks, or simply as a tool to contrast and compare different companies. The document obviously contains important figures such as dividends and earnings per share, but much of the information is conveyed through words and sentiment. Our paper aims to extract this information and test whether the language of 10-K documents, without any detailed numerical indicators of financial performance, correlates with the financial performance of publicly traded companies.

The main contribution of our work is the application of algorithmically generated vectors from an artificial neural network that summarizes the context of a 10-k document. Simply put, by associating words with coordinates based on their meaning and usage within 10-k documents, we can generate vectors that summarizes the context of a company. This tool has been used in patent research but has seen limited use in the study of economics (Johnson & Whitehead, 2016). One of the goals of this paper demonstrate the ability of this tool in a financial context such that future researchers may benefit from innovative technology like it.

Literature Review

Textual analysis is an up and coming qualitative tool used to analyze tone and sentiment in financial documents. Researchers have applied this technique to annual reports, press releases, news articles and more. Recent studies have typically relied on negative word classifications provided by the Harvard Psychosociological Dictionary (Engelberg, 2008; Tetlock et al., 2008). Tetlock, Saar-Tsechansky, and Macskassy (2008) conducted a word count of the number of negative words from the Harvard Dictionary in news articles such as the *Wall Street Journal*. They found that the higher proportion of negative words in a news article, the lower the quarterly earnings of the firm. They point out that additional negative words are not redundant information, but contributes to the explanation of the financial performance of the company that is otherwise uncaptured. Loughran & McDonald (2011) further investigated this negative word list and found that nearly three-quarters of the Harvard Dictionary negative word list are not necessarily negative in a financial context. For instance, words such as *liability*, *tax*, *depreciation* appear frequently in 10-K annual reports but each describe an objective element on the balance sheet. Using similar logic, the authors condensed the word list to a more appropriate one for 10-K documents. They also applied term weighting to this list which increases the impact of low frequency words and lessens the impact of high frequency words. They found their word list to be significantly correlated with announcement returns.

The research to date has consistently used the Harvard word list or has created their own word lists through subjective interpretations. However, none have algorithmically generated a word list from the language of the documents. We will be adding to the literature by using an algorithm used by Johnson & Whitehead (2017) to generate vectors from nearly eight thousand 10-K reports. The literature on this semantic approach is largely technical based rather than application based, discussing skip-gram models and automated indices, meaning that the work is in this paper is largely experimental in nature (Mikolov et al., 2013).

Data & Methodology

We collected financial data for the S&P 500 from MergentOnline, a comprehensive global company database. Using the S&P 500 index, we downloaded various financial figures for the companies. Unfortunately, the database was incomplete, which we considered whilst selecting the dependent variable of earnings per share. We chose this index because it is a popular index that encompass publicly traded companies that are leaders in their industry. As such, their annual reports are more likely to be read by investors.

To gather the 10-K documents for processing, we wrote a script to download and parse through nearly eight thousand 10-K documents from U.S. Securities and Exchange Commission's EDGAR database. We did so by removing all numerical values, symbols, punctuation, leaving only the letters of the alphabet. We also made all the letters lower case since the algorithm would recognize capitalized words as separate words from lower cased words. We also replaced many symbols with spaces to treat them as separate but adjacent words. We then consolidated all the documents into one text file, separating each by a line break, and applied the algorithm to the set.

As described by Johnson & Whitehead (2017), the algorithm uses "artificial neural networks to predict individual word context." The algorithm will consider not only the number of times a word appears in the text, but also where they appear in relation to other words. As the algorithm sees more text, the accuracy of its context prediction increases. Once this process is complete, the neural network will generate semantic embedding vectors for every word that appears in the text based on the internal activations of the network. The result are unique vectors associated with every word that appears in the eight thousand 10-K documents. In our case, we found 118,648 unique words in the dataset, but decided to utilize the 50,000 most frequent word vectors that appear in our dictionary since the rest are largely irrelevant. From these semantic embedding vectors, we create document vectors using the 10-K documents of the S&P 500. By adding up the individual word vectors of all the words that appear in a single 10-K document and dividing it by the document word count, we can generate a semantic document vector, each containing an array of twenty elements, that summarizes the context of that single 10-K document specific to a company. These document vectors are the vectors we use in our regression.

Theoretical Model

We compare the document vectors of the 10-Ks of the S&P 500 to their earnings per share as our primary indicator of financial performance. Since each vector consists of twenty elements that are comparable to one another, we can investigate which element has the most indicative power of financial performance. The number of elements, or dimensions, were chosen to strike a balance between accuracy and practicality. On one hand, the higher the number of elements in a vector, the more accurate the vector will be. On the other hand, the process of generating these semantic vectors require immense computational power, so limiting the number of elements at twenty greatly reduced the processing time of the vectors without sacrificing too much accuracy. However, since these elements explain context that range anywhere from positivity to potential dividend announcements, we cannot justify how one element is more indicative than another in a comparative manner. As such, we take advantage of the step-wise regression method to help us decipher which of the elements are relevant in our model. we begin with no variables in the model, adding in each variable if its significance level is smaller than 0.10. The result is a model with the significant elements in the document vectors that correlate with earnings per share. We acknowledge that the step-wise regression method is a form of data mining and should not be taken lightly. It is a regression method that generates a model that fits the data regardless of relevant theory. However, due to the impossibility of interpreting the elements contextually and the experimental nature of this project, we deemed the step-wise regression method a reasonable and ethical tool for our purposes.

Even though this model is the first to apply this semantic estimation technique to financial documents, we still expect the document vectors to correlate with earnings per share to some degree due to the results of prior research based on negative word lists. Below is a summary statistic table describing all the variables in the model.

Table 1: Summary Statistics

Variable	Description	Mean	Standard Deviation	Min	Max
EPS	Earnings Per Share (TTM)	9.01	859.53	-19.92	14645
x1	Element 1	-2302.34	27589.99	-253418	115444
x2	Element 2	-18854.90	19509.96	-104893	69947
x3	Element 3	42810.23	28397.86	-30004	302436
x4	Element 4	3803.59	18557.92	-63540	126220
x5	Element 5	-16813.99	26970.67	-154601	48694
x6	Element 6	48040.86	28786.90	-14406	225644
x7	Element 7	3339.97	16789.66	-73297	70050
x8	Element 8	-11920.98	13756.10	-90797	24702
x9	Element 9	20567.40	37965.75	-43651	369772
x10	Element 10	12920.60	23674.21	-88960	119450
x11	Element 11	-31225.94	28640.56	-186102	23507
x12	Element 12	36755.76	22033.03	-4983	243191
x13	Element 13	39684.28	30363.56	-16204	261059
x14	Element 14	-14005.50	21826.75	-174509	30745
x15	Element 15	2951.98	14467.09	-49801	105306
x16	Element 16	20259.93	22681.87	-73549	248421
x17	Element 17	-8164.45	15700.98	-76351	87053
x18	Element 18	-4329.20	16931.68	-43098	99589
x19	Element 19	19665.75	20064.60	-34128	204190
x20	Element 20	48701.73	41971.75	-26399	446050

The variation of these variables is quite large which is as expected since the number associated with each element are more akin to coordinates than values. However, to properly compare these elements to our dependent variable without reflecting only the variation in our results, we decided to standardize the variables.

Estimations and Results

Our model examines changes in earnings per share when compared to document semantic embedding vectors. We hypothesized that the elements, though not all, will correlate with earnings per share. The exact elements are not important at this stage since they are contextually indifferentiable, but it will point to further applications of this method in the textual analysis. Below is the model step-wise regression selected.

Table 2: Regression Results

Dependent Variable Explanatory Variable	EPS (standardized) Coefficient	T-Stat
s2	-.221	-3.94
s5	.160	2.96
s7	.221	3.78
s10	-.106	-1.88
s16	-.314	-5.08
s17	.216	4.13
s19	.152	2.86
C	.0005	0.01

Due to problems that occurred in the data consolidation process, we were only able to gather 482 observations out of the 500 companies we originally planned to apply this technique on. The coefficients for the elements in our model are obviously all significant at the 10 percent significance level since that is the criteria the step-wise method uses. These explanatory variables are standardized, as indicated by the 's' in front of the element number. Overall, the coefficients of the elements are consistent in magnitude but differ in signage, with three out of the seven elements being negative. To elaborate on the precise meaning of the coefficients of the elements, for each unit s2 gains, earnings per share will decrease by -.221 standardized units. Since the standardized EPS has a mean of -3.94 x 10^{-9}, this is more impactful than it seems.

Through our results, we show that the elements within our document semantic embedding vectors generated from 10-K documents, without any numeric value that could affect our algorithm, correlates with earnings per share of the S&P 500. Alternatively, the words in 10-K documents alone have indicative power on one financial performance tracker within the same cycle. This conclusion is quite limited but does show that there is value in algorithmic textual analysis which is largely unchartered territory in the financial world.

Conclusion

The purpose of this paper is to demonstrate that the words on a financial document aligns with the financial performance of the publishing company. In this study, we found significant correlation between earnings per share and algorithmically generated semantic document vectors, showing that the words in 10-K documents have indicative power on a financial performance tracker within the same cycle. One of the major shortcomings of this model is the limited financial performance indicator we use as our dependent variable. During our research, we did test the model against percentage price change and other financial indicators. However, due to the incompleteness of these other variables in the dataset, we elected to use earnings per share. Even though EPS is commonly regarded as a helpful tracker of financial health, having a wider diversity of financial indicators would provide more supporting evidence for our case.

The textual analysis tool we used is still in its youth and has many potential applications across all industries. If future research apply this technique to other variables and media types, we may eventually be able to generalize the meaning of the elements and predict a company's future performance using a current text.

References

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements.

Johnson, D. K., & Whitehead, M. (2016). The Technological Core of Apple: Using Artificial Neural Networks and Econometrics to Value Apple's Patents. In *R&D Management Conference, From Science to Society: Innovation and Value Creation*.

Johnson, D. K., & Whitehead, M. (2017). A Tool for Visualizing and Exploring Relationships among Cancer-Related Patents.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10- Ks. The Journal of Finance, 66(1), 35-65.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139-1168.

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

U.S. Securities and Exchange Commission. (2009). SEC.gov | Form 10-K. *Sec.gov*. Retrieved from https://www.sec.gov/fast-answers/answers-form10khtm.html