Final Project

Physics 250 Econophysics, Winter 2017

Dustin Burns

Ph.D. Candidate, UC Davis

Title

Using textual analysis of financial documents to gain insight on a company's market performance

Abstract

Although a causal connection between the tone of a company's SEC filing reports (e.g. 10-Ks) and the returns of the company stock has not been established, McDonald et.al. have shown a correlation between tone and other performance indicators such as filing-period returns, return volatility, trading volume, and accounting fraud [1]. Care has to be taken in defining the dictionary used to measure the tone, as many words that typically have a negative connotation in a general context do not in a financial context, such as *liability* and *foreign.* The goal of this study is to reproduce the correlations found by McDonald et.al. between their custom dictionaries and performance indicators, primarily, the negative correlation between filing period excess return and weighted negative word counts. Additionally, a simplified version of the regression procedure is performed to measure how well the variations in filing period returns can be explained by the proportion of negative words in a document.

Project Background

Tyler Hayes, VP Blackrock, advised me on the subject and technical approach for this project, based on my desire to gain exposure to techniques and software used in industry to address similar language processing problems. The general approach of using nonstandard and unstructured sources of data to gain market insight is gaining traction, and is part of the larger "Big Data" movement in industry. Tyler was confident that this project, combined with my previous background in data analytics, would provide strong evidence to data science employers of my competencies. He also provided me with learning resources, tutorial links and online coursework.

Introduction

Most companies with greater than $10 million in assets are required to file summary documents, collectively called 10-X documents, with the Security and Exchange Commision (SEC) on a quarterly and annual basis [2]. Investors use these documents to gauge the financial performance of the company directly, by observing changes in executive compensation, equity, audits, and other numerical data, and indirectly, by gauging the overall tone of the text in the document. The indirect measures are more difficult to automate, but techniques have been developed to computationally measure the tone in financial text, falling into the general field of natural language processing [3-5].

There are several approaches to quantifying the tone of a document. The simplest approach is called the "bag-of-words" algorithm, whereby a document is tokenized into individual words, then the number of negative words matching those present in a tone dictionary are counted. The major drawback of this algorithm is the loss of information about the presence of words relative to one another.

Care must to be taken in defining the dictionary used to measure tone, as many words that typically have a negative connotation in a general context do not in a financial context, such as *liability* and *foreign.* Additionally, other words that are typically found in negative tone dictionaries can simply describe the company's industry, such as *cancer* and *crude* (oil). McDonald et.al. have developed several word lists that are applicable to financial documents, including the Fin-Neg list used in this study to measure negative tone [1].

Before summing the word counts in each document, the word counts can be weighted in two different ways to account for different effects. The first, called proportional weighting, divides the total number of negative words t_{ij} by the total word count in the document n_j. The negative word yield of the j^{th} document is then the sum of word weights given by

$$Y_j = \sum_i w_{ij} = \sum_i \frac{t_{ij}}{n_j} \qquad (1)$$

The second weighting procedure, called tf.idf (term frequency, inverse document frequency) weights word counts accounting for the frequency of the word relative to the average frequency of other words in a document and the commonality of the word in the set of documents. The weights, which are summed for each document as before, are given by

$$w_{ij} = \frac{1+log(t_{ij})}{1+log(a_j)} log(\frac{N}{d_i}) \qquad (2)$$

Where a_j is the average word count in the j^{th} document, N is the total number of documents, and d_i is the number of documents with at least one occurrence of the i^{th} word.

Once the negative tone of a company's 10-X document is measured, correlations can be measured between the tone and financial indicators, such as stock returns, volume, volatility, and the likelihood of fraud. This study focuses on the correlation to stock return, being potentially the most powerful indicator. In particular, the market response to a 10-X filing can extend up to three days after the filing day [6]. The excess return is defined as the buy-and-hold stock return over the three day filing period, relative to the overall market return, as indicated by the S&P 500 index.

The remainder of this report is organized as follows: the technical approach to scraping and cleaning the required data is outlined, followed by the results of the correlation study between tone and filing period excess returns, ending with a brief discussion and conclusion.

Technical Approach

The data used in this study is scraped from the SEC EDGAR database [7], following the directory convention described in [8], using about 5% of the metadata given in [9]. This metadata describes the data set used in [1], and consists of 10-X documents satisfying basic criteria such as number of words greater than 2000 and the availability of external correlation variables.

Once the document files are downloaded, the text is cleaned, roughly following the procedure outlined in [10]. The HTML blocks corresponding to graphical objects and accounting data (segment <TYPES> of GRAPHIC, ZIP, EXCEL, PDF, XBRL) are removed. <TABLE> segments are not removed do to the possibility of their use in formatting text segments. Header and footer blocks are removed. Next, the document is tokenized into words. Single character words and words containing numbers or special characters (regex [\d\/*\'\-,=;:@<>\._]) are removed, then the list is alphabetized to optimize matching to the negative word dictionary. The total word counts and number of unique words obtained with this procedure are within about 20% of those given in [9], always larger. The total negative word counts are within about 5%.

The external correlation variable used in this study is the filing period excess return, given by the buy-and-hold stock return minus the buy-and-hold market value return as indicated by the S&P 500, over the three day filing period following the filing date. Historical stock data is pulled from Yahoo Finance using the yahoo-finance python plugin [11]. The Yahoo database is queried with the stock ticker symbol, but the metadata file only contains the SEC CIK identification number for each company. A translation from CIK number to ticker symbol is obtained from the document [12]. Files are removed from the data set if their historical stock price data is not available from yahoo-finance.

The Fin-Neg word dictionary used to measure negative tone is obtained from document [13]. A search algorithm is implemented where the 10-X word tokens are sorted alphabetically in a python dictionary. The unique token keys are matched to words in the alphabetized Fin-Neg list, incrementing a counter with the token value. From these counters, the quantities needed to form the weights given in Equations (1)-(2) are calculated.

The final step is determining the linear correlation between the dependant variable filing period excess return and the negative word count yields. A linear regression is performed, minimizing the cost function in Equation (3) to determine the best fit line parameters.

$$C(m,\ b) = \sum_{i} (y_i - \widehat{y_i})^2 \ , \tag{3}$$

where $\widehat{y_i} = mx_i + b$ is the best fit line. The standard error, s, defined in Equation (4) is determined for each fit. The standard error measures the accuracy of the prediction, giving the average distance of a data point from the regression line. The correlation coefficient (coefficient of determination, or r-squared), given in Equation (5), is determined for each fit. The correlation coefficient measures how well the variance in data points is predicted by the independent variables.

$$s^2 = \frac{1}{N} \sum_i (y_i - \widehat{y_i})^2 \tag{4}$$

$$r^2 = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \overline{y_i})^2} \tag{5}$$

The final statistical measure uses a t-test hypothesis test to determine how well the data is modelled by a linear relation. The null hypothesis consists of a linear relation with slope zero, describing data that does not have a strong linear correlation. The alternative hypothesis describes data with a linear correlation with some non-zero slope, m. The null hypothesis is rejected if the p-value, the probability that the t-distribution is greater than the calculated test statistic, is consistent with the desired confidence level, typically 90%.

Results

The binned and unbinned linear regressions for the proportional weighting procedure are shown in Figures 1 and 2, respectively. The best-fit line slope, standard error, r-squared, and p-value are given in Table 1.

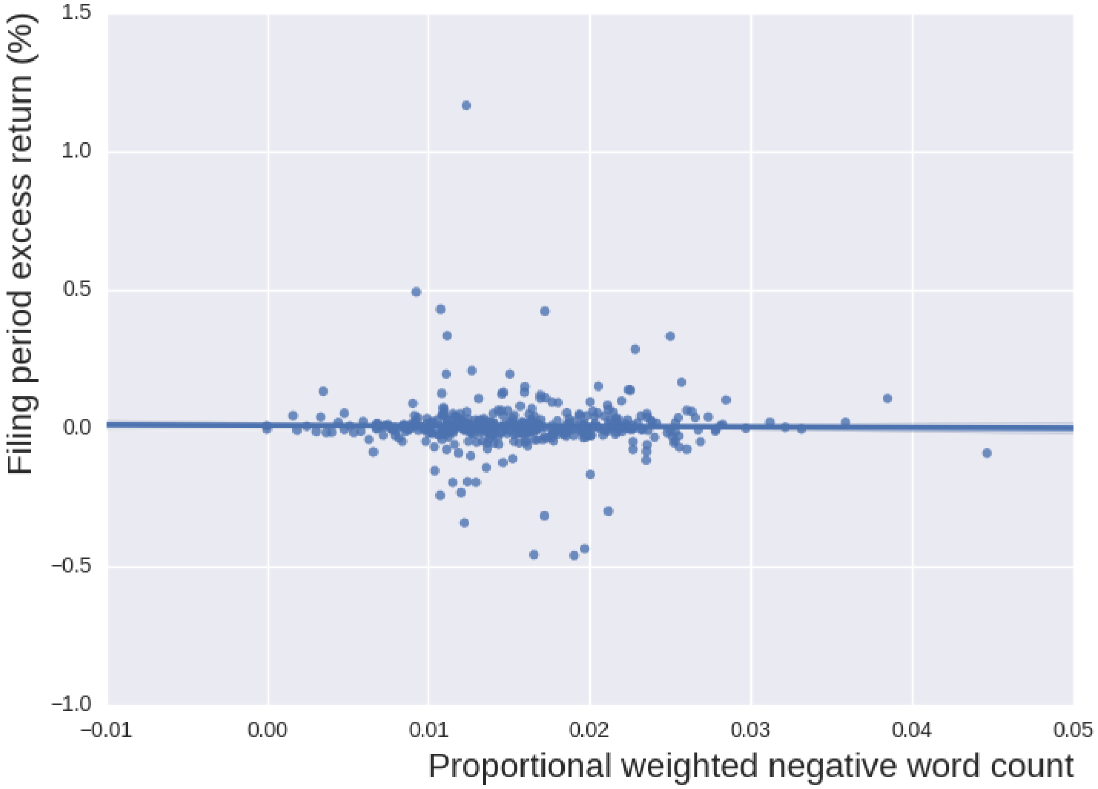

Figure 1: Unbinned proportional weighting correlation.

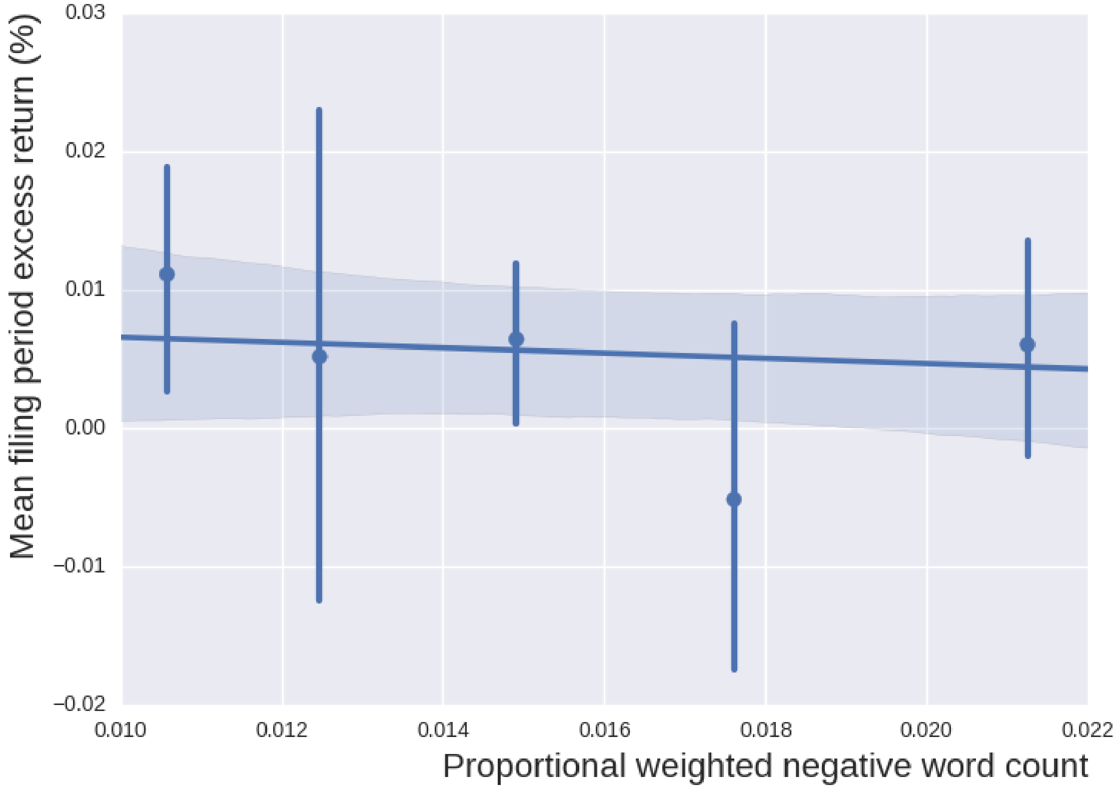

Figure 1: Binned proportional weighting correlation.

Method	**Slope**	**Std error**	**R-squared**	**p-value**
Unweighted	2.8E-6	1.0E-5	1.7E-4	0.78
Proportional weighting	-0.19	0.78	1.4E-4	0.81
Tf.idf weighting	3.5E-4	6.6E-4	6.2E-4	0.60

Table 1: Statistical measures for different weighting strategies

Discussion

The goal of this study was to measure the correlation between filing period excess returns and negative word counts in 10-X documents using the algorithm described in [1]. This correlation is measured in several ways. First, it is observed in Figure 2 that companies with a higher proportion of negative words, on average, have a more negative excess return. This result can be compared to Figure 1 of [1], which indicates a stronger negative correlation, although it is unclear how the bins are formed and what is the uncertainty per bin.

Next, a simplified version of the regression analysis performed in [1] in done using linear regression with one independent variable instead of a multivariate approach. The statistical measures of the regressions for three different weighting methods are shown in Table 1. The proportional weighting method has a best-fit line with negative slope, indicating a negative correlation, while the other methods have a slope nearly zero. The standard errors tend to be large in the units of the dependent variable, supported by the large error bars in Figure 2. This error could be decreased by adding more data to the study. Only about 5% of the total data set was analyzed in this study.

The small r-squared values indicate that the variation in the dependent variable cannot be explained well by the independent variable. This is consistent with the results in section IV.C of [1], namely, “Only a small amount of the variation in filing period returns is explained by the independent variables. Textual analysis is not the ultimate key to the returns cipher.” Finally, the p-values are not low enough to confidently reject the null hypothesis that the data is uncorrelated. In other words, the data does not show a strong linear correlation, although a weak negative correlation is indicated. This is likely due to the lack of events in the high negative word count tail, and the noise and outliers present at low negative word count values.

Conclusion

This study was successful in reproducing the main result of [1], a negative correlation between filing period excess returns and negative tone in SEC 10-X documents. Several shortcomings of the methods used were revealed. Using the “bag-of-words” algorithm looses relational word information. Using a more sophisticated word vectorization method, such as the TensorFlow word2vec method, could lead to stronger results. The use of word lists to gain insight in financial markets has a clear drawback of the existence of feedback loops; if the document preparer is aware that investors will respond in a negative way to certain language, they can simply modify the language to a weaker or alternate form. As pointed out in [1], it is unclear if the results we've obtained imply a causal link between tone and returns, or if tone simply proxies for accounting data in 10-X documents that has a stronger link to returns. Further study is needed to better understand this relationship.

References

[1] Tim Loughran and Bill McDonald, 2011, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," Journal of Finance, 66:1, 35-65.

[2] https://www.sec.gov/files/2017-03/form10-k.pdf

[3] Antweiler, Werner, and Murray Z. Frank, 2004, Is all that talk just noise? The information content of Internet stock message boards, Journal of Finance 59, 1259–1293.

[4] Das, Sanjiv, and Mike Chen, 2001, Yahoo! for Amazon: Opinion extraction from small talk on the web, Working paper, Santa Clara University.

[5] Demers, Elizabeth, and Clara Vega, 2008, Soft information in earnings announcements: News or noise? Working paper, INSEAD.

[6] Griffin, Paul, 2003, Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings, Review of Accounting Studies 8, 433–460.

[7] https://www.sec.gov/edgar/searchedgar/companysearch.html

[8] https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm

[9] http://www3.nd.edu/~mcdonald/Data/LoughranMcDonald_10X_2014.xlsx

[10] http://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/Documentation_StageOne_10-X_Parse.pdf

[11] https://github.com/lukaszbanasiak/yahoo-finance

[12] http://rankandfiled.com/#/data/tickers

[13] http://www3.nd.edu/~mcdonald/Word_Lists_files/LoughranMcDonald_MasterDictionary_2014.xlsx