# “Textual Analysis in Finance”

Tim Loughran
Bill McDonald

University of Notre Dame

# Outline

- 1. Sentiment Analysis
- 2. Zipf's Law
- 3. Financial Document Readability

# Problems at home with a bunch of teenagers

- Why is it that I learn nothing after a 20 minute conversation with my teenager kids?
- The language of teenagers has a high frequency of weak modal words (i.e., weasel words).
- Weak modal words include *may*, *might*, *could*, *depending*, *possibly*, and *appears*.

- What if management used language like a teenager in a conference call, IPO prospectus, or annual report?
- That is, how would investors respond to the use of a large frequency of weasel words by management?
- Academic evidence shows that firms with a high fraction of weak modal words in their annual reports or in an IPO prospectus have higher subsequent stock return volatility.

# Bag of words model

- Bag of words model simply tabulates the frequency of each word within a document.
- Importantly, the word sequence is completely ignored. That is, I do not care about whether the word is used as an adjective or a noun in the sentence.
- Although simplistic, the bag of words approach is quite powerful.

# Tone in Documents

- To measure tone, typically the proportional count of negative words is used.
- More negative words scaled by total document words is gauged as being more pessimistic.

# Capturing Negative Sentiment

- Why not positive words? The framing of negative information is so frequently padded with positive words that the measured positive sentiment is often ambiguous.
- How would you measure the sentiment of a financial document or newspaper column discussing financial news?
- You could survey colleagues about what words they think are negative.
- You could read a few annual reports or newspaper articles and create a list of negative words.

# Available off-the-shelf word lists

- The most common off-the-shelf dictionary is the Harvard General Inquirer (GI).
- However, since the Harvard GI word lists were NOT created with financial documents in mind, serious misclassifications could occur.
- Do you think a word list developed for psychology and sociology would translate well into the realm of business?

# Sentiment misclassifications by the Harvard GI negative word list

- Bill and I find that almost 75% of the Harvard GI negative words do not have pessimistic meaning when used in the context of financial documents.

- Harvard GI negative words like *tax*, *cost*, *capital*, *board*, *liability*, *vice*, *foreign*, and *depreciation*, which are predominate in firms' 10-K filings, do not typically have negative meaning when appearing in an annual report.

- Clearly, these are not negative financial words. The firm is merely naming their *board* of directors or company *vice*-presidents.

- We also document that several of the Harvard negative words are likely to proxy for specific industries.

- For example, management's use of *crude*, *cancer*, and *mine* do not have negative meaning and merely proxy for the oil, pharmaceutical, and mining industries.

# Loughran and McDonald word lists (2011, Journal of Finance)

- We created six different word lists (negative, positive, uncertainty, litigious, strong modal, and weak modal) by examining word usage in at least 5% of 10-Ks (i.e., annual reports) during 1994-2008.
- Our approach was to create a relatively exhaustive list of words that makes avoidance much more challenging.
- The sentiment lists are based on the most likely interpretation of a word in a business context.
- The Loughran and McDonald (LM) word lists are quite extensive: our dictionary contains 354 positive and 2,329 negative words.

# LM (2011) Negative words

- Here are the most frequently occurring LM negative words in a large sample of annual reports: *loss*, *losses*, *claims*, *impairment*, *against*, *adverse*, *restated*, *adversely*, *restructuring*, and *litigation*.

- These 10 words are only 0.4% of the universe of 2,329 LM negative words, yet these 10 words account for more than 33% of occurring negative words in 10-Ks.

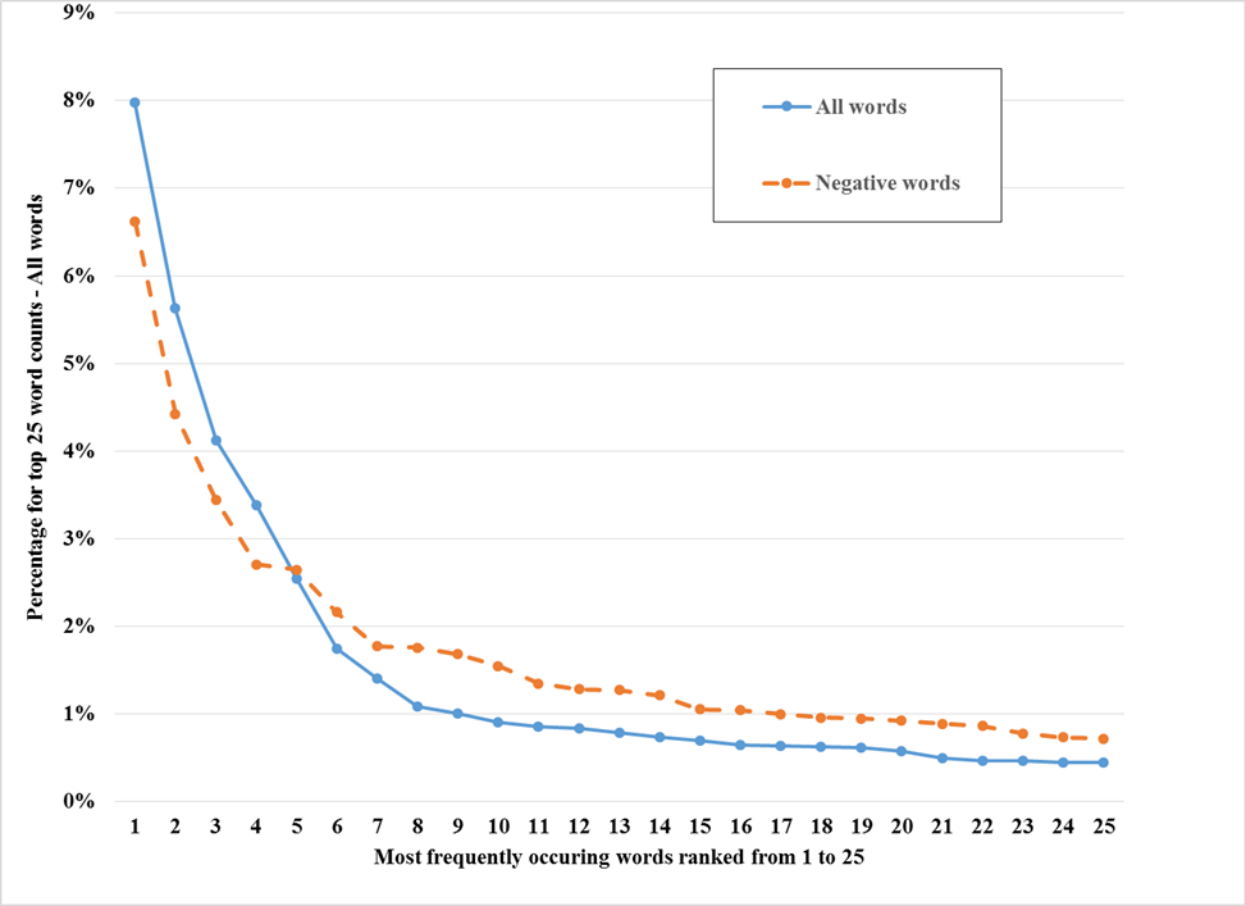
- How is that possible?

# Zipf's Law

- The driving force behind the tripwires of word classifications is that word counts tend to follow a power law distribution, a phenomenon frequently referred to as Zipf's Law.

- That is, a very small number of words will dominate the frequency counts for a given set of words. If one of these words is misclassified it potentially will drive the results.

**The figure shows the proportions for the top 25 most frequently occurring words in all 10-K/Q type SEC filings over the period 1994-2012 for both "All words" and for "Negative words".**



# Measuring the Readability of Financial Documents

- How would you gauge the readability of a document? What makes some text easier to understand than other documents?

- Both financial researchers and government regulators have struggled with the notion of how to define and measure the readability of mandated financial disclosures.

# What is the purpose of financial document disclosures?

- In our 2014 Journal of Finance paper, Bill and I propose defining readability as the effective communication of valuation-relevant information.

- Thus, more readability financial documents should be significantly associated with lower return volatility, earnings forecast errors, and earnings forecast dispersion, after controlling for other variables.

# Fog Index

- In the accounting and finance literature, researchers often use the Fog Index as a measure of document readability.

- The Fog Index's popularity is primarily attributable to its ease of calculation and adaptability to computational measure.

# Fog Index Definition

- The Fog Index is a simple function of two variables: average sentence length (in words) and complex words, defined as the percentage of words with more than two syllables.

- As is common with many readability measures, the two factors are combined in a manner that is intended to predict grade level:

Fog index = 0.4 (average number of words per sentence + percent of complex words)

- Our paper shows that the second component in the Fog Index, “complex words,” is a poorly specified measure in business documents.

- The Fog Index indicates that an increase in the number of complex words (more than two syllables) decreases readability, with this factor accounting for half of the measure’s inputs.

- Business text, however, commonly contains multisyllable words used to describe operations.

- Words like *financial*, *company*, *interest*, *agreement*, *including*, *operations*, *period*, and *related* are predominant complex words occurring in 10-Ks, yet are presumably easy for investors to comprehend.

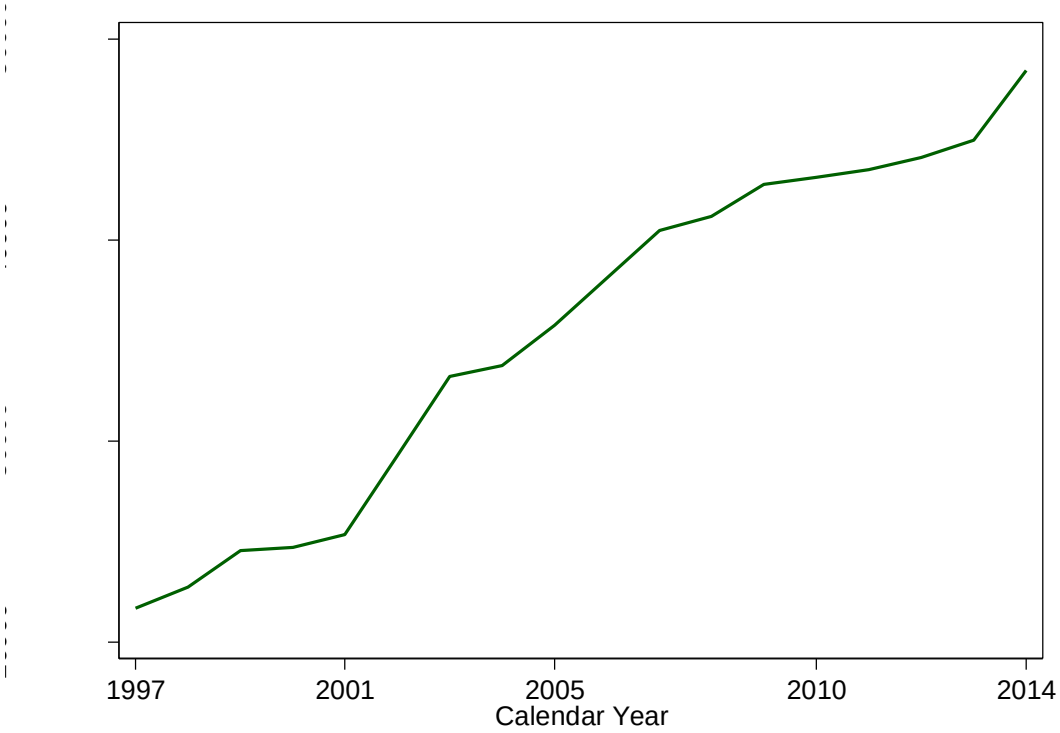
- We propose using the document file size (i.e., the number of megabytes required to store the document as reported on the SEC website) as a simple, and admittedly imperfect, proxy for readability.

- Annual report file size (log) is strongly correlated with number of words in the document. Whether "size" is gross or net does not seem to matter.

- We show that file size relates to post-filing return volatility and other measures of the information environment in a manner consistent with the notion of readability.

- Do you think that annual reports are becoming less readable?

# Trend in average number of 10-K words for publicly-traded firms, 1997-2014



# Conclusion

- Word lists designed specifically for business communication should be used to measure the sentiment of business text.

- Zipf's law—which documents the fact that a very small number of words will dominate the frequency counts—creates a research environment where methodological errors can have a huge effect.

- By its very nature, business text has an extremely high percentage of complex words—one of the Fog Index's two components—that are well understood by investors and analysts.

- When we use the term "readability" in the context of financial documents, think carefully about what that means.

- The file size of the annual report is an easily calculated proxy for document readability.

- Thanks.