1. **Overview of the Study**
   1. Harvard Dictionary location – 3/4s of words there are NOT negative in 10Ks
   2. Why is this the case? Financial statements have a language of their own? How & why has this developed over time? How can you improve the accuracy of such a Dictionary?
   3. Key Financial Variables to examine:
      1. 10-K filing returns,
      2. trading volume,
      3. return volatility,
      4. fraud,
      5. material weakness references and
      6. unexpected earnings.
   4. Other places to evaluate for sentiment include newspapers articles, investor message boards such as WallStreetBets and Reddit.
   5. **Key Result = Corr(Sentiment Measures, Financial Variables) >>> 0**
2. **Harvard Psychological Dictionary** – Word categorization scheme for psychology and not finance.

[General Inquirer Categories (harvard.edu)](https://inquirer.sites.fas.harvard.edu/homecat.htm)

* “Content analysis stands or falls by its categories. Particular studies have been productive to the extent that the categories were clearly formulated and well adapted to the problem.”
* “Misclassified words that are not likely correlated with the variables under consideration—for example taxes or liabilities—simply add noise to the measurement of tone and thus *attenuate the estimates regression coefficients*.”

***Measurement Error = Causes Coefficient Estimates to be Biased Towards Zero***

* Additionally, High Frequency Misclassifications can cause Type 1 errors or FALSE POSITIVES. “mine” & “cancer” relate to industry segments such as mining and drug research/healthcare rather than negative connotations.
* *Polysemes = words that have multiple meanings also complicate our analysis.*
* **Key Result = Fin-Neg Lists >>> Harvard Negative Word List.**

1. ***Dataset construction***, Notre Dame weblink & the CRSP Database => Table 1

* Edgar website, WRDS databases, CIK to PERMNO matching – see WRDS website
* Standard data controls such as Price > $3/share to eliminate illiquid stocks with large bid-ask spreads, no recent IPOs all firms trading > 60 days prior to filing of the 10K reports.

1. ***Term Weighting*** – can have “an enormous impact on the effectiveness of a retrieval system.”

* **Term weighting acknowledges that raw word counts are not the best measure of a word’s information content.**
* Address Three Dimensions with Term Weighting:

1. Importance of a term within a document (often measured by proportional occurrence or the log of frequency)
2. Some form of normalization for document length
3. **The importance of a term within the entire corpus (typically measured by inverse document frequency)**

* tf.idf = Term frequency (accounts for normalization/weighting) vs. Inverse Document Frequency (adjustment for the impact across the entire collection).

“df(i)” = number of documents containing at least one occurrence of the ith word.

“tfi,j” = the raw count of the ith word in the jth document

“a(j)” = average word count in the document

N = total number of 10-Ks in the sample

* Positive words have relatively little to no impact on upside.
* Negative words have impact to the downside.
* Negative words with high term-weighting have significant impacts to the downside.

1. Logistics for the analysis  
   Obtaining the necessary returns data from WRDS/CRSP.

[Historical S&P 500 Index Constituents (upenn.edu)](https://wrds-www.wharton.upenn.edu/pages/wrds-research/applications/python-replications/historical-sp-500-index-constituents/)

[General Inquirer Categories (harvard.edu)](https://inquirer.sites.fas.harvard.edu/homecat.htm)