Analyzing 10Ks using NLP

Hypothesis: 10K's are long and boring to read. Research (and practical experience) suggests that stock price is correlated with the year over year change in 10K language and with negative sentiment increase in the 10K. Fortunately, Natural Language Processing (NLP) is a machine-learning technique that analyzes bodies of texts and their positive and negative sentiments. The aim is to use NLP to analyze 10K's and find possible arbitrage opportunities in the market. Below are some sources to that end.

SOURCES

1. Initial Source (https://www.nytimes.com/2019/01/10/business/secrets-corporate-reports-apple.html)
   New York Times article. When there are a lot of changes in the 10K year over year, something important (and often negative) is going on.

2. Lazy Prices (https://www.nber.org/papers/w25084)
   Paper from National Bureau of Economics cited by NYTimes in #1

3. Microsoft Research (https://www.microsoft.com/developerblog/2017/12/04/predicting-stock-performance-deep-learning/)
   Paper outlining a deep learning CNN framework for analyzing 10Ks (using keras)

4. Stanford University research (https://nlp.stanford.edu/pubs/lrec2014-stock.pdf)
   Text analysis for 8K documents predicting stock prices.

5. Filingsummary.com (https://filingsummary.com/)
   Website that looks similar to what we're trying to do but I think without the NLP anaysis. It's a glorified text searcher.

6. y-combinator note (https://news.ycombinator.com/item?id=17193624)
   Notes that corporations are well aware of NLP and train corporate officers to avoid key phrases that might give too much away.

Overall trend: people know about this technique, but no one has done a large study with lots of year over year data.