Analysing companies 10-Ks using NLP to predict stock price movement

By: Amruthraghav Gopalakrishnan(3035436674)

Introduction:

Publicly traded companies in the United States are required by law to file reports with the Securities and Exchange Commission (SEC) on "10-K" and "10-Q." These reports include qualitative as well as quantitative explanations of the success of the business, from sales estimates to qualitative risk factors. "Companies are required to disclose" important pending litigation or other legal proceedings "details. As such, 10-Ks and 10-Qs also provide useful insights into the success of a company. As such, 10-Ks and 10-Qs often hold valuable insights into a company's performance.

However, these observations can be hard to obtain. The average 10-K is about 42,000 words long. Beyond the sheer length, for many investors, dense terminology and tons of boilerplate will further obscure real sense. In order to extract meaning from the data they contain, we do not need to read the 10-Ks of cover-to-cover businesses.

Hypothesis:

When major changes happen to their business, companies make major textual modifications to their 10-Ks. We consequently consider textual changes to 10-Ks to be a signal of the movement of future share prices. Since the vast majority (86 percent) of textual changes have negative sentiment, we usually expect a decrease in stock price to be signaled by significant textual changes (Cohen et al . 2018).

Methodology:

1. Scrape the 10-K documents from the SEC EDGAR Database for a set of publicly traded firms. Upon scraping, we perform some basic data-cleaning and pre-processing for the 10-K document (removing HTML Tags and numerical tables, converting to txt files, lemmatisation, stemming, removing stop words, and so on).
2. After pre-processing we analyse the textual data from one of the 10-K documents using Exploratory Data Analysis(EDA) techniques such as Bag of Words(BoW), TF_IDF, Wordcloud, LDA modeling with interactive pyLDAvis, Top 20 most frequently used words, and Positivity score of the 10-K document using Textblob library.
3. For a particular company, we calculate the cosine similarity and the Jaccard similarity over the set of 10-Ks that were scraped(cosine and Jaccard similarity are relatively less computationally intensive to compute). We calculate the similarity by comparing each 10-K document with the previous year's 10-K and given it a score. We then calculate the difference between these similarity scores and compare this over the years.
4. Try to map these text changes with the stock price movement for the firm. We download the prices of the stock from the first-time the 10-K document was available till the day the lasted 10-K document was filed. Upon carefully analysing, we can conclude as to whether the results align with the hypothesis or not.

Working:

For the sole purpose of this project, I had strictly restricted my study to just one stock and I had decided to go with “AMZN”(Amazon Inc). After deciding the stock, I was able to complete all the steps in the methodology. The scraper is built to collect multiple companies 10-k documents and also strictly adhere to some requirements by the EDGAR regarding the requests made per second. Upon successful scraping, data cleaning and pre-processing were performed on the HTML pages that were downloaded. Multiple modifications were done to the code to ensure the out of the pre-processing stage was usable for the EDA. Following this, EDA was performed to develop some insights into the textual data, and finally, the similarity scores were calculated. Lastly, I downloaded the AMZN stock price and compared it to the similarity scores to conclude.

Conclusion:

We can observe that the stock price of Amazon has steadily increased over 20 years. From the similarities score that was generated, we know that the changes in the cosine similarities and Jaccard similarities between the years is very low. Drawing this back to the hypothesis we started with, where we suggested 'Major text changes in 10-K over time indicate significant decreases in future returns.' can be proven true for this case using the help of Logic Resolution. Amazon had little text changes in its 10-K documents over time and because of that, we expected an increase in the stock price(vice versa of the hypothesis). To strongly verify this, we need to repeat the same for multiple companies. This hypothesis can be further extended into 10-Qs(10-Ks quarterly) and NLP can be applied to earning call transcripts and quarterly earnings reports to further get a better picture of a company’s future returns.

References:

Cohen, Lauren and Malloy, Christopher J. and Nguyen, Quoc, Lazy Prices (March 7, 2019). 2019 Academic Research Colloquium for Financial Planning and Related Disciplines, Available at SSRN: https://ssrn.com/abstract=1658471 or http://dx.doi.org/10.2139/ssrn.1658471