Looking Beyond Numbers

Li Ma, CFA, Director Quantitative Research

March 2020

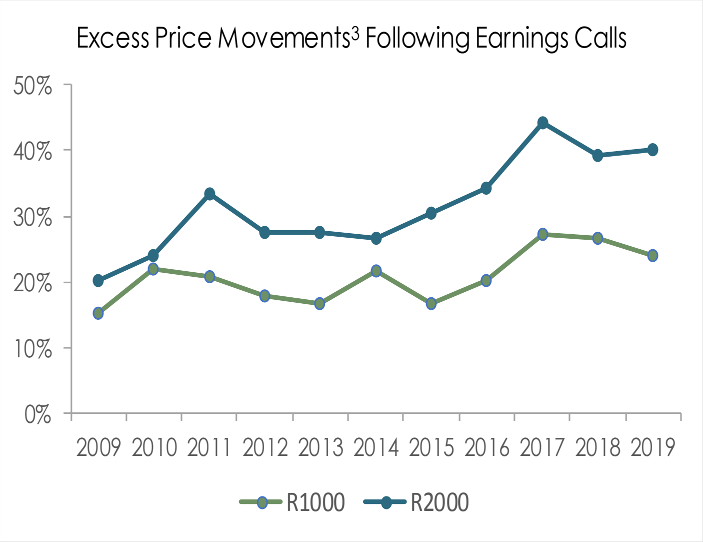
Natural Language Processing (NLP) has become an increasingly popular topic among investors in recent years. The term refers to the process of using computers to analyze human (natural) language and to extract information from large amounts of textual data. As a subfield of linguistics and artificial intelligence, why is NLP attracting the attention of investors? How can NLP be applied in the investment process?

WHY USE NLP IN INVESTING?

In the investing arena, information advantage is the key to success. A huge amount of information is stored as text in today's world. Every quarter, corporate executives file lengthy reports with the U.S Securities and Exchange Commission (SEC), detailing the company's operations, financial performance, potential risks as well as the management's outlook for the firm. While numeric data in the financial reports are easily extracted and readily available from various data vendors, a lot of valuable, forward-looking information is embedded in the vast amount of text in these filings. In fact, in a typical quarterly report, even within the number-heavy Financial Statements section, only 11% of the information is in numeric form while the rest is textual information, in the form of footnotes to the financial statements. The remainder of the report often includes many pages of management discussion of company performance and disclosures about risks. NLP algorithms can help investors effectively process large amounts of textual data and glean useful information, especially when designed by experts with domain knowledge in finance and investing.

In addition to being abundant, textual information is very relevant for stock returns. Investors and analysts make great efforts to gain insights into how a company will perform by reading news, participating in earnings calls, attending shareholder meetings, etc. These events move stock prices. For example, a typical large-cap company holds four earnings calls a year. Using the Russell 1000 Index as a common large-cap equity universe, the price movements[1] during the week following these calls account for 9.4% of the price movements over all trading days of a year. This means such events are 21%[2] more impactful on large-cap stock returns than if information were dispersed uniformly throughout the year. The excess price movements are even more pronounced among small-cap stocks (Russell 2000 Index), and have grown in recent years (Figure 1).

Figure 1 Price Movements[3] Following Earnings Calls



To gain insights from these sources historically required a dedicated team of analysts to collect and extract information from large amounts of written or spoken human language. However, investors have not been able to leverage information in textual data on a large scale due to limited attention and capacity[4]. Relying on human analysis of large amounts of text, results in either huge workloads or coverage of only a fraction of available data. Recent advances in computer science, within both software (proliferation in NLP tools) and hardware (storage and speed), open the door for investors to incorporate this rich dataset into their investment process. Having the ability to access and quantify such meaningful textual data, in addition to the assessment of numerical data, creates an advantage in evaluating the future potential of a company. An additional bonus is less crowded trades for early adopters before other investors begin to hear and learn about the NLP technique, and hence more opportunities of alpha for the first movers.

HOW TO APPLY NLP - CHALLENGES AND METHODOLOGIES

Contrary to numeric data, which is typically well defined, cleaned and standardized by data vendors, textual data is often messy, does not conform to a set format and is sometimes ambiguous. The "unstructured" nature of textual data, together with its vast size, present many challenges for investors who want to leverage its information content. Today's computing power and innovation in parallel processing have largely removed the obstacles to the transfer of and calculation involving large volume of data[5]. Yet a lot of effort is still needed to convert human language to a format which is machine-readable. This includes (1) clean-up of the textual data and extraction of only relevant information, and (2) the transformation of text into quantifiable factors to be consumed in analytical models.

Cleaning the text and extracting relevant information

A few challenges at this stage include:

Retrieving and parsing

The first step of dealing with natural language data is often to extract relevant text. For example, corporate filings are stored on the SEC's EDGAR database, usually in HTML format, and need to be appropriately parsed to retrieve sections of interest. Earnings calls need to be transcribed for the spoken language to be used in subsequent textual analysis.

Removing stop words

"Stop words" are common words in a language, such as "the", "which" and "of" in English. They are so pervasive that they typically do not carry useful information. Removing these words reduces the noise in the textual data.

Stemming and lemmatization

English words have different inflections. For example, "runs" and "ran" have the same root "run", and they tend to convey the same information. Stemming and lemmatization are two ways of returning a word to its root, so that words with the same meaning can be grouped together.

Ensuring consistency

Different people talk and write in a wide variety of styles. Some may use abbreviations while others use formal spellings. Some numbers may be expressed as words (e.g. "a quarter" as opposed to the numerical equivalent "25%"). Some text may even have misspellings. Therefore, it is important to account for these inconsistencies when analyzing a large collection of text.

As the data cleaning efforts directly affect the quality of inputs to subsequent data analysis steps, the designer of the NLP algorithm has to be equipped with not only sound programming skills, but also solid financial knowledge to filter out as much “noise” as possible while retaining relevant facts and insights. Once all text is extracted and cleaned, there are several ways to go about analyzing the data depending upon the goal.

Translating text into quantifiable factors

Fundamental analysts may be interested in quickly identifying the “topic(s)” of a given document, in which case the topic modelling technique, a text-mining tool for discovering abstract topics in an extensive text body, can be applied.

Quantitative investors typically focus on finding the relationship between certain characteristics of the text and future stock returns. For example, the way in which executives deliver messages on earnings calls, such as being straight forward vs. elusive, can signal how confident they are of the company’s outlook. Thus the “readability” of a CEO’s speech may provide additional insight into stock performance not presented in the reported earnings numbers. Another commonly used tool is sentiment analysis. With the help of NLP techniques, one can quantify the magnitude of positivity or negativity in a financial report or executive presentation. Sentiment is presumed to signal the future direction of stock movements. These descriptive characteristics can then be transformed into quantitative scores and analyzed the same way as numeric alpha factors.

The incorporation of sentiment factors derived from textual data is included into our investment process at Chicago Equity Partners. The following case study demonstrates the usefulness of sentiment factors on a stand-alone basis and how they have complemented traditional alpha factors in recent years.

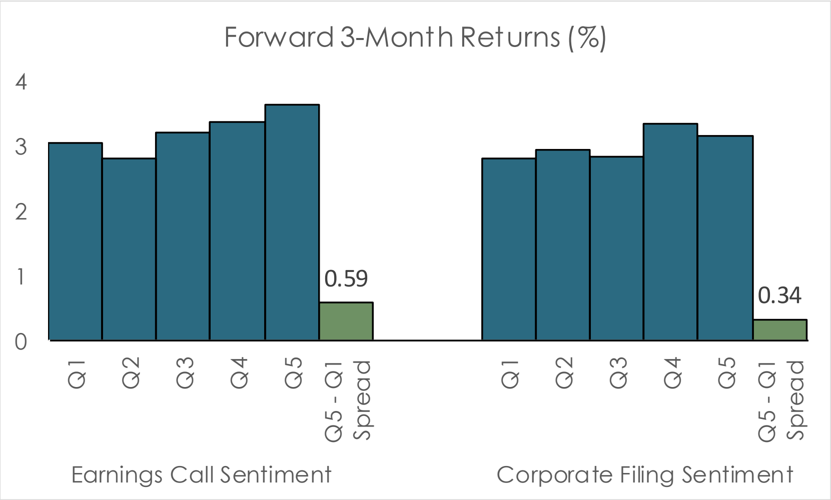
Academic research on sentiment analysis in the context of asset pricing can be traced back to the 2000s[6]. To quantify sentiment analysis a dictionary is required to classify words into positive and negative categories. The most commonly used financial dictionary is the Loughran-McDonald dictionary[7], developed and maintained by two professors at the University of Notre Dame. Their master word list contains more than 85,000 words, including all inflections of a given word and designed for finance-specific contexts. Using the classification of positive and negative words from the dictionary, a sentiment score can be calculated for each quarterly filing, as well as the Q&A section of each earnings call.

NLP FACTOR CASE STUDY: SENTIMENT IN CORPORATE FILINGS AND EARNINGS CALLS

For these illustrations, the Russell 2000 Index (R2000) is utilized to assess the value of NLP sentiment factors in the analysis of U.S. small-cap stocks. Small-cap stocks receive less attention from investors and analysts than their large-cap counterparts, and therefore the potential benefit of digging into the relatively untapped textual information is greater.

A company with a more positive tone in its filings or calls receives a higher sentiment score and these companies are expected to perform better in the future. A univariate sort on the sentiment factors over the last 10 years demonstrates the top 20% of stocks with the highest earnings call sentiment scores (top quintile or Q5) outperformed the bottom 20% of stocks (bottom quintile or Q1) by 59 basis points (bps) over the following 3 months on average (Figure 2). Similarly, those with the most positive tone in quarterly filings (Q5) outperformed the least positive stocks (Q1) by 34 bps in the following 3 months.

Figure 2 Univariate Sort on Sentiment Factors
US Small-Cap Universe; 2009 – 2019



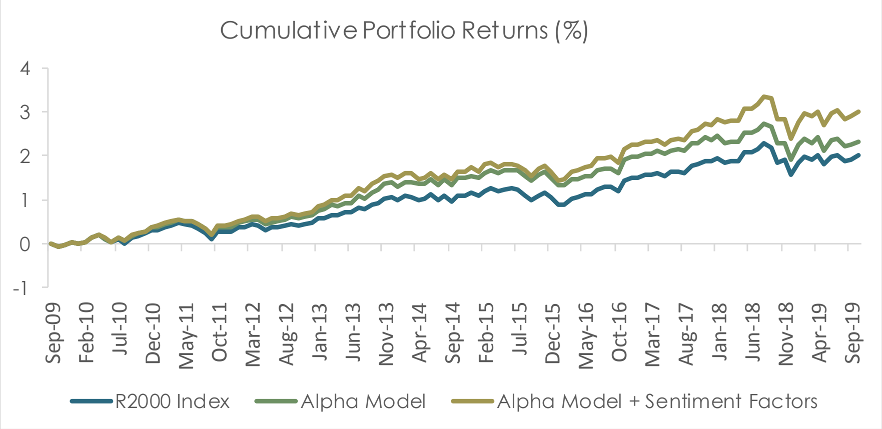
Regression tests also confirm that sentiment factors have predictive powers. In the univariate regressions of forward 3-month returns against the sentiment factors (Table 1), the coefficient of earnings call sentiment averaged 0.28 over the last 10 years and was statistically significant. The earnings call sentiment research also had an impressive hit rate of 72%, which means almost three quarters of the time the regression coefficient was positive. The coefficient of corporate filing sentiment was also positive and statistically significant over the last 10 years.

Table 1 Univariate Regression of Forward 3-Months Returns
US Small-Cap Universe; 2009 – 2019

	Average			Median
Sentiment Factors	Coefficient	T-Stat	Hit Rate	Coefficient
Earnings Call Sentiment	0.28	4.89	72%	0.31
Corporate Filing Sentiment	0.16	2.75	53%	0.06

How do these sentiment factors interact with traditional alpha factors? Do they complement our existing alpha model? A simple simulation consisting of 50% weight in the traditional alpha model and 25% in the two sentiment factors each, is compared to the performance of a 100% traditional alpha model and R2000 consisting of the entire universe (Figure 3).

Figure 3 Simulated Portfolio Performance
Alpha Model with Sentiment Factors vs. Alpha Model



Over the last 10 years, the **combined model** (gold line) consistently outperformed the **traditional alpha model** (green line), demonstrating there is additional information in the textual data not captured by the numbers in traditional quantitative factors.

CONCLUSION

In the era of information overload, the ability to quickly extract and analyze relevant information provides investors an edge in finding the best investment opportunities. Thanks to the rapid development in computing power, we can now leverage natural language processing techniques to gain insights from terabytes of unstructured textual data within hours or even minutes. At Chicago Equity Partners, our domain knowledge combined with programming capabilities allows us to harness this unique source of alpha. The resulting proprietary factors not only improve the performance of our traditional quantitative alpha model, but also afford us the first mover advantage of fishing in less crowded ponds.

This material has been distributed for informational purposes only and does not constitute investment advice or a recommendation of any security or investment service offered by Chicago Equity Partners, LLC. The material presented reflects the opinions of the author and is subject to change without notice. The opinions and themes discussed herein may not be suitable for all investors. The information set forth herein has been obtained or derived from sources believed to be reliable, but its accuracy is not guaranteed. Past performance is not indicative of future results.

ENDNOTES

[1] Price movement is defined as daily absolute value of total return, summed up over the corresponding time period.

[2] The 4 weeks following 4 earnings calls over a 52-week year account for 7.7% of total number of trading days (4 / 52 = 7.7%). The excess impact is 9.4% / 7.7% - 1 = 21%.

[3] Defined as price movements above those achieved if information were dispersed uniformly throughout the year.

[4] Hirshleifer and Teoh (2003)

[5] One has to invest in the infrastructure in order to maintain the storage and computing capacity, but the cost of doing so has decreased dramatically.

[6] Antweiler and Frank (2004), Li (2006).

[7] See Loughran and McDonald (2011). Data available at https://sraf.nd.edu/textual-analysis/

REFERENCES

Antweiler, Werner, and Murray Z. Frank, 2004, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, Journal of Finance.

Hirshleifer, David A. and Teoh, Siew Hong, 2003, Limited Attention, Information Disclosure, and Financial Reporting. JAE Boston Conference October 2002. Available at SSRN: https://ssrn.com/abstract=334940 or http://dx.doi.org/10.2139/ssrn.334940

Li, Feng, 2006, Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Available at SSRN: https://ssrn.com/abstract=898181 or http://dx.doi.org/10.2139/ssrn.898181

Loughran, Tim, and McDonald, Bill, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance 66, 35—65.

Loughran, Tim, and McDonald, Bill, 2016, Textual Analysis in Accounting and Finance: A Survey, Journal of Accounting Research 54, 1187-1230.