Predicting Corporate Acquisitions with Text in News

Name: Daniel Wu
USCID: 1895-2460-22
Email: ddwu@usc.edu

Introduction

Reading the news is a reliable way to stay updated on current financial events, and often stirs speculation on the future of corporations. This document proposes an automated approach using text-patterns in the news to predict corporate acquisitions. An acquisition involves a "bidding firm" purchasing more than fifty percent of the outstanding stock issued by a "target firm", granting the bidding firm complete control rights over the target firm. It is well documented in finance research literature that a significant proportion of these target firms experience a run-up (increase) in their stock price in the days prior to the announcement, suggesting the public knew that an acquisition was imminent despite efforts from company managers hiding it. Thus, an accurate text-based tool to predict acquisition would be valuable information, depending on the incentives of the informed participant. Investors can make easy profits. Regulators can monitor illegal activity more accurately[1]. Competing firms can shield or subtly hint their plans more effectively.

One theory for pre-acquisition stock price run-ups posits that information leaks privately from an employee of the acquiring or target firm, and word travels fast without any public news coverage. Another explanation is that savvy investors infer an impending acquisition based on public information in the news. The second theory entails a text pattern in news reports associated with the target firm. There is some evidence of this from Ahern and Sosyura (2014). They document that companies actively originate more news stories before acquisition announcements, and use substantially fewer understated words such as "*doubtful*", "*uncertain*" and "*weak*". The goal of this proposal is to design a prediction model that captures a broader semantic structure in the news leading up to acquisition announcements.

The main steps are as follows:

1. Convert raw news articles $\mathcal{D}$ to vector representations of text $\mathbf{M}_s$ using TF-IDF.

2. Derive benchmark vectors for acquisitions $\mathbf{v}_{1b}$ and non-acquisitions $\mathbf{v}_{2b}$ from $\mathbf{M}_s$.

3. Compute Cosine Similarity between the benchmark vectors and test instances, and use a classification algorithm to predict firm acquisition.

The Method section will outline the process for each step and provide suggestions on how the prediction model can be evaluated and improved. The Discussion section will point out advantages and shortcomings of the research design.

[1]Suppose a predictive model can be designed using public information. An outcome that predicts nothing, despite key individuals making significant stock transactions is indicative of insider trading.

Method

The Securities Data company (SDC) Platinum database contains comprehensive data on US acquisitions and their announcement dates dating back to the 1970s, which conveniently yields the labels for my data. This data will be useful to identify what company news and what time range to search for. I am interested in relatively recent acquisitions, and will select a random sample of 10,000 acquisitions in 2010 to 2020 to collect news from. For each acquisition in the SDC Platinum database, I will search for two sets of news articles. The first set will include all news articles of the target company from the day before the acquisition announcement to thirty days prior on Factiva[2]. Data collection of the news articles can be expedited with a web-scraping algorithm using the Selenium package on Python. The second set of news articles will include the same thirty day window during the same time, but will apply to a rival company that is not going to be acquired at the end of the thirty day window. The rival company can be identified through a simple match on key characteristics, such as size, industry and product similarity in another standardized database called Compustat. The second set of news provides a sample to eventually train classifications for non-acquisition observations, while the first set provides training for assignment to acquisitions. The combination of the two sets of news articles across all instances in the sample will form the corpus of this project.

In order to convert the news articles to vectors, I will implement the Term Frequency-Inverse Document Frequency (TF-IDF) method for each set s, news article i and acquisition j to construct TF-IDF vectors $\mathbf{u}_{sij}$ at the news article level. For each set s and each observation j, I will construct the TF-IDF vector $\mathbf{v}_{sj}$ at the company level, for each company, by averaging over all news articles $\mathbf{u}_{sij}$ returned in the thirty day window search. Let N denote the number of unique words in the corpus and D denote the number of companies. I will construct two matrices, $\mathbf{M}_1$ and $\mathbf{M}_2$, each of size N x D to store the TF-IDF vectors corresponding to news about the target company being acquired and rival company not being acquired respectively. This procedure fulfills the first step, and yields the following data for $s \in \{1, 2\}$:

$$\mathbf{M}_s = \begin{array}{c} \\ word_1 \\ word_2 \\ \vdots \\ word_N \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{v}_{s1} & \mathbf{v}_{s2} & \cdots & \mathbf{v}_{sD} \end{array} \\ \begin{bmatrix} m_{s11} & m_{s12} & \dots & m_{s1D} \\ m_{s21} & m_{s22} & \dots & m_{s2D} \\ \vdots & \vdots & \ddots & \vdots \\ m_{sN1} & m_{sN2} & \dots & m_{sND} \end{bmatrix} \end{array}$$

This data structure allows the construction of a "benchmark" TF-IDF vector $\mathbf{v}_{sb}$ of length N for each set $s \in \{1, 2\}$ by averaging the corresponding matrix elements across all columns, completing step two. Intuitively, this benchmark vector captures the average text vector representation of the news leading up to an acquisition for set 1, or not leading up to an acquisition for set 2.

[2]Factiva is a global search engine for news articles, available through USC libraries.

Mathematically, the "benchmark" TF-IDF vector is computed as follows:

$$\mathbf{v}_{sb} = \frac{1}{D}\sum_{i=1}^{D}[m_{s1i},\ m_{s2i},\ \ldots,\ m_{sNi}]^T, \qquad \text{for } s \in \{1,2\}$$

These benchmark vectors provide an anchor for test data to apply Cosine Similarity on. Test news articles can be converted to vectors using the same TF-IDF procedure used to process training data, and the following measure can be obtained for each set $s \in \{1,2\}$ and for each observation $j \in \{1,2,\ldots,T\}$ in the test data:

$$Sim_{sj} = \frac{\mathbf{v}_{sj} \cdot \mathbf{v}_{sb}}{\|\mathbf{v}_{sj}\|\|\mathbf{v}_{sb}\|}$$

The classification rule for each training point is as follows:

$$C_j = \mathbb{1}\big[\max\{Sim_{1j}, Sim_{2j}\} = Sim_{1j}\big]$$

Thus, the test observation is predicted to be an acquisition if its text similarity to the acquisition benchmark vector is greater than that of the non-acquisition benchmark vector. This computation completes step three. One way to evaluate the success of this design is to simply calculate the accuracy rate, which is the ratio of correctly classified instances to all test instances. Another interesting performance measure is recall, which provides the ratio of correctly predicted acquisitions to all true acquisitions. Recall is especially important compared to precision, since it accounts for misclassified acquisitions which are missed opportunities. I will initially set a standard for both accuracy and recall at 90%. If this standard is not met, I will increase the training data in increments of 1000 observations. If the standard is still not met after 10 increments, I will redesign the algorithm to take multiple Cosine Similarities of vectors at the observation level, and compute the average of the Cosine Similarity scores instead of using just one score with the benchmark vector at the set level in the classification algorithm. The difference in this new design is that an average is taken over Cosine Similarity scores instead of text vectors, and will preserve the structure of company specific news to be used for classification. I conjecture this new design may be a more accurate model if test news articles are highly similar to news articles from a specific company in the training data, but dissimilar to another company despite having the same label.

Discussion

A common method to predict financial events is to gauge sentiment in the news, which narrows the analysis down to specific words or features. One advantage of synthesizing TF-IDF and Cosine Similarity is that it allows for a broader semantic analysis of the vocabulary used in the news. TF-IDF incorporates information from all words in the news, which is important if non-semantic words are predictive of acquisitions. Moreover, combining Cosine Similarity with TF-IDF is a computationally efficient and reliable way to reduce text information down to an index. The Cosine Similarity score may not be meaningful, but the vector weights responsible for high scores are. I may be able to uncover interesting semantic patterns in the news about what kind of words are used prior to an acquisition announcement, by observing which words have high weights in both vectors before computing Cosine Similarity.

The TF-IDF method to convert raw news articles to text vectors is appropriate for many reasons. Firstly, it automatically penalizes common words across news articles with the IDF component. This provides a safety buffer for stop words that I have neglected to remove during processing, as those stop words will receive a low weight in the vectors. Additionally, boiler plate words in newswires will also receive a low weight, since they don't contain useful signals but appear in multiple news articles. The TF-IDF method loads heavily on "idiosyncratic" words that appear frequently within news articles but rarely across news articles. One potential drawback is that these idiosyncratic words in the training data may be purely random when generalized to the test data, yielding spurious Cosine Similarity scores. Conversely, words that truly signal an impending acquisition may be underrepresented in the training data. One way that I have addressed the former overfit problem is with the rival company's non-acquisition set of news articles that uses the same time window. Thus, time-specific idiosyncratic words such as "*Election*", "*Ballot*" and "*Vote*" during 2016 will equally bias the acquisition and non-acquisition vectors, which is not a problem for classification. However, idiosyncratic words that don't vary over time will not be accounted for, and may require a modification to this design.

References

Ahern, Kenneth R, and Denis Sosyura, 2014, Who writes the news? corporate press releases during merger negotiations, *The Journal of Finance* 69, 241–291.

Wordcount: 1,544