**Predicting Acquisition Targets with 10-K Reports**

Zhihan Li

**Abstract**

Predicting acquisition targets is a topic of great interest as it provides useful implications for buy-side M&A scenarios as well as for investors in the secondary market. Most previous papers about M&A prediction were limited to analysis on financial measures. This study aims to exploit textual data for predicting acquisitions through topic modeling on annual financial disclosures. The result demonstrates that models built upon carefully tailored topic features, together with financial measures, can achieve noticeable improvement in predictability over models incorporating only financial measures.

1. **Introduction**

The term “acquisition” refers to the action of one company purchasing shares or segments of another company. It is a significant component of corporate operation. Acquirers usually seek to grow their current businesses or enter new businesses through acquisitions and expect such transactions to create synergies. On the other hand, acquisition is also a means of capital restructuring. Gugler and Konrad (2002) suggested that companies can utilize acquisitions to adjust to optimal financial structures in the sense that firms with high debt-to-equity ratio tend to target at companies with lower debt-to-equity ratio, and vice versa.

Most previous studies surrounding the topic of M&A prediction focused primarily on financial measures. Palepu (1986) examined six assumptions about takeover targets, including inefficient management hypothesis, growth-resource mismatch hypothesis, industry disturbance hypothesis, size hypothesis, market-to-book hypothesis, and price-earnings hypothesis. Cudd and Duggal (2002) replicated the Palepu (1986) study, and Ambrose and Megginson (1992) extended the Palepu (1986) study by accounting for proportions of tangible assets in the target companies. Hasbrouck (1985) investigated takeover targets’ Tobin’s Q, leverages, and liquidity. Lee, Maeur, and Xu (2018) explored how human capital relatedness affect the likelihood of mergers and acquisitions.

Some papers attempted to approach the subject with textual data. For instance, Routledge, Sacchetto, and Smith (2013) run text regression on 10-K reports with features of unigrams and multiword phrases. Xiang, Zheng, Wen, Hong, Rose, and Liu (2012) tried to exploit textual information from social media sites by building topic models on profiles and news articles from TechCrunch and CrunchBase.

This study aims to contribute to the development of the natural language processing approaches to M&A prediction. Based on the Hasbrouck (1985) study, the current study first investigated Tobin’s Q, leverages, liquidity, and other financial measures. Then, topic modeling on 10-K reports was performed, from which topic distributions were taken as textual features, together with financial measures, to build logit regression models for predicting potential acquisition targets.

1. **Data**

This section describes the data for the study. The procedures involve identifying acquisition targets and control companies and collecting financial measures and annual financial disclosures. All data were collected from the S&P Capital IQ platform.

* 1. **Treatment Group and Control Group**

The data collection started with identifying the acquisition targets and non-targets (control group). From the transaction screening module of Capital IQ, an initial list of 2804 candidate transactions were obtained by selecting only acquisition transactions taking place between Jan. 1, 2005, and Sept. 20, 2021, for which the targets’ primary geographical location and primary listing are both in the United States. The list was then truncated by removing bids for less than 20% interest. From these transactions, target companies that do not belong to the financial industry were taken. Finally, only targets with available 10-K filed within one year prior to the announced transaction date were selected, leading to a total of 345 target companies in the dataset.

A corresponding control group consisted of 345 non-targets to match the 345 targets with one-to-one correspondence were obtained by controlling the variables of primary industry and firm size (total assets). The detailed matching procedure is as follows: for the N target companies in the treatment group with primary industry M, a pool of match-up companies was identified in the Capital IQ company screening module by selecting currently operating public companies in the United States whose primary industry is also M. Then, from the pool, N non-target companies were taken without replacement in the sense that the chosen companies were not involved in any M&A or private placement recorded in Capital IQ during the years of 2020 and 2021 and the total difference between the summed total assets of the N target companies and that of the N selected match-ups is minimized.

* 1. **Financial Measures**

The financial measures considered for this study are Tobin’s Q, market capitalization, net income, cash & cash equivalents, total cash & ST investments, TDETB, TLDEBT, CFIN, CNFIN, TDETB\_mcp, TLDEBT\_mcp, CFIN\_mcp, CNFIN\_mcp, TDETB\_ast, TLDEBT\_ast, CFIN\_ast, and CNFIN\_ast. Details are provided in Table 1. Tobin’s Q, TDEBT, TLDEBT, CFIN, and CNFIN are adapted from Hasbrouck (1985). X\_mcp and X\_ast are simply variants of X with different denominators.

Tobin’s Q reflects whether a company is undervalued or overvalued in the market. TDEBT and TLDEBT inform of leverages. CFIN and CNFIN stand for liquidity. Hasbrouck (1985) explained the interest in investigating these variables: Tobin’s Q less than 1 usually suggests undervaluation and thus signals desirable takeover targets; leverages reflect available debt capacity and link to capital structures; liquidity, according to Palepu (1986), represents resources and is able to indicate growth-resource mismatch; both leverages and liquidity are related to managerial inefficiency.

All the financial measures were obtained from Capital IQ. For the target companies, financial measures were taken from the previous year of their transaction announcement year. For the non-target companies (control group), financial measures were taken from CY2020.

Table 1

|  |  |
| --- | --- |
| Tobin’s Q | (market capitalization + total liabilities) / (total equity + total liabilities) |
| TDEBT | total liabilities / total equity |
| TLDEBT | long-term liabilities / total equity |
| CFIN | total cash &ST investments / total equity |
| CNFIN | (total cash & ST investments – current liabilities) / total equity |
| TDEBT\_mcp | total liabilities / market capitalization |
| TLDEBT\_mcp | long-term liabilities / market capitalization |
| CFIN\_mcp | total cash &ST investments / market capitalization |
| CNFIN\_mcp | (total cash & ST investments – current liabilities) / market capitalization |
| TDEBT\_ast | total liabilities / total assets |
| TLDEBT\_ast | long-term liabilities / total assets |
| CFIN\_ast | total cash &ST investments / total assets |
| CNFIN\_ast | (total cash & ST investments – current liabilities) / total assets |

* 1. **Financial Disclosures**

The textual data for the study were from 10-K forms, or more specifically, the section of Management’s Discussion and Analysis (MD&A) in 10-K. For target companies, 10-K filed within one year prior to their takeover announcement date were taken. For the non-target companies, 10-K filed in 2020 were taken. The rationale is to leave a one-year window for predictions. To further validify the one-year horizon, 15 companies were randomly selected from the 345 target companies. In the News Articles module of the Pitchbook platform, news articles within one year prior to the takeover announcement date for the 15 companies were checked to ensure that there was no takeover information leakage during the one-year window.

1. **Topic Modeling**

This section elaborates on text preprocessing and topic modeling.

* 1. **Text Preprocessing**

The texts come from the MD&A section of 10-K, which is required to file to SEC annually by a public company. The MD&A section contains the management board’s reflection over the company’s operations and performances.

For preprocessing, the texts were tokenized, and punctuations and numerical numbers were removed. Common stop words were removed with the NLTK stop word list. The remaining words were lemmatized. In addition to the common stop words, it was recognized that these texts from financial disclosures contained much industry noise that negatively affect the quality of the topic models. Industry noise refers to words that reflect information specific to each industry. As the variable of the industry was controlled during the data collection process and this study intends to explore firm specific factors relevant to acquisitions, industry noise is irrelevant and was removed in a systematic way as follows.

There were ten industry sectors involved in the dataset, including Communication Service, Consumer Discretionary, Consumer Staples, Energy, Healthcare, Industrials, Information Technology, Materials, Real Estate, and Utilities. For each sub-corpus grouped by industry sector, after a layer of most frequent words with TF-IDF scores lower than a certain threshold (usually at 0.04 or 0.05) was dropped, LDA topic models on the remaining texts generated topics that were mainly about the respective industry. Subsequently, words corresponding to these industry-specific topics were extracted as industry noise and inserted into the stop words list, which helped clean up industry noise from the general corpus.

After the industry noise was handled, bigrams were identified in the corpus. The mechanism was that two tokens appeared together frequently would be combined as a single token. For instance, “new” and “york” will be converted into “new\_york”. A final round of text cleaning was completed by removing all words in the general corpus with TF-IDF scores lower than a threshold S, where the values of S were in the range between 0.0001 and 0.05 and led to different prediction performances described in Section 4. With a threshold of 0.02, the resulting models achieve the best performances with the highest R-squared values. However, this threshold drops a huge number of words and shows a sign of overfitting in terms of the topics generated. The threshold of 0.001 was preferred as it is much less prone to overfitting and the resulting topics are more coherent.

* 1. **Topic Models**

The specific type of topic models considered in this study is the Latent Dirichlet Allocation (LDA) proposed by Blei, Ng, and Jordan (2003). LDA is a generative probabilistic model that takes a document as a mixture of topics and a topic as a unique set of distributions over words. LDA topic models with 10 topics were trained on the preprocessed corpus. The top 10 words for the resulting topics, when the threshold S equals 0.001, are presented in Table 2. The topics are generally associated with company operations and performances, free of industry noise.

Table 2

|  |
| --- |
| 1. expense, net, increase, revenue, base, cost, include, period, service, related |
| 1. cost, price, per, include, interest, increase, total, decrease, change, expenditure |
| 1. increase, revenue, due, cost, primarily, operating, include, change, increase, business |
| 1. net, cost, operation, increase, decrease, due, compare, primarily, sell, high |
| 1. year, end, revenue, increase, cost, compare, period, decrease, total, loss |
| 1. note, interest, net, debt, amount, asset, certain, payment, related, proceeds |
| 1. revenue, service, cost, base, may, use, increase, related, period, estimate |
| 1. year, increase, net, include, change, prior, cost, compare, due, base |
| 1. asset, estimate, financial, include, statement, amount, loss, use, future, liability |
| 1. cash, operating, net, asset, flow, use, financial, include, future, estimate |

1. **Prediction Modeling & Evaluation**

A logit regression classifier with only financial measures as features was trained as a baseline model. The R-square for the model is 0.027, with details presented in Table 3. Market capitalization, net income, cash & cash equivalents, and total cash & short-term investments were found statistically significant through hypothesis testing on the regression coefficients.

Logit regression classifiers with only topic features were trained for comparison. Topic features refer to the distributions of the ten topics in the MD&A documents for each company obtained from the LDA topic model, which led to ten new columns in the data frame. The R-squared for such models range from 0.018 to 0.039. When the TF-IDF threshold S was 0.02, the resulting model achieve the highest R-squared of 0.039, with details presented in Table 4. When the TF-IDF threshold S was 0.001, the resulting model has a R-squared of 0.025, with details presented in Table 5.

Finally, LDA models were trained on both financial measures and topic features. When the TF-IDF threshold S was 0.02, the resulting models obtain the highest R-squared between 0.061 and 0.073, with details displayed in Table 6 & 7. When the TF-IDF threshold S was 0.001, the resulting model have R-squared between 0.051 and 0.067, with details displayed in Table 8 & 9.

The result shows that topic features can largely enhance model predictability, increasing R-squared measure by about 150%. This huge increase in predictive power further proves the necessity of incorporating textual data for M&A prediction. The study also attempted support vector machine and decision tree for prediction, but they led to negative R-squared values.

Table 3

表格

描述已自动生成

Table 4

表格

描述已自动生成

Table 5

表格

描述已自动生成

Table 6

表格

描述已自动生成

Table 7

表格

描述已自动生成

Table 8

表格

描述已自动生成

Table 9

表格

描述已自动生成

1. **Conclusion**

This study proposes to augment the model predictability of acquisition targets by delving into 10-K corpus in addition to revisiting financial measures previous studies emphasized on. Topic features from topic modeling prove to be effective in building stronger prediction models. However, one of the weaknesses in the study is that the generated topics are not semantically coherent enough to help identify firm-specific signals indicative of potential acquisitions. Future directions for improvement include increasing the number of companies in the dataset and experimenting with different textual sources, such as financial news sites. Meanwhile, it is necessary to extend the study to cover private companies which do not have much information publicly available.

**References**

Ambrose, B. W., and Megginson, W. L. 1992. The Role of Asset Structure, Ownership Structure, and Takeover Defenses in Determining Acquisition Likelihood. *Journal of Financial and Quantitative Analysis, VOL. 27, NO. 4.*

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3 (2003) 993 – 1022.*

Cudd, M., and Duggal, R. 2000. Industry Distributional Characteristics of Financial Ratios: An Acquisition Theory Application. *The Financial Review, 41 (2000) 105-120.*

Gugler, K., and Konrad, K. A. 2002. Merger Target Selection and Financial Structure. *<https://www.wu.ac.at/fileadmin/wu/d/i/iqv/Gugler/Artikel/rio.pdf>*

Hasbrouck, J. 1984. The Characteristics of Takeover Targets q and Other Measures. *Journal of Banking and Finance, 9 (1985) 351-362.*

Lee, K. H., Mauer, D. C., and Xu, E. Q. 2018. Human Capital Relatedness and Mergers and Acquisitions. *Journal of Financial Economics, 129 (2018) 111-135.*

Palepu, K. G. 1985. Predicting Takeover Targets a Methodology and Empirical Analysis. *Journal of Accounting and Economics, 8 (1986) 3-35.*

Routledge, B. R., Sacchetto, S., and Smith, N. A. Predicting Merger Targets and Acquirers from Text. *<http://sulawesi.tepper.cmu.edu/pdf/ma_ste_latest.pdf>*

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., and Liu, C. 2012. A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. *The International AAAI Conference on Web and Social Media (ICWSM).*