Predicting Merger Targets and Acquirers from Text

Bryan R. Routledge, Stefano Sacchetto, and Noah A. Smith*

Abstract

We explore the use of a U.S. firm's SEC filings to predict whether the firm will be an acquirer or a target of an acquisition within a year of the filing. Our approach uses text regression, in which frequencies of words and phrases in the document are used as independent variables in a logistic regression model. We find that word and phrase features have significant predictive power in models of being an acquirer or a target. In each case, the best performing models involve a different use of text alongside standard financial variables.

JEL Classifications: G34, C25
Keywords: Mergers and Acquisitions; Textual Analysis; Machine Learning

*Routledge, Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh 15213, PA, email: routledge@cmu.edu, phone: +1-412-268-7588; Sacchetto (corresponding author), IESE Business School, University of Navarra, Av. Pearson 21, Barcelona 08034, Spain, email: ssacchetto@iese.edu, phone: +34-626-268-745; Smith, Computer Science & Engineering, University of Washington, 185 Stevens Way, Seattle 98195, WA, email: nasmith@cs.washington.edu, phone: +1-206-685-3134. All remaining errors are our responsibility.

1 Introduction

Mergers and acquisitions (M&As) play a key role in the economy. At the aggregate level, M&A transactions represent the main mechanism for consolidation and restructuring within industries, and their value as a fraction of U.S. GDP is substantial (5.8% between 1980 and 2011).[1] At the individual company level, takeovers constitute major investment decisions and an effective way to discipline inefficient managers. Given their importance in the economy, it is no surprise that M&As have attracted a great deal of attention among researchers—there is a wide body of theoretical and empirical research surrounding mergers.

Mergers are important, but they are also a relatively infrequent event: An average of about 5% of public firms have been acquired every year between 1980 and 2011.[2] Interestingly, but perhaps not surprisingly, mergers are difficult to predict. Table 1 lists prior research aimed at predicting targets where the general conclusion is that "predicting target firms with any accuracy has proven difficult" (Betton, Eckbo, and Thorburn, 2008). As the table shows, the explanatory power of the models (typically a logistic regression) is relatively low, with some evidence of interesting time-series properties (mergers come in "waves").[3] More directly, merger announcements typically involve a large premium over current prices (between 40% and 50% on average—see Eckbo, 2014) and lead to a large and rapid change in market prices suggesting the announcement is news to the market.[4] Accordingly, any improvement in the ability to predict which firms will be involved in a merger deal would prove to be very profitable for an investor in the stock market.

In this study, we exploit *text data* to predict whether a U.S. firm will participate in a merger or acquisition. Specifically, we use the disclosure in the firm's Management Discussion

[1] M&A volume is computed from SDC Platinum data as the aggregate value of completed acquisitions of U.S. target companies.

[2] This is the average fraction of firms that are dropped from the Center for Research in Security Prices (CRSP) sample during a year because they are acquired.

[3] See Andrade, Mitchell, and Stafford (2001), Harford (2005), and Rhodes-Kropf, Robinson, and Viswanathan (2005).

[4] Not coincidentally, many insider trading cases involve suspicious trades of insiders around the date of a merger announcement. See Keown and Pinkerton (1981) or `http://www.sec.gov/spotlight/insidertrading/cases.shtml` for recent examples.

Study	Sample	Pseudo R^2
Hasbrouck (1985)	258 firms for 1976–1981	0.054 to 0.102
Palepu (1986)	419 firms for 1971–1979	0.0695 to 0.1245
Morck, Shleifer, and Vishny (1988)	371 firms for 1981–1985	Not reported
Ambrose and Megginson (1992)	475 firms for 1981–1986	0.03 to 0.078
Shivdasani (1993)	346 firms for 1980–1988	Not reported
Comment and Schwert (1995)	21,887 firm-year observations for 1977–1991	Not reported
Cremers, Nair, and John (2009)	83,752 firm-year observations for 1981–2004	0.0176
	15,332 firm-year observations in IRRC for 1991–2004	0.0495
Hoberg and Phillips (2010)	50,104 firm-year observations for 1997–2006	Not reported
Edmans, Goldstein, and Jiang (2012)	100,160 firm-year observations for 1980–2007	0.015 to 0.02
Chatterjee, John, and Yan (2012)	68,950 firm-year observations for 1984–2004	Not reported
Cocco and Volpin (2013)	319 UK firms for 2002–2008	0.088 to 0.095

Table 1: Summary of prior results. Predicting the binary event "was target" (T) or "not" using various financial variables. The common measure of performance in this area is the "pseudo R^2," defined in Equation 4. The second sample reported for Cremers, Nair, and John (2009) uses data from the Investor Responsibility Research Center (IRRC).

and Analysis (MD&A), a section of the annual Form 10-K filing. We consider two predictions: (i) will the firm be an acquirer in the subsequent year after the filing? and (ii) will the firm be the target of an acquisition in the subsequent year? We explore the usefulness of text in making these predictions, and demonstrate how text-based forecasting models can offer intuitive hints about upcoming mergers.

Our methodology uses a large sample (tens of thousands of disclosures) to infer predictive cues in text; this is accomplished by estimating a regularized logistic regression model. We measure the pseudo R^2 on a held-out sample of disclosures, showing how our model compares with a baseline financial forecasting approach that does not use text. We find that combining the two kinds of information, i.e., financial variables and text, gives a marked performance increase over both. In particular, when predicting acquirers, text adds substantially to the performance of the baseline model, but it also performs well on its own: The pseudo R^2 increases from 0.0696 in the baseline model to 0.1022 in the text and financial model, and it is 0.0528 in the text only model. For the target prediction task, the pseudo R^2's are almost identical in the baseline and text only models (0.0262 and 0.0267, respectively) and the pseudo R^2 increases to 0.0294 in a model that combines information from financial variables

and text.

We also estimate a predictive model based on multiword phrases. This "phrases only" model provides better interpretable results than predictive models based on individual words, and it allows us to investigate the drivers of takeover decisions. Interestingly, we find that firms that use phrases associated with poor economic performance (such as 'net loss') have higher probability of becoming takeover targets. This result is consistent with the predictions of the Q-theory of M&As (Jovanovic and Rousseau, 2002): Poor-performing firms are likely to be targeted by companies, or investors, that acquire inefficiently managed assets and put them to a more productive use. The phrases only model also highlights other determinants of merger synergies: Phrases that point to potential tax benefits (e.g., 'effective tax rate' or 'loss carryforwards') and those that indicate the presence of financial constraints ('credit agreement') imply a higher probability of becoming a takeover target. This empirical finding confirms the importance of financial synergies as a driver of mergers and acquisitions (see Lewellen, 1971 and Leland, 2007).

We then propose a new way to capture interactions between continuous independent variables and high-dimensional textual variables that sacrifices neither interpretability nor computational efficiency. We find that, for target prediction, a model that interacts Tobin's Q with text generates the highest predictive power among all the models that we consider, with a pseudo R^2 of 0.0342. Moreover, many text features present significant interactions with Q: For instance, the word 'value' has a positive effect on the probability of becoming a target when used by low Q firms, and a negative effect for high Q firms.

The central contributions of this study are (i) a demonstration of the predictive utility of text disclosures for takeover bids, and (ii) modeling innovations for scalable text-driven forecasting. This paper contributes to the literature that studies the determinants of corporate acquisition decisions (see studies reported in Table 1). To the best of our knowledge, this is the first paper that uses text regressions to predict takeover targets and acquirers. The only other paper that uses variables constructed from text to predict M&A events is

Hoberg and Phillips (2010). However, the approach used by Hoberg and Phillips (2010) is substantially different from ours: They build measures of product market similarity across firms from 10-k product descriptions, and employ them as explanatory variables in standard logistic regressions to predict targets and acquirers. Instead, we use text regressions to study the predictive power of words and phrases used by the management in their annual 10-k discussion and analysis for takeover events.

More in general, we contribute to the growing literature that uses data extracted from text to study corporate finance decisions. Loughran and McDonald (2013) and Jegadeesh and Wu (2013) perform textual analyses of the initial public offering prospectuses to study stock returns of IPO firms. Hoberg and Maksimovic (2015), Bodnaruk, Loughran, and McDonald (2013) and Buehlmaier and Whited (2014) generate text-based measures of financial constraints to study corporate investment, financing decisions, and stock returns, respectively. Our paper shows that information from text is useful to predict mergers and acquisitions, which represent one of the most important investment decisions that firms undertake.

The paper proceeds as follows. Section 2 describes the data and the construction of the financial and text variables used for estimation. Section 3 presents the estimation methodology, Section 4 the baseline regression results using standard financial variables, and Section 5 the results incorporating text variables. Section 6 develops and estimates a predictive model that interacts financial and text variables, and Section 7 concludes.

2 Data

The study is based on two kinds of data: M&A events and financial disclosures.

2.1 Takeover Bid Event Data

Our data are drawn from Thomson Reuters' SDC Platinum database. Here we focus on the list of pairs of companies that have made (acquirer) and received (target) a takeover offer in

Dataset	Number of Firm-Year Observations	Number of Acquirers	Number of Targets
Training (for parameter estimation)	33,085	2170	2145
Development (for hyperparameter tuning)	5,687	369	240
Test (for measuring R^2)	5,647	1013	400

Table 2: Summary of datasets used in this study.

a given year from the SDC Platinum database. Such offers may eventually be unsuccessful or successful. Each offer occurs on a specific date; though data are available going back to 1978, we focus on 1995–2011, the period that overlaps with our disclosure dataset (Section 2.3).

In order to exclude acquisitions of minority interest in the target or stock repurchases, we drop cases in which the bidder already owns more than 50% of the target shares prior to the announcement of the bid. We also drop a bid if the percentage of shares that the bidder is seeking to acquire is less than 50% of the target shares, and, if this information is missing, if the fraction of shares held by the bidder after a completed transaction is less than 50%. We also drop observations that SDC labels as "block purchases," "creeping acquisition," and "privatization."[5] This definition of a "takeover" is standard in the literature (see Betton, Eckbo, and Thorburn, 2008). Over the 1995–2011 period, we have 142,454 takeover bids. Many of these, however, involve private or non-US companies that do not file with the US Securities and Exchange Commission (SEC). Takeovers where at least one of the parties (target or acquirer) is public number 55,508. Lastly, we focus on transactions where we have both financial data (via Compustat) and text data (via the 10-K annual report) so that we can fairly compare our text model with existing studies. The final size of the data is summarized in Table 2.

2.2 Financial Data

The financial data is from Compustat. The specific explanatory variables we use are the usual and standard ones in this literature: Tobin's Q (ratio of the market value of company

[5]More details can be found at `http://mergers.thomsonib.com/td/DealSearch/help/def.htm`.

assets to book value), PPE (the book value of property plant and equipment), log of cash balance, the size of leverage (book value of debt over book value of assets), size (market value of equity) and return on assets (operating income divided by year-end book value of assets). For ease of interpretation we standardize these variables to have mean 0 and variance 1.[6]

2.3 Form 10-K Text Corpus

Our text data comes from the annual report, the "Form 10-K" that each publicly company files with the SEC. Inside the 10-K is a section called "Management's Discussion and Analysis" (MD&A). The MD&A section (about 6,000 to 9,000 words per company per year) is where management reviews the past year's financial and other results and discusses forecasts of the future.[7] From 1995 to 2011 we have 83,349 firm-year observations.

The text was processed similar to many studies that use text as a regressor. Punctuation was removed and all words were down-cased. Numerical, percentage, and dollar figures were replaced with a token. For example, '$50,000' and '$2.00' are both replaced by '$#' (recall that we have high quality financial information from the Compustat data). These individual words (and tokens) are called *unigrams*. To capture multiple word chains, we constructed multiword *phrases* from our data. Phrases are identified by applying a program that identifies each word's part of speech,[8] then conjoins common phrase patterns (adjective-noun, for example). Common phrases include: 'financial condition,' 'capital resources,' 'common stock,' 'qualitative disclosures,' 'market risk,' 'fair value,' and 'financial statements.' Combined, there are 236,480 unigrams and phrases in our "training" data; as described below. We discard features that are used less than 500 times and in less than 200 documents. This leaves 12,243 unique terms. Let $\text{freq}(j, n)$ denote the frequency of the jth term type

[6]Some of these variable are ratios and so are occasionally ill-defined. Following standard conventions, we winsorized the highest 1% of the observations, but the results are not sensitive to this assumption.

[7]The data were downloaded from the SEC's "Edgar" system. The relevant MD&A section was extracted with a regular expression. We obtain about 80% success extracting this section; some companies include the text "by reference." We tend to be slightly better with bigger companies, but were unable to find correlation between success and other company attributes.

[8]This is part of the Stanford Core NLP package, available at `http://nlp.stanford.edu/software/corenlp.shtml` and described in Toutanova, Klein, Manning, and Singer (2003).

in the nth document. Since these counts are highly skewed, we log-transform them. Our right-hand-side textual independent variables are $x_{n,j} = \log(1 + \text{freq}(j, n))$.

2.4 Joining the Data

For each disclosure, we determine whether the firm was the acquirer (A) in a takeover bid in the 364 days following the filing of the disclosure. Whether the firm was a target (T) was defined similarly. Not all events in our takeover bid data have a corresponding MD&A. Of the 65,000 observations we have 4,711 instances of "acquirer" and 3,816 instances of "target."[9] We reserved about 20% of these observations as a "blind test" sample we can use for later evaluations. None of the results in this article use any of that data.

3 Methodology: Regularized Logistic Regression

Given explanatory variables $\boldsymbol{x} \in \mathbb{R}^P$, our prediction $\hat{y}$ will range over binary values, 1 for a prediction that the firm will be a target of a takeover bid (or will make a takeover bid), 0 for a prediction that it will not. The prediction formula is:

$$\hat{y} = \arg\max_y p(y \mid \boldsymbol{x}) \tag{1}$$

We use a logistic regression model to define the probability distribution:

$$p(y = 1 \mid \boldsymbol{x}) = \frac{\exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}\right)}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}\right)} = \frac{1}{1 + \exp\left(-\beta_0 - \boldsymbol{\beta}^\top \boldsymbol{x}\right)} \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^P$ are the parameters and $\beta_0 \in \mathbb{R}$ is a bias term.

Given a training sample $\langle\langle \boldsymbol{x}_1, y_1\rangle, \langle \boldsymbol{x}_2, y_2\rangle, \ldots, \langle \boldsymbol{x}_N, y_N\rangle\rangle$, there is no closed-form maximum likelihood solution for $\boldsymbol{\beta}$; one must typically solve a convex optimization problem to

[9] These figures do not match, since many mergers involve non-U.S.-listed companies that do not file disclosures with the SEC.

find a global solution, which is in general not unique. Our approach is to seek a *regularized* estimate, penalizing solutions that make use of "extreme" values of $\boldsymbol{\beta}$. Regularization is extremely important in high-dimensional (large P) regression problems like the ones we face here.

A general class of regularized likelihood objectives is:[10]

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \tag{3}$$

This type of regularization is known as "elastic net" regularization Zou and Hastie (2005). It makes use of both the classic "ridge" regularizer Hoerl and Kennard (1970) as well as the more recently proposed "lasso" regularizer Tibshirani (1996). By interpolating between them, the elastic net can achieve sparse solutions, in which many dimensions of $\boldsymbol{\beta}$ are driven to zero, while avoiding the lasso's tendency to choose one among a set of highly correlated independent variables. Sparse solutions are attractive in settings like ours, where we wish to be able to inspect the model to gain an understanding of the signals on which it bases its predictions. Dimensions of $\hat{\boldsymbol{\beta}}$ that are zero have no impact on prediction, so that even if P is large, the "effective" P, which we denote $\tilde{P}$, is small enough for human inspection.

We use the `creg` package to fit our model parameters.[11]

Evaluation To evaluate the quality of a predictive model, we apply it to a test sample (a portion of our dataset separate from the data used to estimate parameters $\boldsymbol{\beta}$ and hyperparameters $\langle \lambda_1, \lambda_2 \rangle$; see Table 2). The pseudo R^2 measure is used as a measure of model

[10] In this notation, $\|\boldsymbol{z}\|_1$ is the ℓ_1 norm of the vector $\boldsymbol{z} \in \mathbb{R}^D$, $\sum_{d=1}^{D} |z_d|$.

[11] Implemented by Chris Dyer and available at `https://github.com/redpony/creg`.

quality:

$$R^2 = 1 - \frac{\sum_{n=1}^{N'} \log p(y_n \mid \boldsymbol{x}_n)}{\sum_{n=1}^{N'} \log \tilde{p}(y_n)} \tag{4}$$

where N' is the size of the test set and $\tilde{p}$ is the empirical marginal distribution over y in the training dataset.

Hyperparameters The values of the hyperparameters $\langle \lambda_1, \lambda_2 \rangle$ can have a large effect on what is learned. If either is too large, then $\boldsymbol{\beta}$ will be too constrained to fit the data; if either is too small, then our estimate will overfit the training data and not generalize well out of sample. We therefore require a method for selecting $\langle \lambda_1, \lambda_2 \rangle$. Our budgeted stochastic approach is as follows:

1. Define the range of values for $\log \lambda_1$ to be $(-\infty, \log \lambda_1^{max}]$, where λ_1^{max} is the smallest value that forces $\hat{\boldsymbol{\beta}} = \mathbf{0}$ (when $\lambda_2 = 0$). This can be calculated given the training sample (Park and Hastie, 2007). There is no analogous value λ_2^{max}, since ridge regularization does not drive $\boldsymbol{\beta}$ all the way to zero. Instead, we select λ_2^{max} to a value large enough such that, at λ_2^{max}, the sample variance of the fitted model's $\log p(y \mid \boldsymbol{x})$ is small.

2. Our search begins by solving for $\hat{\boldsymbol{\beta}}$ at K values of $\langle \lambda_1, \lambda_2 \rangle$ that are evenly spaced in the range defined above. These values are sorted by on the pseudo R^2 values they obtain for the *development* data (see Table 2) and stored in a list $\mathcal{L}$.

3. We propose a new value of $\langle \lambda_1, \lambda_2 \rangle$ and solve for the corresponding $\hat{\boldsymbol{\beta}}$. The new value is proposed based by (i) selecting two values from the top $E = 4$ elements of $\mathcal{L}$, (ii) taking a linear combination (interpolation), and (iii) adding noise. The new value of $\langle \lambda_1, \lambda_2 \rangle$ is added to $\mathcal{L}$ based on the pseudo R^2 value it obtains on the development data.

4. We repeat step 3, decreasing the amount of noise at each iteration, until the performance on the development data shows improvement in pseudo R^2 smaller than 10^{-4}.

4 Baseline: Financial Model

Our first model, which establishes baseline performance on our dataset, uses explanatory variables that are standard in the literature. We use an indicator variable for each calendar year (the year is the year the report is published: 1995–2011). Financial variables are all measured for the year-end of the 10-K report.[12] Our financial variables are: Q, the ratio of the market value of company's equity value to book; PPE, the book value of property plant and equipment; cash balance (logarithm); size of leverage (book value of debt over book value of assets); size (market value of equity); and return on assets (operating income divided by year-end book value of assets). For each of these, we transform with the z-score (normalize mean and variance) to ease interpretation.

For this 25-parameter model (β_0 plus $P = 24$ explanatory variables in $\boldsymbol{\beta}$), we estimated both an unregularized logistic regression model and a regularized one using the method described in Section 3. The regularized results are shown in Table 3. Results for the unregularized model were indistinguishable in performance; the coefficients were also similar.

For the acquirer prediction task, we achieve a pseudo R^2 just below 0.07, with the year and firm size as the strongest effects. For the target prediction task, we find that this model obtains a pseudo R^2 of 0.0262, which is comparable to, and perhaps stronger than, the values reported by the only two studies whose data temporally overlaps ours: Cremers, Nair, and John (2009), with a pseudo R^2 of 0.0176 on data from 1981–2004, and Edmans, Goldstein, and Jiang (2012), with a pseudo R^2 of 0.02 for the 1980–2007 period.[13] Overall, these experiments suggest that the target task is more difficult than the acquirer task.

[12] For a company with a year-end December 31, 2005, the 10-K report is typically filed in February 2006.

[13] The regressions in Cocco and Volpin (2013) have higher pseudo R^2 (0.088 to 0.095), but they use data for firms in the United Kingdom.

	Acquirer	Target
	Coefficient	Coefficient
Intercept	-2.8549	-4.6228
Year (max. coeff.)	(1995) 0.2459	(1998) 1.3441
Year (min. coeff.)	(2011) -1.0954	(2011) -0.9374
Q	0.0041	-67.9058
PPE	-0.2201	-0.0303
log Cash	0.0986	0.0256
Leverage	-0.0175	0.3660
Size	0.7867	-0.0300
ROA	-0.0603	-0.0029
Pseudo R^2	0.069568	0.026245

Table 3: The regularized maximum likelihood estimates for the baseline logistic regression model, and performance on test (out of sample) data. Here $\lambda_1 = 6.14 \times 10^{-6}$ and $\lambda_2 = 0.368$. Only the strongest positive- and negative-weighted year coefficients are shown for each model.

5 Prediction from Text

We next consider models that use text to make the same predictions. We incorporate explanatory variables corresponding to words and phrases selected as discussed in Section 2.3.

Table 4 summarizes the performance and effective size (number of nonzero coefficients) for various models that use text on its own or with financial data. As noted in Section 3, for regressions with large P (number of independent variables), regularization is crucial. The values of λ_1 and λ_2 are therefore reported.

For the acquirer prediction task, our key finding is that text adds substantially to the performance of the baseline model ("text and financial" line in Table 4), but performs well on its own ("text" line in Table 4). Ablating words and predicting based only on multiword phrases is highly detrimental ("phrases only" line in Table 4).

The results for target prediction are more nuanced. We find that text *on its own* performs on par with the baseline, but combining the two is harmful. Noting the large selected value for λ_2, we conjecture that the combined model performs poorly because the text variables demand strong regularization (there are many of them, most irrelevant), while the financial

		Acquirer				Target			
		Pseudo R^2	$\tilde{P}$	λ_1	λ_2	Pseudo R^2	$\tilde{P}$	λ_1	λ_2
§4	Financial (baseline)	0.069568	25	ϵ	ϵ	0.026245	25	ϵ	0.368
§5	Text	0.052793	118	75.5	64.9	0.026663	481	62.4	89.0
	Phrases only	0.036865	4,394	0.350	99.4	0.023285	1,816	5.99	89.1
	Text and financial	*0.102176	12,134	ϵ	ϵ	0.016735	12,182	0.368	90.2
	Text and 1000×financial	0.069393	1,849	90.1	0.368	0.029355	621	83.7	0.178
§6	Text/time	0.054632	157	48.6	2.19	0.017928	972	90.1	90.2
	Text/Q	0.053583	108	87.7	0.269	*0.034244	1,071	23.5	89.2

Table 4: Test (out of sample) data performance of models that incorporate text. $\tilde{P}$ is the number of nonzero coefficients. ϵ here denotes the value 6.14×10^{-6}. * Denotes strongest results, and § the section in the text where results are discussed.

variables do not. We therefore experimented with a model that increases each financial variable by a factor of 1,000, effectively weakening the regularization for financial variables. This led to the strongest performance for the target prediction task.

As noted in Section 1, text-based forecasting models can be explored to understand what textual hints the model uses. To do this, we consider the impact of each variable on the predictions. Impact is defined by Yano, Smith, and Wilkerson (2012) for the jth predictive variable as the model beta parameter times the sample mean of that feature. Here, we use a slightly different definition and consider impact as

$$\hat{\beta}_j \times \widehat{\text{std}}(X_{\bullet,j}) \tag{5}$$

where $\widehat{\text{std}}(X_{\bullet,j})$ is the sample standard deviation of the jth feature, measured on the test data.

$$\widehat{\text{std}}(X_{\bullet,j}) = \sqrt{\frac{1}{N'}\sum_{n=1}^{N'} x_{n,j}^2 - \left(\frac{1}{N'}\sum_{n=1}^{N'} x_{n,j}\right)^2}$$

Our impact can be understood as the effect a one standard deviation increase in a feature has on the predictive log-odds.

Past work has shown that, frequently, the most interpretable models tend not to be those

with the strongest performance (Yano, Yogatama, and Smith, 2013). This is because some valuable features are simply difficult to interpret (e.g., words in isolation whose actual usage patterns may not be obvious) and also because models can overfit the coefficients for spurious effects. We show in Table 5 the twenty phrases with highest positive and negative impact from the "phrases only," model. These phrases point to several potential determinants of merger decisions.

First, the phrases suggesting an increased likelihood of takeover highlight poor performance ('net loss', 'material effect', 'valuation allowance', 'financial condition'), while positive income phrases like 'net income' are associated with less chance of a takeover. This result is in line with the Q-theory of mergers and acquisitions (Jovanovic and Rousseau, 2002), which predicts that profitable companies are eager to acquire poor-performing firms and generate operational gains by putting their assets to a more productive use. Indeed, we find that, when managers point to poor performance in the annual MD&A, the company is more likely to be identified as a target by a potential acquirer.

Second, the high impact of 'credit facility' and 'credit agreement' (as well, perhaps, as 'accounts receivable' and 'convertible notes') suggest that financial constraints are important drivers of M&A decisions. This finding is consistent with merger theories based on the presence of financial synergies (Lewellen, 1971 and Leland, 2007): When two firms have imperfectly correlated cash flows and face financing frictions, a merger can generate a coinsurance effect that decreases expected bankruptcy costs and raises debt capacity, along with the potential benefits due to higher interest tax shields. Interestingly, and on a related note, we find that phrases such as 'tax benefit' and 'effective tax rate,' which highlight tax-driven motives for takeovers (e.g., purchasing tax deductions), are associated with a higher probability of becoming a target. In the list of phrases, 51 included 'tax' and both 'loss carryforwards' and 'net operating loss carryforwards.'

Finally, although not among the twenty with most impact, phrases including 'strategic' were strong indicators of a potential target. 'Strategic alternatives' and 'strategic initiatives'

Positive-impact phrases	Impact	Negative-impact phrases	Impact
credit facility	0.0593	net income	-0.0793
credit agreement	0.0534	marketable securities	-0.0531
total revenues	0.0502	financing activities	-0.0528
tax benefit	0.0486	intangible assets	-0.0508
effective tax rate	0.0452	financial statements	-0.0449
revenue growth	0.0423	business combination	-0.0434
net cash	0.0416	variable interest entities	-0.0393
financial results	0.0404	market conditions	-0.0384
net revenues	0.0399	market risk	-0.0360
financial condition	0.0386	private placement	-0.0351
net loss	0.0381	cash flow	-0.0333
prior year	0.0354	qualitative disclosures	-0.0332
material effect	0.0314	economic conditions	-0.0324
financial accounting standards	0.0313	balance sheet	-0.0322
valuation allowance	0.0306	operating costs	-0.0321
capital expenditures	0.0282	accounting principles	-0.0313
third quarter	0.0277	other expense	-0.0288
process research	0.0276	sales volume	-0.0284
accounts receivable	0.0270	managements discussion	-0.0276
convertible notes	0.0260	plan assets	-0.0275

Table 5: Phrases with greatest positive and negative impact (equation 5), from the "phrases only" target prediction model. Positive impact phrases on the left strongly encourage the model to predict that the firm will be a target; negative phrases on the right encourage the opposite.

are phrases often associated with companies floundering for direction and considering putting themselves up for sale (see Boone and Mulherin, 2007).

6 Smooth Interaction Model

The combined financial and text model in Section 5 exploits both kinds of independent variables, but it does not capture interactions between the two, which we expect to vary smoothly across continuous-valued financial and context variables. Given the high dimensionality of the latter, computational efficiency is a key concern. Let P_f and P_t denote the dimensionalities of the independent financial and textual variables (i.e., $P = P_f + P_t$). Further, let $\boldsymbol{x} = \langle \boldsymbol{x}_f, \boldsymbol{x}_t \rangle$ and $\boldsymbol{\beta} = \langle \boldsymbol{\beta}_f, \boldsymbol{\beta}_t \rangle$, with the same subscripts.

First, consider a straightforward approach to interactions that uses $O(P_f P_t)$ parameters, due to Yogatama, Heilman, O'Connor, Dyer, Routledge, and Smith (2011). Rather than financial variables, Yogatama et al. considered indicator variables for different (discrete) timesteps; they sought to model temporal variation in the association between a textual variable[14] and a binary response. Let:

$$p(y = 1 \mid \boldsymbol{x}) \propto \exp\left(\beta_0 + \boldsymbol{\beta}_f^\top \boldsymbol{x}_f + \boldsymbol{\beta}_t^\top \boldsymbol{x}_f + \boldsymbol{x}_t^\top \boldsymbol{\Gamma} \boldsymbol{x}_t\right) \tag{6}$$

where $\boldsymbol{\Gamma} = [\gamma_{i,j}]$ is a $\mathbb{R}^{P_f \times P_t}$ matrix of interaction coefficients. For example, if $x_{t,j}$ indicates the presence of the word *inflation* in the input document and $x_{f,i}$ is 1 if and only if the document was authored in 1999, then $\gamma_{i,j}$ is the *inflation* coefficient for 1999. Using a regularization function that penalized the squared difference between $\gamma_{i,j}$ and $\gamma_{i+1,j}$ (i.e., the interaction coefficients at *adjacent* timesteps), Yogatama, Heilman, O'Connor, Dyer, Routledge, and Smith (2011) estimated a model that captured smooth variation across time.

The main disadvantages of the Yogatama et al. model are (i) that it requires $O(P_f P_t)$ parameters, and (ii) the interaction between contextual variables $\boldsymbol{x}_f$ (such as our financial variables or Yogatama et al.'s temporal variables) is mediated through a discretized version of those variables.

We introduce a $O(P_f + P_t)$-parameter model that more naturally captures continuity in $\boldsymbol{x}_f$. The key idea is to assign to each $\gamma_{i,j}$ a parametric form that depends on K "basis" values for $\boldsymbol{x}_f$, $\{\boldsymbol{b}_k\}_{k=1}^K$, each in $\mathbb{R}^{P_f}$; per-basis parameters $\boldsymbol{\Theta} \in \mathbb{R}^{K \times P_t}$; and a similarity function *sim*:

$$\gamma_{i,j} = \sum_{k=1}^{K} \theta_{k,j} sim(\boldsymbol{x}_f, \boldsymbol{b}_k) \tag{7}$$

We can control the smoothness properties of the interaction by varying K: smaller K increases the expected covariance between "nearby" interaction coefficients. The selection of

[14] For simplicity of exposition, we assume these are binary.

the bases and the similarity function provide additional ways to incorporate structure into the estimate. Here, we explore the temporal (year) variable and Q, as these are the most predictive non-text features; more details are provided below.

The computational advantages of this approach are twofold. First, we now have KP_f interaction parameters, rather than P_tP_f. Second, given the K basis values, we can avoid explicitly representing $\mathbf{\Gamma}$ in our model, instead working with $\mathbf{\Theta}$. To the independent variables for instance $\boldsymbol{x}$, we concatenate

$$\langle sim(\boldsymbol{x}_f, \boldsymbol{b}_1) \cdot \boldsymbol{x}_t, \quad sim(\boldsymbol{x}_f, \boldsymbol{b}_2) \cdot \boldsymbol{x}_t, \quad \ldots, \quad sim(\boldsymbol{x}_f, \boldsymbol{b}_K) \cdot \boldsymbol{x}_t \rangle \tag{8}$$

with corresponding coefficients

$$\langle \boldsymbol{\theta}_{1,*}, \boldsymbol{\theta}_{2,*}, \ldots, \boldsymbol{\theta}_{K,*} \rangle.$$

How might the bases $\{\boldsymbol{b}_k\}_{k=1}^K$ be selected? We discuss below our heuristic methods for selecting bases for time and Q, which essentially space the bases out evenly through the data. In future work, they might be placed close to dense regions, or inferred together with the parameters of the model.

This model also lends itself well to interpretation. Given $\boldsymbol{x}_f$, the log-odds effect for a particular textual independent variable $x_{t,j}$ is

$$\beta_{t,j} + \sum_{k=1}^{K} \theta_{k,j} sim(\boldsymbol{x}_f, \boldsymbol{b}_k)$$

This can be inspected for a particular data instance, or plotted as a function of $\boldsymbol{x}_f$ to visualize trends.

How does this interact with regularization? We propose to regularize the per-basis interaction parameters $\boldsymbol{\theta}_{k,j}$. This equates to a prior expectation of marginal independence among $\boldsymbol{\theta}_{k,*}$, across k. If we consider two arbitrary values $\boldsymbol{x}_f$ and $\boldsymbol{x}'_f$, their interactions with the

textual variables are expected to covary the more similar they are, mediated through *sim* and the bases.

We present experiments with two variations of this smooth interaction model. In both cases, we select a single important non-text variable and work with text features selected by our "text only" model from Section 5.

6.1 Words and Time

We first consider applying the smooth interaction model to time and text. The date (day, month, and year) of each document is known. The bases for the interactions are pinned to January 1 of each year from 1992 to 2013. We define the similarity between a filing date d and a basis date b_k as (measured in days):

$$\frac{\max\{C - |d - b_k|, 0\}}{\sum_{k'=1}^{K} \max\{C - |d - b_{k'}|, 0\}} \tag{9}$$

where C equals 2.5 years. Intuititvely, d is most similar to basis dates that are closer, but only within a five-year window, outside of which the similarity is zero. Further, d's similarities to all basis dates sum to one. The advantage of this function is that the transformed data representation (equation 8) is *sparse* (when similarity is zero, the new variables take the value zero).

One drawback to this approach is that the feature vectors grow in length by as much as a factor of six (since each text feature interacts with as many as five time bases). We therefore pruned the 12,243-dimensional text feature vector to only include the features with non-zero coefficients in our text-only forecasting models. This reduced our feature vectors, prior to interaction with the 18 time-posts, to 481 and 118 for our prediction of target and acquirer, respectively.

The results are reported in Table 4 ("text/time" line). For our model predicting acquirer, we find a very slight improvement in pseudo R^2, compared to the text only model. Few of

the time interaction terms have non-zero weight in the regularized model. Those that are non-zero do not vary substantially across time-posts. This suggests that the text features have constant weight across time on the odds a firm will make an acquisition. This is surprising since there was substantial regulatory change in takeovers over our sample period. (For example, the method of accounting for a takeover changed in the post-Sarbanes-Oxley period after 2002.) One possibility for this finding is the aggressive pruning of the text features.

In our model predicting targets, we see a drop in performance relative to the text only model. 95 of the 482 text features have non-zero interactions with time. For example, the word 'cash' has a weight that steadily moves from negative to positive. That is, in 2000 'cash' is associated with a lower chance of being a target but by 2004 it is associated with a higher likelihood, *ceteris paribus*.

6.2 Words and Q

The smooth interaction model can also be used to capture the notion that words may be used differently depending on the situation the company is in. Noting the importance of Q in the baseline model (Section 4), we apply the smooth interaction model to text and Q. (Recall that Q is the market value to book value ratio.) The bases are the quintiles for the Q variable; since it is standardized, these are $\{-0.84, -0.25, 0.25, 0.84\}$. For a firm whose Q ratio is q, we let its similarity to a basis b_k be:

$$\frac{|q - b_k|^{-1}}{\sum_{k'=1}^{K} |q - b_{k'}|^{-1}}. \tag{10}$$

Again, for tractability we implement this model on highly pruned text feature vectors, of dimensionality 482 and 119 for prediction of target and acquirer, respectively. These are features that had non-zero weight in their respective text-only prediction models. The results are reported in Table 4 ("text/Q" line).

For acquisition predication, the interaction with the Q had no impact on the performance. In fact, none of the interaction terms have non-zero weight in the regularized model. This is evidence that the text based prediction of whether a company will make an acquisition is not dependent on the level of Q—text feature weights are constant across Q quintiles.

For target prediction, a model that interacts Q with text gives the strongest results observed so far. 353 of the 482 text features have non-zero dependence on Q. For example, the term 'value' has positive impact on the likelihood of being a takeover target when used by low Q firms and negative impact when used by high Q firms. The term 'tax,' interestingly, was most important and a positive takeover indicator for Q values close to the median (that is, zero for quintile 0 and 4). However, many of the terms have non-zero weights only for 2 of the 4 quintile bases we used, making it difficult to see patterns of Q and text interaction.

7 Conclusions

Mergers and acquisitions are important, infrequent events that are difficult to predict. In this study, we explore the use of a publicly traded U.S. firm's Management Discussion and Analysis (part of its SEC-mandated Form 10-K annual filing) to predict whether the firm will be an acquirer or a target of an acquisition within a year of the filing. Our approach uses text regression, which provides both strong performance and interpretability. In combination with standard financial variables, we find that word and phrase features can give much stronger acquirer predictive accuracy (measured using pseudo R^2, out of sample) than the former alone. For the more difficult task of predicting acquisition targets, we find that text interacts with the ratio of the market value of the company's assets to book value, and that the best predictive accuracy comes from our novel approach of efficiently capturing smooth interactions between text cues and a continuous-valued variable. Moreover, we find that phrases that indicate poor firm performance, tax benefits and the presence of financial constraints—thus pointing to potentially valuable operational improvements and financial

synergies that an acquirer can generate—are associated with a higher probability of becoming a takeover target. Overall, we conclude that text regression is a useful tool for predicting important corporate events such as mergers and acquisitions and to study their determinants.

References

Ambrose, Brent W., and William L. Megginson, 1992, The role of asset structure, ownership structure, and takeover defenses in determining acquisition likelihood, *Journal of Financial and Quantitative Analysis* 27, 575–589.

Andrade, Gregor, Mark Mitchell, and Erik Stafford, 2001, New evidence and perspectives on mergers, *Journal of Economic Perspectives* 15, 103–120.

Betton, Sandra, B. Espen Eckbo, and Karin S. Thorburn, 2008, Corporate takeovers, *Handbook of Corporate Finance: Empirical Corporate Finance* vol. 2, edited by B. E. Eckbo, Elsevier, Oxford, UK, 291–416.

Bodnaruk, Andriy, Tim Loughran, and Bill McDonald, 2013, Using 10-k text to gauge financial constraints, *Working Paper*.

Boone, Audra L., and Harold J. Mulherin, 2007, How are firms sold?, *Journal of Finance* 62, 847–875.

Buehlmaier, Matthias M. M., and Toni M. Whited, 2014, Looking for risk in words: A narrative approach to measuring the pricing implications of financial constraints, *Working Paper*.

Chatterjee, Sris, Kose John, and An Yan, 2012, Takeovers and divergence of investor opinion, *Review of Financial Studies* 25, 227–277.

Cocco, João F., and Paolo F. Volpin, 2013, Corporate pension plans as takeover deterrents, *Journal of Financial and Quantitative Analysis* 48, 1119–1144.

Comment, Robert, and G. William Schwert, 1995, Poison or placebo? Evidence on the deterrence and wealth effects of modern antitakeover measures, *Journal of Financial Economics* 39, 3–43.

Cremers, K. J. Martijn, Vinay B. Nair, and Kose John, 2009, Takeovers and the cross-section of returns, *Review of Financial Studies* 22, 1409–1445.

Eckbo, B. Espen, 2014, Corporate takeovers and economic efficiency, *Annual Review of Financial Economics* 6, 51–74.

Edmans, Alex, Itay Goldstein, and Wei Jiang, 2012, The real effects of financial markets: The impact of prices on takeovers, *Journal of Finance* 67, 933–971.

Harford, Jarrad, 2005, What drives merger waves?, *Journal of Financial Economics* 77, 529–560.

Hasbrouck, Joel, 1985, The characteristics of takeover targets q and other measures, *Journal of Banking & Finance* 9, 351–362.

Hoberg, Gerard, and Vojislav Maksimovic, 2015, Redefining financial constraints: A text-based analysis, *Review of Financial Studies* 28, 1312–1352.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies and competition in mergers and acquisitions: A text-based analysis, *Review of Financial Studies* 23, 3773–3811.

Hoerl, Arthur E., and Robert W. Kennard, 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12, 55–67.

Jegadeesh, Narasimhan, and Di Wu, 2013, Word power: A new approach for content analysis, *Journal of Financial Economics* 110, 712–729.

Jovanovic, Boyan, and Peter L. Rousseau, 2002, The Q-theory of mergers, *American Economic Review* 92, 198–204.

Keown, Arthur J., and John M. Pinkerton, 1981, Merger announcements and insider trading activity: An empirical investigation, *The Journal of Finance* 36, 855–869.

Leland, Hayne E., 2007, Financial synergies and the optimal scope of the firm: Implications for mergers, spinoffs, and structured finance, *Journal of Finance* 42, 765–807.

Lewellen, Wilbur, 1971, A pure financial rationale for the conglomerate merger, *Journal of Finance* 26, 521–537.

Loughran, Tim, and Bill McDonald, 2013, IPO first-day returns, offer price revisions, volatility, and form S-1 language, *Journal of Financial Economics* 109, 307–326.

Morck, Randall, Andrei Shleifer, and Robert W. Vishny, 1988, Characteristics of targets of hostile and friendly takeovers, *Corporate Takeovers: Causes and Consequences* edited by Alan J. Auerbach, University of Chicago Press, Chicago, IL, 101–136.

Palepu, Krishna G., 1986, Predicting takeover targets: A methodological and empirical analysis, *Journal of Accounting and Economics* 8, 3–35.

Park, Mee Young, and Trevor Hastie, 2007, l_1-regularization path algorithm for generalized linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69, 659–677.

Rhodes-Kropf, Matthew, David T. Robinson, and S. Viswanathan, 2005, Valuation waves and merger activity: The empirical evidence, *Journal of Financial Economics* 77, 561–603.

Shivdasani, Anil, 1993, Board composition, ownership structure, and hostile takeovers, *Journal of Accounting and Economics* 16, 167–198.

Tibshirani, Robert, 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 267–288.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer, 2003, Feature-rich part-of-speech tagging with a cyclic dependency network, in *Proceedings of the Human Language Technologies Conference of the North American Association for Computational Linguistics* pp. 252–259.

Yano, Tae, Noah A. Smith, and John D. Wilkerson, 2012, Textual predictors of bill survival in Congressional committees, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* pp. 793–802 Montréal, Québec.

Yano, Tae, Dani Yogatama, and Noah A. Smith, 2013, A penny for your tweets: Campaign contributions and Capitol Hill microblogs, in *Proceedings of the International AAAI Conference on Weblogs and Social Media* Boston, MA.

Yogatama, Dani, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith, 2011, Predicting a scientific community's response to an article, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Zou, Hui, and Trevor Hastie, 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320.