Predicting Risk from Financial Reports with Regression

Shimon Kogan
McCombs School of Business
University of Texas at Austin
Austin, TX 78712, USA
shimon.kogan@mccombs.utexas.edu

Dimitry Levin
Mellon College of Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
dimitrylevin@gmail.com

Bryan R. Routledge
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213, USA
routledge@cmu.edu

Jacob S. Sagi
Owen Graduate School of Management
Vanderbilt University
Nashville, TN 37203, USA
Jacob.Sagi@Owen.Vanderbilt.edu

Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

Abstract

We address a text regression problem: given a piece of text, predict a real-world continuous quantity associated with the text's meaning. In this work, the text is an SEC-mandated financial report published annually by a publicly-traded company, and the quantity to be predicted is *volatility* of stock returns, an empirical measure of financial risk. We apply well-known regression techniques to a large corpus of freely available financial reports, constructing regression models of volatility for the period following a report. Our models rival past volatility (a strong baseline) in predicting the target variable, and a single model that uses both can significantly outperform past volatility. Interestingly, our approach is more accurate for reports after the passage of the Sarbanes-Oxley Act of 2002, giving some evidence for the success of that legislation in making financial reports more informative.

1 Introduction

We consider a text regression problem: given a piece of text, predict a $\mathbb{R}$-valued quantity associated with that text. Specifically, we use a company's annual financial report to predict the financial risk of investment in that company, as measured empirically by a quantity known as *stock return volatility*.

Predicting financial risk is of clear interest to anyone who invests money in stocks and central to modern portfolio choice. Financial reports are a government-mandated artifact of the financial world that—one might hypothesize—contain a large amount of information about companies and their value. Indeed, it is an important question whether mandated disclosures are informative, since they are meant to protect investors but are costly to produce.

The intrinsic properties of the problem are attractive as a test-bed for NLP research. First, there is no controversy about the usefulness or existential reality of the output variable (volatility). Statistical NLP often deals in the prediction of variables ranging from text categories to linguistic structures to novel utterances. While many of these targets are uncontroversially useful, they often suffer from evaluation difficulties and disagreement among annotators. The output variable in this work is a statistic summarizing facts about the real world; it is not subject to any kind of human expertise, knowledge, or intuition. Hence this prediction task provides a new, objective test-bed for any kind of linguistic analysis.

Second, many NLP problems rely on costly annotated resources (e.g., treebanks or aligned bilingual corpora). Because the text and historical financial data used in this work are freely available (by law) and are generated as a by-product of the American

economy, old and new data can be obtained by anyone with relatively little effort.

In this paper, we demonstrate that predicting financial volatility automatically from a financial report is a novel, challenging, and easily evaluated natural language understanding task. We show that a very simple representation of the text (essentially, bags of unigrams and bigrams) can rival and, in combination, improve over a strong baseline that does not use the text. Analysis of the learned models provides insights about what can make this problem more or less difficult, and suggests that disclosure-related legislation led to more transparent reporting.

2 Stock Return Volatility

Volatility is often used in finance as a measure of *risk*. It is measured as the standard deviation of a stock's returns over a finite period of time. A stock will have high volatility when its price fluctuates widely and low volatility when its price remains more or less constant.

Let $r_t = \frac{P_t}{P_{t-1}} - 1$ be the return on a given stock between the close of trading day $t-1$ and day t, where P_t is the (dividend-adjusted) closing stock price at date t. The measured volatility over the time period from day $t-\tau$ to day t is equal to the sample s.d.:

$$v_{[t-\tau,t]} = \sqrt{\sum_{i=0}^{\tau}(r_{t-i} - \bar{r})^2 \Big/ \tau} \qquad (1)$$

where $\bar{r}$ is the sample mean of r_t over the period. In this work, the above estimate will be treated as the true output variable on training and testing data.

It is important to note that predicting volatility is not the same as predicting *returns* or *value*. Rather than trying to predict how well a stock will perform, we are trying to predict how stable its price will be over a future time period. It is, by now, received wisdom in the field of economics that predicting a stock's *performance*, based on easily accessible public information, is difficult. This is an attribute of well-functioning (or "efficient") markets and a cornerstone of the so-called "efficient market hypothesis" (Fama, 1970). By contrast, the idea that one can predict a stock's level of *risk* using public information is uncontroversial and a basic assumption made by many economically sound pricing models. A large body of research in finance suggests that the two types of quantities are very different: while predictability of returns could be easily traded away by the virtue of buying/selling stocks that are under- or over-valued (Fama, 1970), similar trades are much more costly to implement with respect to predictability of volatility (Dumas et al., 2007). By focusing on volatility prediction, we avoid taking a stance on whether or not the United States stock market is informationally efficient.

3 Problem Formulation

Given a text document $\mathbf{d}$, we seek to predict the value of a continuous variable v. We do this via a parameterized function f:

$$\hat{v} = f(\mathbf{d}; \mathbf{w}) \qquad (2)$$

where $\mathbf{w} \in \mathbb{R}^d$ are the parameters or weights. Our approach is to learn a human-interpretable $\mathbf{w}$ from a collection of N training examples $\{\langle \mathbf{d}_i, v_i \rangle\}_{i=1}^N$, where each $\mathbf{d}_i$ is a document and each $v_i \in \mathbb{R}$.

Support vector regression (Drucker et al., 1997) is a well-known method for training a regression model. SVR is trained by solving the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \underbrace{\max\left(0, \left|v_i - f(\mathbf{d}_i; \mathbf{w})\right| - \epsilon\right)}_{\epsilon\text{-insensitive loss function}} \qquad (3)$$

where C is a regularization constant and ϵ controls the training error.[1] The training algorithm finds weights $\mathbf{w}$ that define a function f minimizing the (regularized) empirical risk.

Let h be a function from documents into some vector-space representation $\subseteq \mathbb{R}^d$. In SVR, the function f takes the form:

$$f(\mathbf{d}; \mathbf{w}) \;=\; h(\mathbf{d})^\top \mathbf{w} \;=\; \sum_{i=1}^{N} \alpha_i K(\mathbf{d}, \mathbf{d}_i) \qquad (4)$$

where Equation 4 re-parameterizes f in terms of a kernel function K with "dual" weights α_i. K can

[1] Given the embedding h of documents in $\mathbb{R}^d$, ϵ defines a "slab" (region between two parallel hyperplanes, sometimes called the "ϵ-tube") in $\mathbb{R}^{d+1}$ through which each $\langle h(\mathbf{d}_i), f(\mathbf{d}_i; \mathbf{w})\rangle$ must pass in order to have zero loss.

year	words	documents	words/doc.
1996	5.5M	1,408	3,893
1997	9.3M	2,260	4,132
1998	11.8M	2,462	4,808
1999	14.5M	2,524	5,743
2000	13.4M	2,425	5,541
2001	15.4M	2,596	5,928
2002	22.7M	2,846	7,983
2003	35.3M	3,612	9,780
2004	38.9M	3,559	10,936
2005	41.9M	3,474	12,065
2006	38.8M	3,308	11,736
total	247.7M	26,806	9,240

Table 1: Dimensions of the dataset used in this paper, after filtering and tokenization. The near doubling in average document size during 2002–3 is possibly due to the passage of the Sarbanes-Oxley Act of 2002 in the wake of Enron's accounting scandal (and numerous others).

be seen as a similarity function between two documents. At test time, a new example is compared to a subset of the training examples (those with $\alpha_i \neq 0$); typically with SVR this set is sparse. With the linear kernel, the primal and dual weights relate linearly:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i h(\mathbf{d}_i) \quad (5)$$

The full details of SVR and its implementation are beyond the scope of this paper; interested readers are referred to Schölkopf and Smola (2002). SVM^{light} (Joachims, 1999) is a freely available implementation of SVR training that we used in our experiments.[2]

4 Dataset

In the United States, the Securities Exchange Commission mandates that all publicly-traded corporations produce annual reports known as "Form 10-K." The report typically includes information about the history and organization of the company, equity and subsidiaries, as well as financial information. These reports are available to the public and published on the SEC's web site.[3] The structure of the 10-K is specified in detail in the legislation. We have collected 54,379 reports published over the period 1996–2006 from 10,492 different companies. Each report comes with a date of publication, which is important for tying the text back to the financial variables we seek to predict.

From the perspective of predicting future events, one section of the 10-K is of special interest: Section 7, known as "management's discussion and analysis of financial conditions and results of operations" (MD&A), and in particular Subsection 7A, "quantitative and qualitative disclosures about market risk." Because Section 7 is where the most important forward-looking content is most likely to be found, we filter other sections from the reports. The filtering is done automatically using a short, hand-written Perl script that seeks strings loosely matching the Section 7, 7A, and 8 headers, finds the longest reasonable "Section 7" match (in words) of more than 1,000 whitespace-delineated tokens.

Section 7 typically begins with an introduction like this (from ABC's 1998 Form 10-K, before tokenization for readability; boldface added):

> The following discussion and analysis of ABC's consolidated financial condition and consolidated results of operation should be read in conjunction with ABC's Consolidated Financial Statements and Notes thereto included elsewhere herein. This discussion contains certain **forward-looking statements which involve risks and uncertainties**. ABC's actual results could differ materially from the results expressed in, or implied by, such statements. See "Regarding Forward-Looking Statements."

Not all of the documents downloaded pass the filter at all, and for the present work we have only used documents that do pass the filter. (One reason for the failure of the filter is that many 10-K reports include Section 7 "by reference," so the text is not directly included in the document.)

In addition to the reports, we used the Center for Research in Security Prices (CRSP) US Stocks Database to obtain the price return series along with other firm characteristics.[4] We proceeded to calculate two volatilities for each firm/report observation: the twelve months prior to the report ($v^{(-12)}$) and the twelve months after the report ($v^{(+12)}$).

[2] Available at `http://svmlight.joachims.org`.

[3] `http://www.sec.gov/edgar.shtml`

[4] The text and volatility data are publicly available at `http://www.ark.cs.cmu.edu/10K`.

Tokenization was applied to the text, including punctuation removal, downcasing, collapsing all digit sequences,[5] and heuristic removal of remnant markup. Table 1 gives statistics on the corpora used in this research; this is a subset of the corpus for which there is no missing volatility information. The drastic increase in length during the 2002–2003 period might be explained by the passage by the US Congress of the Sarbanes-Oxley Act of 2002 (and related SEC and exchange rules), which imposed revised standards on reporting practices of publicly-traded companies in the US.

5 Baselines and Evaluation Method

Volatility displays an effect known as autoregressive conditional heteroscedasticity (Engle, 1982). This means that the variance in a stock's return tends to change gradually. Large changes in price are presaged by other changes, and periods of stability tend to continue. Volatility is, generally speaking, not constant, yet prior volatility (e.g., $v^{(-12)}$) is a very good predictor of future volatility (e.g., $v^{(+12)}$). At the granularity of a year, which we consider here because the 10-K reports are annual, there are no existing models of volatility that are widely agreed to be significantly more accurate than our historical volatility baseline. We tested a state-of-the-art model known as GARCH$(1,1)$ (Engle, 1982; Bollerslev, 1986) and found that it was no stronger than our historical volatility baseline on this sample.

Throughout this paper, we will report performance using the mean squared error between the predicted and true log-volatilities:[6]

$$\text{MSE} = \frac{1}{N'} \sum_{i=1}^{N'} (\log(v_i) - \log(\hat{v}_i))^2 \qquad (6)$$

where N' is the size of the test set, given in Table 1.

6 Experiments

In our experiments, we vary h (the function that maps inputs to a vector space) and the subset of the data used for training. We will always report performance over test sets consisting of one year's worth of data (the subcorpora described in Table 1). In this work, we focus on predicting the volatility over the year following the report ($v^{(+12)}$). In all experiments, $\epsilon = 0.1$ and C is set using the default choice of SVM^{light}, which is the inverse of the average of $h(\mathbf{d})^\top h(\mathbf{d})$ over the training data.[7]

6.1 Feature Representation

We first consider how to represent the 10-K reports. We adopt various document representations, all using word features. Let M be the vocabulary size derived from the training data.[8] Let $\text{freq}(x_j; \mathbf{d})$ denote the number of occurrences of the jth word in the vocabulary in document $\mathbf{d}$.

- **TF**: $h_j(\mathbf{d}) = \frac{1}{|\mathbf{d}|}\text{freq}(x_j; \mathbf{d}), \forall j \in \{1, ..., M\}$.
- **TFIDF**: $h_j(\mathbf{d}) = \frac{1}{|\mathbf{d}|}\text{freq}(x_j; \mathbf{d}) \times \log(N/|\{\mathbf{d} : \text{freq}(x_j; \mathbf{d}) > 0\}|)$, where N is the number of documents in the training set. This is the classic "TFIDF" score.
- **LOG1P**: $h_j(\mathbf{d}) = \log(1 + \text{freq}(x_j; \mathbf{d}))$. Rather than normalizing word frequencies as for TF, this score dampens them with a logarithm. We also include a variant of LOG1P where terms are the union of unigrams and bigrams.

Note that each of these preserves sparsity; when $\text{freq}(x_j; \mathbf{d}) = 0$, $h_j(\mathbf{d}) = 0$ in all cases.

For interpretability of results, we use a linear kernel. The usual bias weight b is included. We found it convenient to work in the logarithmic domain for the predicted variable, predicting $\log v$ instead of v, since volatility is always nonnegative. In this setting, the predicted volatility takes the form:

$$\log \hat{v} \;=\; b + \sum_{j=1}^{M} w_j h_j(\mathbf{d}) \qquad (7)$$

Because the goal of this work is to explore how text might be used to predict volatility, we also wish

[5] While numerical information is surely informative about risk, recall that our goal is to find indicators of risk expressed in the text; automatic predictors of risk from numerical data would use financial data streams directly, not text reports.

[6] We work in the log domain because it is standard in finance, due to the dynamic range of actual volatilities; the distribution over $\log v$ across companies tends to have a bell shape.

[7] These values were selected after preliminary and cursory exploration with 1996–2000 as training data and 2001 as the test set. While the effects of ϵ and C were not large, further improvements may be possible with more careful tuning.

[8] Preliminary experiments that filtered common or rare words showed a negligible or deleterious effect on performance.

	features	2001	2002	2003	2004	2005	2006	micro-ave.
history	$v^{(-12)}$ (baseline)	0.1747	0.1600	0.1873	0.1442	0.1365	0.1463	0.1576
	$v^{(-12)}$ (SVR with bias)	0.2433	0.4323	**0.1869**	0.2717	0.3184	5.6778	1.2061
	$v^{(-12)}$ (SVR without bias)	0.2053	0.1653	0.2051	**0.1337**	0.1405	0.1517	0.1655
words	TF	0.2219	0.2571	0.2588	0.2134	0.1850	0.1862	0.2197
	TFIDF	0.2033	0.2118	0.2178	0.1660	0.1544	0.1599	0.1842
	LOG1P	0.2107	0.2214	0.2040	0.1693	0.1581	0.1715	0.1873
	LOG1P, bigrams	0.1968	0.2015	***0.1729**	0.1500	0.1394	0.1532	0.1667
both	TF+	0.1885	0.1616	0.1925	***0.1230**	***0.1272**	***0.1402**	***0.1541**
	TFIDF+	0.1919	0.1618	0.1965	***0.1246**	***0.1276**	***0.1403**	***0.1557**
	LOG1P+	0.1846	0.1764	***0.1671**	***0.1309**	***0.1319**	**0.1458**	***0.1542**
	LOG1P+, bigrams	0.1852	0.1792	***0.1599**	***0.1352**	***0.1307**	**0.1448**	***0.1538**

Table 2: MSE (Eq. 6) of different models on test data predictions. Lower values are better. Boldface denotes improvements over the baseline, and * denotes significance compared to the baseline under a permutation test ($p < 0.05$).

to see whether text adds information beyond what can be predicted using historical volatility alone (the baseline, $v^{(-12)}$). We therefore consider models augmented with an additional feature, defined as $h_{M+1} = \log v^{(-12)}$. Since this is historical information, it is always available when the 10-K report is published. These models are denoted TF+, TFIDF+, and LOG1P+.

The performance of these models, compared to the baseline from Section 5, is shown in Table 2. We used as training examples all reports from the five-year period preceding the test year (so six experiments on six different training and test sets are shown in the figure). We also trained SVR models on the single feature $v^{(-12)}$, with and without bias weights (b in Eq. 7); these are usually worse and never signficantly better than the baseline.

Strikingly, the models that use *only* the text to predict volatility come very close to the historical baseline in some years. That a text-only method (LOG1P with bigrams) for predicting future risk comes within 5% of the error of a strong baseline (2003–6) shows promise for the overall approach. A combined model improves substantially over the baseline in four out of six years (2003–6), and this difference is usually robust to the representation used. Table 3 shows the most strongly weighted terms in each of the text-only LOG1P models (including bigrams). These weights are recovered using the relationship expressed in Eq. 5.

6.2 Training Data Effects

It is well known that more training data tend to improve the performance of a statistical method; however, the standard assumption is that the training data are drawn from the same distribution as the test data. In this work, where we seek to predict the future based on data from past, that assumption is obviously violated. It is therefore an open question whether more data (i.e., looking farther into the past) is helpful for predicting volatility, or whether it is better to use only the most recent data.

Table 4 shows how performance varies when one, two, or five years of historical training data are used, averaged across test years. In most cases, using more training data (from a longer historical period) is helpful, but not always. One interesting trend, not shown in the aggregate statistics of Table 4, is that recency of the training set affected performance much more strongly in earlier train/test splits (2001–3) than later ones (2004–6). This experiment leads us to conclude that temporal changes in financial reporting make training data selection nontrivial. Changes in the macro economy and specific businesses make older reports less relevant for prediction. For example, regulatory changes like Sarbanes-Oxley, variations in the business cycle, and technological innovation like the Internet influence both the volatility and the 10-K text.

6.3 Effects of Sarbanes-Oxley

We noted earlier that the passage of the Sarbanes-Oxley Act of 2002, which sought to reform financial reporting, had a clear effect on the *lengths* of the 10-K reports in our collection. But are the reports more informative? This question is important, because producing reports is costly; we present an empirical argument based on our models that the legis-

1996–2000		1997–2001		1998–2002		1999–2003		2000–2004		2001–2005		high $\hat{v}$
net loss	0.026	year #	0.028	loss	0.023	loss	0.026	loss	0.025	loss	0.026	↑
year #	0.024	net loss	0.023	net loss	0.020	net loss	0.020	net loss	0.017	net loss	0.018	
loss	0.020	expenses	0.020	expenses	0.017	expenses	0.017	year #	0.016	going concern	0.014	
expenses	0.019	loss	0.020	year #	0.015	going concern	0.015	expenses	0.015	expenses	0.014	
covenants	0.017	experienced	0.015	obligations	0.015	year #	0.015	going concern	0.014	a going	0.014	
diluted	0.014	of $#	0.015	financing	0.014	financing	0.014	a going	0.013	personnel	0.013	
convertible	0.014	covenants	0.015	convertible	0.014	a going	0.014	administrative	0.013	financing	0.013	
date	0.014	additional	0.014	additional	0.014	additional	0.013	personnel	0.013	administrative	0.012	
longterm	-0.014	merger agreement	-0.015	unsecured	-0.012	distributions	-0.012	distributions	-0.011	policies	-0.011	
rates	-0.015	dividends	-0.015	earnings	-0.012	annual	-0.012	insurance	-0.011	by the	-0.011	
dividend	-0.015	unsecured	-0.017	distributions	-0.012	dividend	-0.012	critical accounting	-0.012	earnings	-0.011	
unsecured	-0.015	dividend	-0.017	dividends	-0.015	dividends	-0.012	lower interest	-0.012	dividends	-0.012	
merger agreement	-0.017	properties	-0.018	income	-0.016	rates	-0.013	dividends	-0.013	unsecured	-0.012	
properties	-0.018	net income	-0.019	properties	-0.016	properties	-0.015	properties	-0.014	properties	-0.013	
income	-0.021	income	-0.021	net income	-0.019	rate	-0.019	rate	-0.017	rate	-0.014	↓
rate	-0.022	rate	-0.025	rate	-0.022	net income	-0.023	net income	-0.021	net income	-0.018	low $\hat{v}$

Table 3: Most strongly-weighted terms in models learned from various time periods (LOG1P model with unigrams and bigrams). "#" denotes any digit sequence.

features	1	2	5
TF+	0.1509	**0.1450**	0.1541
TFIDF+	0.1512	**0.1455**	0.1557
LOG1P+	0.1621	0.1611	**0.1542**
LOG1P+, bigrams	0.1617	0.1588	**0.1538**

Table 4: MSE of volatility predictions using reports from varying historical windows (1, 2, and 5 years), micro-averaged across six train/test scenarios. Boldface marks best in a row. The historical baseline achieves 0.1576 MSE (see Table 2).

lation has actually been beneficial.

Our experimental results in Section 6.1, in which volatility in the years 2004–2006 was more accurately predicted from the text than in 2001–2002, suggest that the Sarbanes-Oxley Act led to more informative reports. We compared the learned weights (LOG1P+, unigrams) between the six overlapping five-year windows ending in 2000–2005; measured in L_1 distance, these were, in consecutive order, $\langle 52.2,\ 59.9,\ 60.7,\ 55.3,\ 52.3 \rangle$; the biggest differences came between 2001 and 2002 and between 2002 and 2003. (Firms are most likely to have begun compliance with the new law in 2003 or 2004.) The same pattern held when only words appearing in all five models were considered. Variation in the recency/training set size tradeoff (§6.2), particularly during 2002–3, also suggests that there were substantial changes in the reports during that time.

6.4 Qualitative Evaluation

One of the advantages of a linear model is that we can explore what each model discovers about different unigram and bigram terms. Some manually selected examples of terms whose learned weights (w) show interesting variation patterns over time are shown in Figure 1, alongside term frequency patterns, for the text-only LOG1P model (with bigrams). These examples were suggested by experts in finance from terms with weights that were both large and variable (across training sets).

A particularly interesting case, in light of Sarbanes-Oxley, is the term *accounting policies*. Sarbanes-Oxley mandated greater discussion of accounting policy in the 10-K MD&A section. Before 2002 this term indicates high volatility, perhaps due to complicated off-balance sheet transactions or unusual accounting policies. Starting in 2002, explicit mention of accounting policies indi-

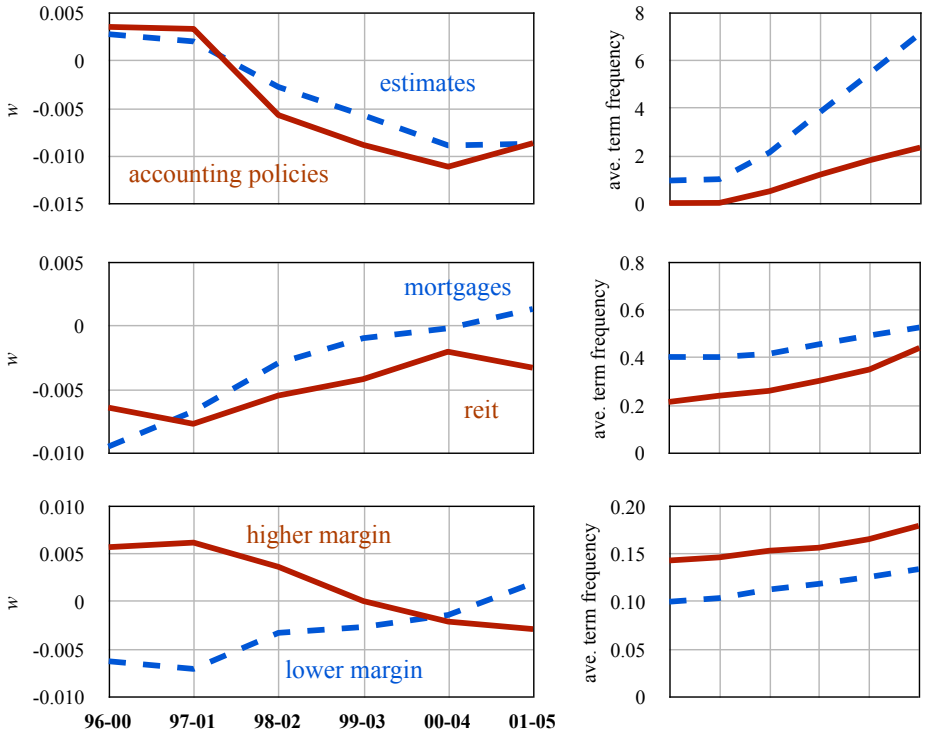

Figure 1: Left: learned weights for selected terms across models trained on data from different time periods (x-axis). These weights are from the LOG1P (unigrams and bigrams) models trained on five-year periods, the same models whose extreme weights are summarized in Tab. 3. Note that all weights are within 0 ± 0.026. Right: the terms' average frequencies (by document) over the same periods.

cates lower volatility. The frequency of the term also increases drastically over the same period, suggesting that the earlier weights may have been inflated. A more striking example is *estimates*, which averages one occurrence per document even in the 1996–2000 period, experiences the same term frequency explosion, and goes through a similar weight change, from strongly indicating high volatility to strongly indicating low volatility.

As a second example, consider the terms *mortgages* and *reit* (Real Estate Investment Trust, a tax designation for businesses that invest in real estate). Given the importance of the housing and mortgage market over the past few years, it is interesting to note that the weight on both of these terms increases over the period from a strong low volatility term to a weak indicator of high volatility. It will be interesting to see how the dramatic decline in housing prices in late 2007, and the fallout created in credit markets in 2008, is reflected in future models.

Finally, notice that *high margin* and *low margin*, whose frequency patterns are fairly flat "switch places," over the sample: first indicating high and low volatility, respectively, then low and high. There is no *a priori* reason to expect high or low margins would be associated with high or low stock volatility. However, this is an interesting example where bigrams are helpful (the word margin by itself is uninformative) and indicates that predicting risk is highly time-dependent.

6.5 Delisting

An interesting but relatively infrequent phenomenon is the **delisting** of a company, i.e., when it ceases to be traded on a particular exchange due to dissolution after bankruptcy, a merger, or violation of exchange rules. The relationship between volatility and delisting has been studied by Merton (1974), among others. Our dataset includes a small number of cases where the volatility figures for the period following the publication of a 10-K report are unavailable because the company was delisted. Learning to predict delisting is extremely difficult because fewer than 4% of the 2001–6 10-K reports precede delisting.

Using the LOG1P representation, we built a linear SVM classifier for each year in 2001–6 (trained on the five preceding years' data) to predict whether a company will be delisted following its 10-K report. Performance for various precision measures is shown in Table 5. Notably, for 2001–4 we achieve

precision (%) at ...	'01	'02	'03	'04	'05	'06
recall = 10%	80	93	79	100	47	21
n = 5	100	100	40	100	60	80
n = 10	80	90	70	90	60	70
n = 100	38	48	53	29	24	20
oracle F_1 (%)	35	42	44	36	31	16

6	bulletin, creditors, dip, otc
5	court
4	chapter, debtors, filing, prepetition
3	bankruptcy
2	concern, confirmation, going, liquidation
1	debtorinpossession, delisted, nasdaq, petition

Table 5: Left: precision of delisting predictions. The "oracle F_1" row shows the maximal F_1 score obtained for any n. Right: Words most strongly predicting delisting of a company. The number is how many of the six years (2001–6) the word is among the ten most strongly weighted. There were no clear patterns across years for words predicting that a company would *not* be delisted. The word *otc* refers to "over-the-counter" trading, a high-risk market.

above 75% precision at 10% recall. Our best (oracle) F_1 scores occur in 2002 and 2003, suggesting again a difference in reports around Sarbanes-Oxley. Table 5 shows words associated with delisting.

7 Related Work

In NLP, regression is not widely used, since most natural language-related data are discrete. Regression methods were pioneered by Yang and Chute (1992) and Yang and Chute (1993) for information retrieval purposes, but the predicted continuous variable was not an end in itself in that work. Blei and McAuliffe (2007) used latent "topic" variables to predict movie reviews and popularity from text. Lavrenko et al. (2000b) and Lavrenko et al. (2000a) modeled influences between text and time series financial data (stock prices) using language models. Farther afield, Albrecht and Hwa (2007) used SVR to train machine translation evaluation metrics to match human evaluation scores and compared techniques using correlation. Regression has also been used to order sentences in extractive summarization (Biadsy et al., 2008).

While much of the information relevant for investors is communicated through text (rather than numbers), only recently is this link explored. Some papers relate news articles to earning forecasts, stock returns, volatility, and volume (Koppel and Shtrimberg, 2004; Tetlock, 2007; Tetlock et al., 2008; Gaa, 2007; Engelberg, 2007). Das and Chen (2001) and Antweiler and Frank (2004) ask whether messages posted on message boards can help explain stock performance, while Li (2005) measures the association between frequency of words associated with risk and subsequent stock returns. Weiss-Hanley and Hoberg (2008) study initial public offering disclosures using word statistics. Many researchers have focused the related problem of predicting sentiment and opinion in text (Pang et al., 2002; Wiebe and Riloff, 2005), sometimes connected to extrinsic values like prediction markets (Lerman et al., 2008).

In contrast to text regression, text *classification* comprises a widely studied set of problems involving the prediction of *categorial* variables related to text. Applications have included the categorization of documents by topic (Joachims, 1998), language (Cavnar and Trenkle, 1994), genre (Karlgren and Cutting, 1994), author (Bosch and Smith, 1998), sentiment (Pang et al., 2002), and desirability (Sahami et al., 1998). Text categorization has served as a test application for nearly every machine learning technique for discrete classification.

8 Conclusion

We have introduced and motivated a new kind of task for NLP: *text regression*, in which text is used to make predictions about measurable phenomena in the real world. We applied the technique to predicting financial volatility from companies' 10-K reports, and found text regression model predictions to correlate with true volatility nearly as well as historical volatility, and a combined model to perform even better. Further, improvements in accuracy and changes in models after the passage of the Sarbanes-Oxley Act suggest that financial reporting reform has had interesting and measurable effects.

Acknowledgments

The authors are grateful to Jamie Callan, Chester Spatt, Anthony Tomasic, Yiming Yang, and Stanley Zin for helpful discussions, and to the anonymous reviewers for useful feedback. This research was supported by grants from the Institute for Quantitative Research in Finanace and from the Center for Analytical Research in Technology at the Tepper School of Business, Carnegie Mellon University.

References

J. S. Albrecht and R. Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proc. of ACL*.

W. Antweiler and M. Z. Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance*, 59:1259–1294.

F. Biadsy, J. Hirschberg, and E. Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proc. of ACL*.

D. M. Blei and J. D. McAuliffe. 2007. Supervised topic models. In *Advances in NIPS 21*.

T. Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

R. Bosch and J. Smith. 1998. Separating hyperplanes and the authorship of the disputed Federalist papers. *American Mathematical Monthly*, 105(7):601–608.

W. B. Cavnar and J. M. Trenkle. 1994. n-gram-based text categorization. In *Proc. of SDAIR*.

S. Das and M. Chen. 2001. Yahoo for Amazon: Extracting market sentiment from stock mesage boards. In *Proc. of Asia Pacific Finance Association Annual Conference*.

H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1997. Support vector regression machines. In *Advances in NIPS 9*.

B. Dumas, A. Kurshev, and R. Uppal. 2007. Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. Swiss Finance Institute Research Paper No. 07-37.

J. Engelberg. 2007. Costly information processing: Evidence from earnings announcements. Working paper, Northwestern University.

R. F. Engle. 1982. Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation. *Econometrica*, 50:987–1008.

E. F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

C. Gaa. 2007. Media coverage, investor inattention, and the market's reaction to news. Working paper, University of British Columbia.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of ECML*.

T. Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

J. Karlgren and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of COLING*.

M. Koppel and I. Shtrimberg. 2004. Good news or bad news? let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. 2000a. Language models for financial news recommendation. In *Proc. of CIKM*.

V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. 2000b. Mining of concurrent text and time series. In *Proc. of KDD*.

K. Lerman, A. Gilder, M. Dredze, and F. Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *COLING*.

F. Li. 2005. Do stock market investors understand the risk sentiment of corporate annual reports? Working Paper, University of Michigan.

R. Merton. 1974. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29:449–470.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP*.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. 1998. A Bayesian approach to filtering junk email. In *Proc. of AAAI Workshop on Learning for Text Categorization*.

B. Schölkopf and A. J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3):1437–1467.

P. C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

K. Weiss-Hanley and G. Hoberg. 2008. Strategic disclosure and the pricing of initial public offerings. Working paper.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing*.

Y. Yang and C. G. Chute. 1992. A linear least squares fit mapping method for information retrieval from natural language texts. In *Proc. of COLING*.

Y. Yang and C. G. Chute. 1993. An application of least squares fit mapping to text information retrieval. In *Proc. of SIGIR*.