Credit Rating Change Modeling Using News and Financial Ratios

HSIN-MIN LU, National Taiwan University
FENG-TSE TSAI, Asia University
HSINCHUN CHEN, University of Arizona
MAO-WEI HUNG and SHU-HSING LI, National Taiwan University

Credit ratings convey credit risk information to participants in financial markets, including investors, issuers, intermediaries, and regulators. Accurate credit rating information plays a crucial role in supporting sound financial decision-making processes. Most previous studies on credit rating modeling are based on accounting and market information. Text data are largely ignored despite the potential benefit of conveying timely information regarding a firm's outlook. To leverage the additional information in news full-text for credit rating prediction, we designed and implemented a news full-text analysis system that provides firm-level coverage, topic, and sentiment variables. The novel topic-specific sentiment variables contain a large fraction of missing values because of uneven news coverage. The missing value problem creates a new challenge for credit rating prediction approaches. We address this issue by developing a missing-tolerant multinomial probit (MT-MNP) model, which imputes missing values based on the Bayesian theoretical framework. Our experiments using seven and a half years of real-world credit ratings and news full-text data show that (1) the overall news coverage can explain future credit rating changes while the aggregated news sentiment cannot; (2) topic-specific news coverage and sentiment have statistically significant impact on future credit rating changes; (3) topic-specific negative sentiment has a more salient impact on future credit rating changes compared to topic-specific positive sentiment; (4) MT-MNP performs better in predicting future credit rating changes compared to support vector machines (SVM). The performance gap as measured by macroaveraging F-measure is small but consistent.

Categories and Subject Descriptors: H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*

General Terms: Management, Economics, Performance

Additional Key Words and Phrases: Credit rating changes, missing-tolerant multinomial probit, news coverage, news sentiment, topic-specific news coverage, topic-specific news sentiment, latent Dirichlet allocation, SVM

ACM Reference Format:
Lu, H.-M., Tsai, F.-T., Chen, H., Hung, M.-W., and Li, S.-H. 2012. Credit rating change modeling using news and financial ratios. ACM Trans. Manage. Inf. Syst. 3, 3, Article 14 (October 2012), 30 pages.
DOI = 10.1145/2361256.2361259 http://doi.acm.org/10.1145/2361256.2361259

This work was supported in part by the following grants: 99-2410-H-002-261 (Taiwan National Science Council), 100-2410-H-002-025-MY3 (Taiwan National Science Council), 100-2420-H-002-028-MY2 (Taiwan National Science Council), HDTRA-09-0058 (Defense Threat Reduction Agency, Basic and Applied Research Directorate), and CNS-0709338 & CBET-0730908 (National Science Foundation).

Author's address: H.-M. Lu, Department of Information Management, National Taiwan University; email: luim@ntu.edu.tw; F.-T. Tsai (corresponding author), Department of Finance, Asia University; email: fttsai@asia.edu.tw; H. Chen, Department of Management Information Systems, University of Arizona; email: hchen@eller.arizona.edu; M.-W. Hung, Department of International Business, National Taiwan University; email: hung@ntu.edu.tw; S.-H. Li, Department of Accounting, National Taiwan University; email: shli@management.ntu.edu.tw.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permission may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2012 ACM 2158-656X/2012/10-ART14 $15.00

DOI 10.1145/2361256.2361259 http://doi.acm.org/10.1145/2361256.2361259

1. INTRODUCTION

Credit risk management is one of the core issues in the banking and insurance industries. Since credit evaluation is a complicated process and requires special knowledge, credit market participants rely heavily on professional credit rating services. Credit ratings are important for both financial and nonfinancial companies because issuer credit ratings determine a company's cost of debt.

In order to manage credit risks, researchers and rating agencies have developed statistical models to measure the creditworthiness of obligors and their obligations. Current credit risk models have mainly adopted two types of input variables [Das et al. 2009]. The first type of variable is accounting numbers published in financial reports. Previous studies often assume that the performance reported in the financial statements can truly reflect worsening credit quality in vulnerable companies (e.g., [Altman 1968; Beaver 1966; Ohlson 1980]). The other type of variable is obtained from financial markets. Examples include stock returns and their volatilities, debt or credit default swap prices, and information from related derivatives (e.g., implied volatilities). The relative performance of market-based and accounting-based prediction models in the literature are still under debate [Agarwal and Taffler 2008; Das et al. 2009; Hillegeist et al. 2004].

Current credit rating studies, nonetheless, ignore the fact that qualitative public information sources such as newspapers may complement two existing types of variables and improve the performance of credit rating models. The necessity of qualitative factors is suggested by Standard & Poor's [2003], study in which they note that "... there are many nonnumeric distinguishing characteristics that determine a company's creditworthiness." Public qualitative information may improve credit rating models for the following reasons. First, news about a firm may provide early warnings or clues about its deteriorating credit situation before accounting numbers in financial statements are communicated to investors. In regard to privately held firms, news is even more valuable because these firms lack market information and they have limited accounting data as well. Second, news sources may provide useful information for firms with illiquid stocks since their market prices can deviate from their true values. News may provide additional information for firms with poor accounting quality caused by earnings management. Third, financial news may play an important role in conveying value-related information to the markets [Chen et al. 2011]. Finally, media reports and exaggerations can influence the beliefs of depositors and lenders, and in extreme cases, may induce depositors or lenders to withdraw their funds and further cause bankruptcy of unhealthy firms or a run on the bank.

Motivated by the deficiency of existing credit rating research, this study aims at utilizing rich information in news articles to create additional predictors for credit rating changes. We designed and implemented a firm-level news analysis system to study whether news coverage, topics, and sentiment can be used to model the changes in rating agencies' future credit rating assignments.

Modeling and predicting credit rating changes using additional news-based variables creates new challenges to the core classification algorithms. When aggregated by firms, those without news coverage have missing sentiment values. Deleting observations with missing values (listwise deletion) may substantially reduce model performance especially when the percentage of missing values is high. To address the missing-value problem, we developed the missing-tolerant multinomial probit (MT-MNP) model, which imputes missing sentiments based on the Bayesian theoretical framework. Our missing imputation approach is integrated into the Gibbs sampling algorithm previously proposed to estimate multinomial probit models [Imai and van Dyk 2005]. Our extension allows the model to explicitly consider the uncertainty caused by the missing values and make use of other nonmissing covariates that come along with the missing values. Our approach is able to leverage all information available to the model and addresses the overoptimistic problem of filling in missing values using predefined values [Greene 2008].

We contribute to the credit rating literature by expanding the information set for credit rating modeling and develop a novel statistical approach that complements the characteristics of the news-based variables. The remainder of this article is organized as follows. In Section 2, we briefly review the studies on the economic impacts of text data in mass media, the credit rating predictors used in previous studies, and the credit rating modeling techniques. In Section 3 we present our research questions, followed by a discussion of the firm-level news analysis system, the MT-MNP approach and evaluation design in Section 4. We summarize the results in Section 5, and we conclude in Section 6.

2. LITERATURE REVIEW

We first review recent studies on the economic impact of text data in mass media, followed by a summary of financial ratios adopted in credit rating modeling literature. The credit rating modeling techniques are then discussed.

2.1 Economic Impacts of Text Data in Mass Media

The economic impact of text data in mass media has drawn attention from researchers in economics, finance, and information systems. Previous studies mainly focused on news coverage and news sentiment. We discuss these two focuses in sequence.

An interesting discovery in recent studies is the economic impact of news coverage. Fang and Peress [2009] documented that firms with no news coverage earn higher returns than firms with high news coverage. The difference is economically and statistically significant after controlling for well-known risk factors. Another similar study was conducted by Chan [2003], who documented a strong price drift after bad news. The news coverage effects often last for months. These empirical results challenge the conventional efficient market belief that financial markets can incorporate new information fairly quickly.

The other research focus is on the economic impacts of news sentiment. Sentiment analysis aims at identifying positive and negative opinions, emotions, and evaluations [Bai 2011; Wiebe et al. 2005]. At market-level, Tetlock [2007] found that the sentiment of a Wall Street Journal (WSJ) column “Abreast of the Market” can be used to predict short-term market returns. News sentiment is also useful in explaining the variation of aggregated private consumption [Uhl 2011]. At firm-level, the news sentiment is shown to predict firms’ future earnings [Tetlock et al. 2008]. Similar results were confirmed by Kothari et al. [2009], who documented a statistically significant relationship between news sentiment and stock return volatility.

Note that text mining modules are required to extract relevant information from the full-text of news articles [Lau et al. 2011]. Two techniques are often adopted to achieve this goal. The first technique adopts dictionaries that contain lists of positive and negative words [Zhang et al. 2012]. This dictionary-based approach is straightforward to implement and can often provide good results [Tetlock et al. 2008]. Both individual words and multiword phrases can be included in a dictionary. One of the most popular dictionaries is the General Inquirer (GI) dictionary.[1] Other dictionaries tailored for financial reports have been proposed as well [Cecchini et al. 2010; Loughran and McDonald 2011]. Subsequent aggregation of dictionary matching results can be conducted at sentence-, paragraph-, or document-level.

[1] http://www.wjh.harvard.edu/~inquirer/

The second approach adopts machine learning techniques to construct classifiers that can be used to predict sentiments of unseen text. Naive Bayesian classifiers [Antweiler and Frank 2004; Das and Chen 2007], support vector machines [Abbasi and Chen 2008; Abbasi et al. 2008], and maximum entropy classifiers [Pang et al. 2002] have been adopted in previous studies. Among others, the OpinionFinder[2] system provides a synthetic approach that refines the sentiment recognition process by considering the context around known sentiment words [Wilson et al. 2005b]. This approach delivers better performance by combining the benefits of a keyword-based approach and machine learning techniques.

A large collection of training examples is a prerequisite for machine learning approaches. Preparing training examples, however, is often time consuming and costly despite the potential benefits of achieving good classification results. One way to address the issue of costly training examples is to adopt unsupervised learning approaches that automatically group words of similar meanings. Based on the latent Dirichlet allocation (LDA) model [Blei et al. 2003], Lin and He [2009] proposed identifying fine-grained sentiment by estimating a joint topic-sentiment model that only requires a general sentiment dictionary as the prior information. Experiments show that the proposed approach can extract coherent sentiment-topic information.

2.2 Inputs for Credit Rating Modeling

Credit rating modeling approaches can be roughly divided into two types according to the input variables. The first type adopts accounting variables as the inputs while the other type adopts market-based variables such as stock price and return volatility. Previous studies found conflicting evidence on the relative performance of these two approaches [Agarwal and Taffler 2008; Das et al. 2009; Hillegeist et al. 2004]. For the purpose of validating the value of text data in credit rating modeling, this study mainly focuses on the accounting-based models, which are widely accepted among researchers and practitioners [Altman and Hotchkiss 2005].

One of the classical accounting-based approaches is Altman's Z-Score model [Altman 1968], which discriminate defaulting firms from healthy ones using a linear combination of five financial ratios. Altman's Z-Score improves on its predecessors that use single financial ratios to predict bankruptcy [Beaver 1966]. Altman's Z-Score model is widely applied in industry because of the availability of financial reports and its effectiveness in identifying failing firms.

Altman's Z-Score uses a cutoff value to determine the classification of a firm and outputs dichotomous scores. Ohlson [1980] addresses this shortcoming by wrapping the linear combination of financial ratios with a logistic transformation. The new approach, Ohlson's O-Score, outputs probability instead of a binary decision. The classification accuracy, however, is similar to that of the Altman's Z-Score.

Merton's distance to default (DTD) [Merton 1974], a market-based model, identifies distressed firms based on observed market variables such as returns, volatility, and the book value of total liabilities. Merton's model assumes that the total value of a firm follows geometric Brownian motion and the equity of the firm is a call option on the underlying value of the firm with a strike price equal to the face value of the firm's debt. The pricing model leads to a functional form of expected default frequency (π_{Merton}) that can be computed based on observed market data. Previous studies show that π_{Merton} can be used to predict firm defaults [Agarwal and Taffler 2008; Bharath and Shumway 2008].

[2] http://www.cs.pitt.edu/mpqa/opinionfinder.html

Table I. List of Financial Ratios and Market-Based Variables

Category	Variable	Financial Ratio
Size	Z1	Asset Accounting Value
	Z2	Total Liabilities
	Z3	Asset Market Value
	Z4	Book to Market Value
Financial leverage	Z5	Long-Term Debts / Total Invested Capital
	Z6	Debt Ratio
	Z7	Debt / Equity
Profitability	Z8	Return on Total Assets
	Z9	Return on Equity
	Z10	Operating Income Before Depreciation / Sales
	Z11	Operating Income / Received Capitals
	Z12	Non-Operating Income / Sales
	Z13	Net Income Before Tax / Received Capitals
	Z14	Net Income Before Tax / Sales
	Z15	Gross Profit Margin
	Z16	Net Profit Margin
	Z17	Earnings Per Share
	Z18	Retained Earnings / Total Assets
Interest coverage	Z19	EBIT Interest Coverage
	Z20	EBIT / Total Debt
Liquidity	Z21	Current Ratio
	Z22	Quick Ratio (Acid Test Ratio)
Market-based	π_{Merton}	Merton's Distance to Default Measure

Treating credit rating as a classifcation problem, recent studies applied machine learning approaches to improve credit rating modeling. SVM has been adopted in credit rating modeling [Huang et al. 2004]. Experiments suggest that SVM provides some level of improvement compared to other baseline approaches [Bellotti et al. 2011].

Table I summarizes important financial ratios considered in the previous studies. We divide these ratios into five categories: size [Ohlson 1980], financial leverage [Altman 1968], profitability [Altman 1968], interest coverage [Huang et al. 2004], liquidity [Ohlson 1980], and market-based according to previous credit rating studies. These variables have been found to be associated with subsequent default and loss experience. The size variables measure a company's book value, market value, and debt capacity. The financial leverages measure the proportion of funds coming from borrowing. The profitability ratios measure different kinds of profits before and after costs relative to a company's assets, equities, or sales. A company's return on sales and return on equity, for example, provide information about profitability. The interest coverage ratios measure earning capacity to cover its interest expense or debt. The liquidity ratios measure the ability of a firm to meet its short-term financial obligations in the future. The Merton's distance to default measures the expected default frequency based on a bond pricing model.

While financial ratios and market valuables that reflect different aspects of firms have been incorporated into credit rating models, few studies have investigated the value of text data in credit rating modeling. It is our intended contribution to better understand how to leverage text data in news articles to address credit rating modeling and prediction problems.

2.3 Credit Rating Modeling Techniques

Credit rating and bankruptcy prediction studies adopt either statistical methods or AI methods. Statistical methods include ordinary least square, discriminant analysis, logit, and probit models [Altman 1968; Beaver 1966; Ohlson 1980]. AI methods cover neural networks, expert systems, decision trees, and support vector machines (e.g., Huang et al. [2004]). Other statistical methods such as the hazard model [Shumway 2001], modified ordered probit model [Hwang et al. 2009], ordered semiparametric probit model [Hwang et al. 2010], random effect ordered probit model [Afonso et al. 2009], and mixed logit model [Jones and Hensher 2004] have also been proposed.

When prediction accuracy is the only concern, AI methods may be a good choice. However, AI methods are often treated as a black box and offer limited model interpretability. For instance, support vector machines are known to deliver superior prediction ability when nonlinear kernels are adopted. These kernel functions transfer input features into higher dimensional spaces and allow better separation between different classes. These transformations, nonetheless, also have a negative impact on our ability to understand how input features contribute to the final classification outcomes.

Statistical methods, on the other hand, often provide good model interpretability. Statistical tests can be used to directly identify the effects of input features. Our study aims at investigating the relationship between credit rating and qualitative information in news articles. The missing values in news-based variables, nonetheless, prevent the direct application of existing qualitative response models. The shortcomings of existing approaches motivates the development of MT-MNP in this study.

3. RESEARCH QUESTIONS

Based on our review, it is clear that text data in news articles are valuable in modeling firm earnings, stock returns, and volatility. Market-level studies also indicate the usefulness of news articles in predicting market returns. Few credit rating modeling studies, however, have investigated the potential value of news articles. Our study aims at bridging this gap by leveraging statistical modeling, text mining, and sentiment analysis techniques to provide novel input variables and classification approaches for credit ratings.

Most previous studies on the economic value of news data have focused on coverage and sentiment. In addition to these two dimensions, we believe that fine-grained sentiment on different topics may also be valuable. These novel topic-specific variables, nonetheless, contain a large portion of missing values because many firms were at most mentioned in a few topics during a given time period. Dealing with missing sentiment of firms without topic-specific news coverage creates another technical challenge.

To support credit rating modeling and prediction using news coverage, topics, and sentiment, our study focuses on the following research questions. First, how to develop rigorous missing imputation approaches that allow the estimation of explanatory power of news variables and conduct out-sample prediction based on covariates that may contain missing values?

After the technical framework is developed, we can investigate the value of news variables. The second research question is the explanatory power of news coverage, sentiment, and topics. Finally, the ability to predict future credit rating changes (upgrade and downgrade) provides a meaningful way to verify the performance of the proposed approach.

4. CREDIT RATING CHANGE MODELING USING NEWS-BASED VARIABLES

To answer these questions, we designed and implemented a system that supports credit rating modeling using text data from news articles. Our system identifies public firms in news articles and extracts topic and sentiment associated with these firms. The firm-level news coverage, sentiment, topic-specific coverage, and topic-specific sentiment variables are then included in credit rating models as additional inputs. To address the missing value problem, we developed the missing-tolerant multinomial probit (MT-MNP) model, which can be used to investigate the explanatory and prediction power of news-based variables.

Our current design computes news-based variables by quarter. Together with financial ratios, the MT-MNP predicts the credit rating changes in the following quarter at the end of current quarter. We selected the quarterly frequency based on the availability of financial ratios. Moving to a higher frequency, such as monthly or weekly, can easily be achieved under the current design. In this section we first introduce the credit rating outcomes, followed by a description of our system.

4.1 Credit Rating Outcomes

Credit rating outcomes are provided by credit rating agencies such as Standard and Poor's (S & P) or Moody's. S & P is one of the largest rating agencies in the world. Our discussion is based on S & P's rating system. Other rating systems are similar to the one explained in the following.

S & P rates borrowers on a scale from AAA (best rating) to D (worst rating). There are 22 levels in total: AAA, AA+, AA, AA−, A+, A, A−, BBB+, BBB, BBB−, BB+, BB, BB−, B+, B, B−, CCC+, CCC, CCC−, CC, C, and D. The ratings of BBB- and above are investment grade classifications while the rest are speculative grade classifications. The rating agency explains that firms with investment grade ratings have the capacity to meet their financial commitments, while firms with speculative grade ratings are less likely to pay back creditors and are not suitable for investment by fiduciary organizations.

The 22 rating levels can be combined into coarser groups when conducting credit rating modeling [Duffie and Singleton 2003]. A more interesting question, however, is to predict future upgrades and downgrades since these events often have significant economic impacts. We define three possible rating changes by the end of the following quarter: −1 (downgrade), 0 (unchanged), and 1 (upgrade). The goal is to predict one of the three outcomes, given news-based variables and financial ratios available in the current quarter.

4.2 System Design

Figure 1 depicts our system design. Our system consists of two major components. The first component conducts text analysis using news articles and outputs news coverage, topic, and sentiment for the second component, which constructs credit rating prediction models via the missing-tolerant multinomial probit (MT-MNP) model and evaluates performance. We summarize these two components in the following.

4.2.1 News Full-Text Analysis. Our first component can be further divided into three modules: firm name extraction, topic clustering, and sentiment analysis. The first module performs named entity recognition[3] [Finkel et al. 2005] and standardizes the recognized company names by consulting the stocknames table in CRSP's monthly stock price dataset. Standardized firm IDs (PERMCO) are then attached to the input

[3] http://nlp.stanford.edu/ner/index.shtml

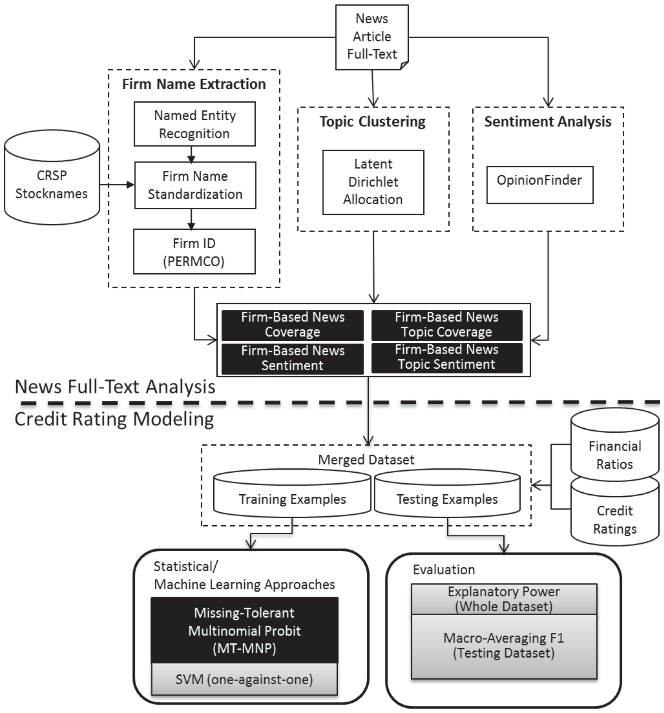


Fig. 1. Design framework for credit rating modeling using news-based variables.

news articles if the firm is mentioned at least once. Several rules that consider the variation caused by abbreviations and acronyms are applied during the process. The stocknames table covers most firms traded in the NYSE, AMEX, and NASDAQ in the past 20 years. Conducting firm name extraction based on the stocknames table allows us to efficiently identify publicly traded firms. The name standardization routine was implemented via Java.

The second module conducts topic clustering using the latent Dirichlet allocation (LDA) model [Blei et al. 2003; McCallum 2002]. The LDA is a generative model that allows the latent topic variables to determine the observed word distribution in a document. LDA is an unsupervised learning algorithm, which requires no costly manual labeling. We adopt the LDA model to provide topical information. Combined with the sentiment analysis results, fine-grained topic-specific sentiment variables can be computed for a subsequent credit rating modeling module.

One prerequisite for the LDA approach is to determine T, the number of topics. We select the value by conducting a grid search (from 1 to 500) and choosing the number that maximizes the model posterior. We used $\beta = 0.1$ and $\alpha = 50/T$ as the prior parameters, following a previous study [Griffiths and Steyvers 2004]. Based on their procedure, we set the number of topics to 83 and estimated a final model. The topic of a document is assigned to the one with the highest posterior probability. We adopted the Mallet[4] toolkit for LDA estimation.

The third module conducts sentence-level sentiment analysis based on the OpinionFinder [Wilson et al. 2005a] system. An input document goes through several major processing steps, such as sentence splitting, part of speech (POS) tagging, stemming,

[4] http://mallet.cs.umass.edu/

and shallow parsing, to extract features for subsequent polarity classification. OpinionFinder takes the context into consideration when tagging the sentiment of words in a sentence. As a result, the polarity classification results are more accurate than with the dictionary-based approach.

One interesting question is whether OpinionFinder considers negations when conducting sentiment tagging. Our observations suggest that the negation of positive adjectives (e.g. “not good”) is tagged as negative while the negation of negative adjectives (e.g. “not bad”) is still tagged as negative. Double negation (e.g. “not as bad as expected”) is not handled by the polarity classifier. To understand the effects of negations in our testbed, we conducted a preliminary study and manually analyzed a random sample of 30 news articles from the WSJ. We found no usage of double negation and the cases of negation were rare. We thus decided to use the output of OpinionFinder directly without performing any postprocessing.

Combining the outputs from the three modules, we obtain the counts of positive and negative sentences in a news article, its topic, and the public firm IDs associated with the news articles. The positive (negative) sentiment score of an article z is the fraction of positive (negative) sentences

$$pos_z = \frac{pos\_sent_z}{total\_sent_z}; \qquad neg_z = \frac{neg\_sent_z}{total\_sent_z},$$

where pos_sent_z (neg_sent_z) are the number of positive (negative) sentences tagged by OpinionFinder, and $total\_sent_z$ is the total number of sentences in article z. To provide a concert example for this process, consider the following news segment published on Feb. 6, 2002:

> “AES Corp.’s profit fell 80% in the fourth quarter, reflecting the power producer’s losses from discontinued operations. ... AES shares fell nearly 50% on Sept. 26 after the company issued a **profit warning** for 2001 results. Since then, the collapse of Enron Corp. and subsequent investor **concerns** about the energy sector and companies with hard-to-understand financial results have further depressed the stock. ... ‘It may be a **prudent** decision to exit the investment’ in Argentina, said Chris Budzynski, analyst at Legg Mason in Baltimore. Since the fall, AES has pledged to explain its complex business more clearly to investors. In a first step yesterday, AES reported its results within four new business lines: growth energy distribution, contract generation, competitive supply and large utilities. ...”

This news article contains 23 sentences. Four negative words tagged by OpinionFinder are marked in bold. Three sentences in which these negative words appeared are marked as negative by our system. The negative sentiment score of this article is $neg_z = 3/23 = 0.13$ and the positive score is $pos_z = 0$. The “AES Corp” in this article was identified as a publicly traded firm with PERMCO=10996. “Legg Mason” was identified by our system but not included in the analysis because it belonged to the financial sector. Note that “Enron Corp” was not identified because it went bankrupt in 2001. Our topic clustering module assign this article to Topic 11 (Earnings (1)). The negative and positive scores are thus associated with AES Corp.

It is interesting to note that the “fell” in the first sentence was not tagged by OpinionFinder. The tagged sentiment words, however, seem to be accurate. The sentiment analysis module clearly did not provide perfect tagging results. By aggregating across large amounts of news articles, the drawback of individual tagging errors can be substantially reduced.

We aggregate the article-level sentiment scores to firm-level by computing the average of positive and negative scores within a topic g

$$pos\_score_{i,t} = \frac{1}{N_{i,t}} \Sigma_{z \in S_{i,t}} pos_z; \quad neg\_score_{i,t} = \frac{1}{N_{i,t}} \Sigma_{z \in S_{i,t}} neg_z,$$

where $S_{i,t}$ is the set of all articles mentioning firm i in quarter t and $N_{i,t}$ is the size of $S_{i,t}$. We also compute the topic-specific sentiment of a firm i in quarter t by averaging article-level scores with respect to these two dimensions

$$pos\_score_{i,t,g} = \frac{1}{N_{i,t,g}} \Sigma_{z \in S_{i,t,g}} pos_z; \quad neg\_score_{i,t,g} = \frac{1}{N_{i,t,g}} \Sigma_{z \in S_{i,t,g}} neg_z,$$

where $S_{i,t,g}$ is the set of articles about topic g and firm i in quarter t.

We note that the sentiment scores $pos\_score_{i,t}$ and $neg\_score_{i,t}$ become missing if the firm i receives no news coverage in period t. The coverage is further diluted when topic-specific sentiment is considered. A firm may have non-missing $pos\_score_{i,t}$ but missing $pos\_score_{i,t,g}$ when a topic g is not discussed in quarter t.

4.2.2 Credit Rating Modeling. As discussed in the preceding, one of our main research questions is to leverage news-based variables for modeling and predicting credit rating changes. There are three possible future rating changes: 1 (upgrade), −1 (downgrade), and 0 (unchanged). Discrete choice models are a natural choice for this type of problem. Our dataset, however, contains missing sentiment values from firms without news coverage. Direct application of standard multinomial probit (MNP) or multinomial logit models is not possible because these models cannot handle missing values properly. One of the commonly used approaches is to conduct listwise deletion (deleting observations that contain missing values). The listwise deletion approaches, however, severely reduce the sample size because topic-specific sentiment variables often contain more than 90% of missing values. Single imputation (replacing missing values with a given constant) leads to overoptimistic estimation results [Greene 2008] and is not suitable to study the effect of news-based variables.

To overcome the problems, we developed a novel missing-tolerant multinomial probit (MT-MNP) model, which extends the existing Gibbs sampling estimation approach by designing the multiple imputation mechanism based on the Bayesian theoretical framework. Our approach overcomes the overoptimistic problem of single imputation by explicitly considering the uncertainty associated with missing values. Other nonmissing values that come along with missing covariates can still contribute to the estimation process.

Our MT-MNP model is derived based on the assumption that the probability of having missing sentiment values is not dependent on the values of the missing sentiment but may depend on other observed values such as firm size and other financial ratios. This is commonly known as the missing at random (MAR) hypothesis [Rubin 1976] in the incomplete data literature.

The Bayesian inference theoretical framework adopted in this study have been proposed to handle the missing value problem in binary logit regression [Ibrahim et al. 2002]. Few studies, however, have discussed the case of more than 2 outcomes. Our MT-MNP model extends the previous studies by integrating the missing imputation procedure into the data augmentation [Meng and van Dyk 1999] estimation method for multinomial probit [Imai and van Dyk 2005]. The resulting MT-MNP model can handle three or more discrete outcomes and leverage all nonmissing input covariates that come along with the missing values.

4.2.3 Missing-Tolerant Multinomial Probit (MT-MNP) Model. We developed the missing-tolerant multinomial probit (MT-MNP) to model the classification problem involving $p \geq 3$ categories. In the problem of modeling credit rating changes, the outcome (Y_i) is -1 (downgrade), 1 (upgrade), and 0 (unchanged). To streamline the discussion in this section, we relabeled the outcome so that 2 indicates downgrade. Upgrade and unchanged cases are still labeled as 1 and 0 respectively. Each outcome variable is associated with a k × 1 vector X_i', containing news-based variables, financial ratios, and a mark-based variable. For each observed Y_i, there is a latent vector $W_i = (W_{i,1}, W_{i,2}, \ldots, W_{i,p-1})$. The latent vector determines the observed outcome via

$$\mathrm{Y}_i(W_i) = \begin{cases} 0, \text{ if } \max(W_i) \leq 0 \\ j, \text{ if } \max(W_i) = W_{i,j>0} \end{cases}, \tag{1}$$

where (W_i) is the largest element of the vector W_i. The latent variable is modeled as

$$W_{i,j} = X_i'\beta_j + e_{i,j},\ j = 1, 2, \ldots, p-1; \quad \begin{pmatrix} e_{i,1} \\ e_{i,2} \\ \vdots \\ e_{l,p-1} \end{pmatrix} \sim N(0, \Sigma), \tag{2}$$

where B_j is a 1×k vector and $e_{t,j}$, $j = 1, 2, \ldots, p-1$, are white noise with mean 0 and covariance Σ, and a p-1 by p-1 matrix. The trace of Σ is normalized to p-1 so that the model is identifiable [Imai and van Dyk 2005]. Note that this setting assumes that all elements in X_i' are available. When some elements of X_i' are missing, we assume that the equation still holds but those missing elements cannot be observed. For the sake of clarity, we will first present the model with no missing values in X_i', and then extend it to the case of missing values.

This setting suggests that the probability of the j-th category (j > 0), given the covariate X_i', is $Prob(Y_i = j|X_i') = P(X_i'B_j + \epsilon_{i,j} > 0 \textit{ and } X_i'B_j + \epsilon_{i,j} \geq X_i'B_k + \epsilon_{i,k} \, \forall \, k \neq j)$. The joint posterior of the latent $W = (W_1, W_2, \ldots, W_n)$ and $B = (B_1, B_2, \ldots, B_{p-1})$, given the data $Y = (Y_1, Y_2, \ldots, Y_n)$ and $X = (X_1, X_2, \ldots, X_n)$; is

$$\begin{aligned} &P(\mathrm{B}, \Sigma, W|Y, X) \propto \\ &\quad \left\{\textstyle\prod_{i=1}^{n} \left[1(Y_i = 0)1(\max(W_1) < 0) + \right.\right. \\ &\quad \left.\left.\textstyle\sum_{j=1}^{p-1} 1(Y_1 = j)1(W_{i,j} = \max(W_i) > 0)\right] MN(W_i; X_i'B, \Sigma)\right\} P(\Sigma)P(B), \end{aligned} \tag{3}$$

where $P(\mathrm{B})$ and $P(\Sigma)$ are the priors of B and Σ. The function $1(d \in A)$ is the indicator function that equals 1 if the random variable d is in the set A (0 otherwise), and $MN(\cdot; X_i'\mathrm{B}, \Sigma)$ is the PDF of a multivariate normal distribution with mean $X_i'\mathrm{B}$, a (p-1) by 1 vector, and a (p-1) by (p-1) covariance matrix Σ.

The summation of the multiplications of indicator functions in Equation (3) deserves a more detailed explanation. Recall that Equation (1) defines the connection between Y_i and W_i. If there is no such connection, then the distribution of W_i, given $X_i'\mathrm{B}$ and Σ, is just a multivariate normal distribution. However, since $Y_i' = 0$ if and only if $\max(W_i) < 0$, and $Y_i = j$ if and only if $W_{ij} = \max(W_i) > 0$ for $j = 1, 2, \ldots, p-1$, we know that regions not satisfying these conditions on the (Y_i, W_i) plane have a zero probability. The summation involving the indication functions defines this connection formally.

As proposed by previous studies [Albert and Chib 1993; Imai and van Dyk 2005], sampling parameters and latent variables can be achieved by Gibbs sampling [Geman and Geman 1984]. Gibbs sampling is an iterated simulation algorithm that repeatedly draws random variables from the conditional posterior distributions of parameters (B and Σ) and latent variables W. Following Algorithm 1 of Imai and van Dyk [2005], we augment the original model with its twin sister, which corresponds to the setting with no constraint on the trace of Σ. This twin sister model can be written as

$$\widetilde{W_{i,j}} \equiv \alpha W_{i,j} = X_i'(\alpha \mathrm{B}_j) + \alpha e_{i,j} \equiv X_i'\widetilde{B}_j + \widetilde{e_{i,j}}, \tag{4}$$

where α is a positive scalar. Note that this model is equivalent to the original model because $Y(W_{i,j}) = Y(\widetilde{W_{i,J}})$. By augmenting parameter α, the estimation procedure is able to handle an improper prior and increase the speed of convergence [Imai and van Dyk 2005].

If there are missing values in X_i', Equation (3) needs to be adjusted to incorporate this effect. Let $x_{i,obs}$ and $x_{i,miss}$ denote all nonmissing and missing elements in X_i'. Each $x_{i,miss}$ is a $q_i \times 1$ vector. Note that X_i' and X_j' $(j \neq i)$ may not have the same missing elements. We denote the collection of all $x_{i,miss}$ and $x_{i,obs}$ as X_{miss} and Y_{obs}, respectively. The joint posterior of parameters and latent variables is:

$$\begin{aligned} &P(\beta, \Sigma, W|Y, X_{obs}) \\ &\propto \left\{ \sum_{i-1}^{n} \left[1(Y_i = 0)1(\max(W_i) < 0) + \sum_{j-1}^{p-1} 1(Y_i = j)1\left(W_{i,j} = \max(W_i) > 0\right) \right] \right. \\ &\left. \int_{x_{i,miss}} MN(W_i; X_i'\beta, \Sigma) P\left(x_{i,miss}|x_{i,obs}\right) dx_{i,miss} \right\} P(\Sigma)P(\beta). \end{aligned} \tag{5}$$

Compared to a posterior without a missing value (Equation (3)), $MN(\cdot)$ (the PDF of the multivariate normal distribution) is now part of a larger integral that deals with the effect of missing values in X_i'. The following section summarizes the Gibbs sampling approach proposed to estimate the joint posterior of Equation (5).

4.2.4 Gibbs Sampling for MT-MNP. The Gibbs sampling algorithm iterates through a sequence of conditional posterior to approximate the joint posterior distribution defined in Equation (5). Let $\theta^{(t-1)} = \left(\beta^{(t-1)}, \Sigma^{(t-1)}, W^{(t-1)}, X_{miss}^{(t-1)}\right)$ be the values of the parameters and latent variable at iteration t-1. At the iteration t, the values are updated by

(1) Draw $W^{(t)}$ from $P\left(W\middle|Y, X_{obs}, \beta^{(t-1)}, \Sigma^{(t-1)}, X_{miss}^{(t-1)}\right)$.
(2) Draw $\beta^{(t)}$ from $P\left(\beta\middle|Y, X_{obs}, \Sigma^{(t-1)}, W^{(t)}, X_{miss}^{(t-1)}\right)$.
(3) Draw $\Sigma^{(t)}$ from $P\left(\Sigma\middle|Y, X_{obs}, \beta^{(t)}, W^{(t)}, X_{miss}^{(t-1)}\right)$.
(4) Draw $X_{miss}^{(t)}$ from $P\left(X_{miss}\middle|Y, X_{obs}, \beta^{(t)}, \Sigma^{(t)}, W^{(t)}\right)$.

It has been shown that this process forms a Markov chain that converges to the joint posterior distribution $P(\beta, \Sigma, W, X_{miss}|Y, X_{obs})$ [Tanner 1996]. Model estimation and prediction can be achieved based on the collection of random variables accumulated through this process. We first present the 4 steps in detail, followed by a discussion on related estimation and prediction issues.

Before getting into the details of the conditional distributions in steps 1–4, it is important to note that the posteriors in step 1, $P\left(W\middle|Y, X_{obs}, \beta^{(t-1)}, \Sigma^{(t-1)}, X_{miss}^{(t-1)}\right)$, depend on both the observed covariates X_{obs} and the imputed missing values $X_{miss}^{(t-1)}$. With these two sets of variables combined, the conditional posterior in step 1 is the same as the model without missing values. Similar observations can be made for the posteriors in steps 2 and 3. It suggests that the four steps can be divided into two groups. The first group consists of steps 1, 2, and 3, which draws W, β, and Σ conditional on the observed and imputed missing values. The missing values are then updated based on the most recent draws of W, β, and Σ. We adopted the procedure proposed by Imai and van Dyk [2005] for the first three steps. We extended the existing approach by adding step 4, which handles missing values. These 4 steps are discussed in the following. The online appendix provides more technical details on these steps.

Recall that for the i-th observation, the latent $W_i = \left(W_{i,1}, W_{i,2}, \ldots, W_{i,p-1}\right)$ determines the observed Y_i. When drawing W_i conditional on Y_i, and other parameters (including the imputed missing valuables), the linkage between W_i and Y_i is preserved by truncating drawings of W_i, which is inconsistent with the observed Y_i, according to Equation (1). One way to achieve this is to draw $W_{i,j}|W_{i,-j}$ from a truncated normal distribution, where $W_{i,-j}$ is the W_i vector with $W_{i,j}$ removed and the dot $(\cdot)$ indicates other relevant parameters and latent variables. The truncation threshold depends on the value of Y_i. If $Y_i = j$, then the posterior is truncated below at $\max(W_{i,-j}, 0)$. If $Y_i \neq j$, the posterior is truncated above at $\max(W_{i,-j}, 0)$. The updated values are referred to as $W_i^{(t)}$.

Before moving to step 2, we need to draw the working parameter α. This parameter creates a mapping to the twin sister model, as defined in Equation (4). It should be noted that α is not identifiable. The reason to incorporate this parameter is to improve the rate of convergence [Imai and van Dyk 2005; Meng and van Dyk 1999]. The working parameter α^2 is the normalization coefficient of the covariance matrix and it follows an inverse Chi-square distribution: $\alpha^2|\Sigma \sim \mathit{trace}\left(\Sigma^{-1}\right)/\chi^2_{v(p-1)}$, where v is set to p-1.

We divide β into p-1 chunks, $\beta_1, \beta_2, \ldots, \beta_{p-1}$, and update them in sequence. It should be noted that, conditioned on other parameters, these p-1 chunks of variables are independent of each other if the off-diagonal terms in Σ are zero. The intuition is that when different outcomes are not correlated with each other, then $\beta_i, i = 1, 2, \ldots, p-1$, are independent of each other conditional on other relevant parameters. If this is the case, then the posterior of each β_i can be derived following the classical regression model. However, since the off-diagonal terms in Σ are not zero in general, adjustments are required if we update β in p-1 chunks. The adjustment can be derived based on the conditional distribution of multivariate normal distributions. The posterior of β_i also follows a multivariate normal distribution. Following Algorithm 1 of Imai and van Dyk [2005], β_i is drawn from the twin sister model and reverts back to the original model.

In step 3, we first draw the covariance matrix of the twin sister model from an inverse Wishart distribution and normalize it so that the trace of the covariance matrix equals p-1. The same constant is used to adjust $W_i^{(t)}$.

One of our contributions is the addition of a missing imputation procedure that is integrated into the existing Gibbs sampling algorithm. To apply the procedure, the location of missing values needs to be recorded and filled with initial values before running the Gibbs sampling algorithm. Steps 1 to 3 then can be executed as if there are no missing values. At this step, the missing values are updated by drawing from their posteriors, given current parameters $\beta^{(t)}$, $\Sigma^{(t)}$, $W_{ij}^{(t)}$. Our procedure draws missing covariates column-by-column. That is, when drawing the missing values of a covariate, other missing covariates are filled with the latest drawings from their posteriors. To derive the posterior of missing covariates, we rewrite the latent equation as $W_{i,j} = x'_{ij}\beta_j + e_{i,j} \equiv q'_{ij}\alpha_j + z_j\gamma_j + e_{ij}$, where $z_i = x_{i\alpha}$ is a covariate with missing values and $q'_{ij} = (x_{i1}, x_{i2}, \ldots, x_{i,\alpha-1}, x_{i,\alpha+1}, \ldots, x_{ik})$, $\gamma_j = \beta_{j\alpha}$ is the slope of the missing covariate, $\alpha_j = \left(\beta_{j1}, \beta_{j2}, \ldots, \beta_{j,\alpha-1}, \beta_{j,(\alpha+1)}, \ldots, \beta_{jk}\right)'$ are the remaining coefficients in β_j. The conditional posterior of the missing value is

$$p(z_i|W_i, \beta, q_{ij}, \Sigma) \propto p(W_i|z_i, q_{ij}, \beta, \Sigma)p(z_i|q_{ij}, \beta, \Sigma).$$

The first term on the right-hand side is the multivariate normal PDF, $MN\left(W_i; q'_{ij}\alpha_j + z_i\gamma_j, \Sigma\right)$. We assume that all covariates follow a multivariate normal distribution. As a result, $p\left(z_i|q_{ij}, \beta, \Sigma\right)$ follows a univariate normal distribution. The mean $\bar{\mu}_i^{(t)}$ and variance $1/\bar{s}_i^{2(t)}$ of $p\left(z_i|q_{ij}, \beta, \Sigma\right)$ can be computed following the standard regression model by treating z_i as the dependent variable and q_{ij} as the independent variable and use all nonmissing z_i to estimate $z_i = q_{ij}\gamma_i + \epsilon_{zi}$. The estimated slope $\gamma_i^{(t)}$ and $var(\epsilon_{zi}) = s_i^{2(t)}$ can then be used to compute $\bar{\mu}_i^{(t)} = q_{ij}\gamma_i^{(t)}$ and $1/\bar{s}_i^{2(t)} = s_i^{2(t)} + q'_{ij} Var\left(r_i^{(t)}\right) q_{ij}$. Combining these two distributions using the standard Bayesian update formula, we have $z_i| \cdot \sim N\left(\frac{i'\Omega^{-1}\mu_i + \bar{\mu}_i\bar{s}_i^{-2}}{i'\Omega^{-1}i + \bar{s}_i^2}, \frac{1}{i'\Omega^{-1}i + \bar{s}_i^2}\right)$, where i is a vector of ones, and

$$\Omega = \begin{bmatrix} \frac{\sigma_{11}}{\gamma_1^2} & \frac{\sigma_{12}}{\gamma_1\gamma_2} & \cdots & \sigma_1, p_1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{\sigma_{p-1,1}}{\gamma_{p-1}\gamma_1} & \cdots & \cdots & \frac{\sigma_{p-1,p-1}}{\gamma_{p-1}^2} \end{bmatrix}, \mu_i = \begin{bmatrix} \frac{(W_{i1} - q'_{i1}\alpha_1)}{\gamma_1} \\ \vdots \\ \frac{(W_{i,p-1} - q'_{i,p-1}\alpha_{p-1})}{\gamma_{p-1}} \end{bmatrix}.$$

The parameters and latent variables $\theta^{(t)} = \left(\beta^{(t)}, \Sigma^{(t)}, W^{(t)}, X_{miss}^{(t)}\right), \gamma_i^{(t)}$, and $s_i^{2(t)}$ are then recorded. The process repeats for a predefined number of iterations. We implemented the MT-MNP estimation and prediction routines in R[5].

4.2.5 Rating Prediction. Rating prediction is straightforward, given the collection of parameters and latent variables from the Gibbs samplings approach. For a data point with covariate g'_b from the testing dataset, the probability that this data point belongs to class c, $P\left(y_b = c|g'_b\right)$, can be computed by repeatedly applying Equations (2) and (1) using $\theta^{(t)}$, $\gamma_i^{(t)}$, and $s_i^{2(t)}$ from the previously recorded dataset. Specifically, let M denote the total number of Gibbs sampling iterations minus the burn-in runs (to reduce the effect of initial values), for t = 1, 2, ..., M. We compute the prediction by first filling in the missing values in g'_b using corresponding $\gamma_i^{(t)}$ and $s_i^{2(t)}$. If there is more than one missing value in g'_b, each missing value is first filled with the mean value of the corresponding column, then the corresponding $\gamma_i^{(t)}$ and $s_i^{2(t)}$ are used to update the missing values in sequence. The process is repeated for a small number of iterations. The missing-free g'_b and $\theta^{(t)}$ can then be used to compute the W_b vector using Equation (2). The predicted outcome $y_b^{(t)}$ then is determined according to Equation (1).

This process is repeated for t = 1, 2,... M, and each $y_b^{(t)}$ is recorded. The posterior probability $P(y_b = c|g'_b)$ is estimated by the distribution of $\{y_b^{(t)}\}$. The outcome with the highest posterior probability is our prediction.

4.2.6 SVM for Credit Rating. We adopted the support vector machine (SVM) classifier [Joachims 1999; Vapnik 1995] as a baseline approach. Previous studies have shown that SVM achieved good performance on credit rating [Huang et al. 2004] as well as

[5] http://www.r-project.org/

international bank ratings [Bellotti et al. 2011] predictions. A salient characteristic of the SVM model is that it maximizes the margin between two classes when choosing the decision hyperplane. An SVM model can be written as

$$\min_{S,S_0} \tfrac{1}{2}||S||^2_{L2} + C\textstyle\sum_{i=1}^{N}\xi_i$$
$$\text{subject to } \xi_i > 0, Y_i(X_i \cdot S + S_0) \geq 1 - \xi_i \quad \forall i,$$

where dot $(\cdot)$ is the inner product operator; S_0 and S are the intercept and slope of the decision hyperplane. In a binary classification setting, we assign the outcome variable $Y_i = 1$ or -1 for a positive or a negative case. The constraints say that given the decision hyperplane, each training example needs to reside at the correct side of the hyperplane and the distance to the hyperplane needs to be at least $1/||S||_{L2}$. If the condition is not satisfied, a penalty of ξ_i will be charged. The training process searches for the hyperplane that minimizes the total penalty of training examples and the L2 norm of the normal vector S. Note that minimizing the L2 norm of the normal vector S is equivalent to maximizing the margin between two classes, which is $2/||S||_{L2}$.

Since we have more than 2 outcomes for credit rating changes, several binary classification models need to be combined to generate multiclass outcomes. We adopted the one-against-one approach following a previous credit rating study [Huang et al. 2004]. For predicting credit rating changes, we trained 3 classifiers for discriminating between -1 (downgrade) and 0 (unchanged); 0 (unchanged) and 1 (upgrade); and -1 (downgrade) and 1 (upgrade). For a test case, each of the three classifiers votes on the final outcomes. For example, if the classification results of the three classifiers are unchanged, unchanged, and downgrade, then the final output is unchanged. Ties are broken randomly.

We filled missing sentiment values with zero, based on the assumption that a missing value provides information neutral to the credit rating prediction. We adopted the linear kernel in this study based on preliminary experiments on our dataset. The cost parameter C is determined by conducting a grid search on a separate tuning dataset with respect to the performance measure. We used a popular implementation, LibSVM [Chang and Lin 2011], to solve the optimization problem.

4.3 Performance Evaluation

To understand the value of news coverage and news sentiment in helping credit rating modeling, we consider both the explanatory power and prediction performance in our experiments. In this context, the explanatory power of news-based variables, including overall coverage, overall sentiment, topic-specific coverage, and topic-specific sentiment, can be investigated by looking at the confidence intervals associated with a conventional confidence level. For a given coefficient, its $(1 - \alpha)$ confidence interval is simply the $\frac{\alpha}{2}$ percentile and $1 - \frac{\alpha}{2}$ percentile values from the outcomes accumulated by the Gibbs sampling algorithm. The explanatory power was estimated using the whole dataset.

To emphasize the importance of predicting rating changes, we adopted the macroaverage F-measure (see, e.g., Manning et al. [2009]) as our performance measure. The macroaveraging F-measure computes a simple average of the F-measures of the three classes. It gives equal weights for the minority class (upgrade, downgrade) and the majority class (unchanged). The importance of rating changes can thus be better represented compared to other measures such as accuracy.

Specifically, we define the macroaveraging F-measure as: $F_{MAC} \equiv 1/3[F(Upgrade) + F(Downgrade) + F(Unchanged)]$. The F-measure for each outcome is: $F(C) = 2\frac{Precision_C \times Recall_C}{Precision_C + Recall_C}, C \in \{Upgrade, Downgrade, Unchanged\}$. Note that $Precision_C$ may be undefined if the classifier makes no prediction for class C. This will lead to undefined

Table II. Industry Distribution by the First Digit of the SIC Code

Leading SIC Digit	Num. of Firms-Quarters
0 (Agriculture, Forestry and Fishing)	157
1 (Mining and Contruction)	2,532
2 (Manufacturing; Consumer Goods)	6,304
3 (Manufacturing; Machinery and Equipment)	6,950
4 (Transportation and Communications)	6,303
5 (Wholesale and Retail)	2,800
7 (Business Services)	2,574
8 (Health and Education Services)	911
9 (Public Administration)	212

$F(C)$ and F_{MAC}. We avoid the problem by defining $Precision_C = 0$ when a classifier predicts no class C. This definition is reasonable because 0 is the lowest possible precision. The corresponding $F(C)$ is also 0 by definition. We define $F(C) = 0$ if both P$recision_C = 0$ and R$ecall_C = 0$. This can be justified by fixing precision at zero and taking recall to approach 0 from above. The limit is 0 because the numerator is always 0 and the denominator is always positive.

5. EXPERIMENTAL STUDIES

In this section, we present the experimental results to answer the research questions investigated in this study. We first summarize our research testbed, followed by the estimation results and prediction performance.

5.1 Research Testbed

We created the research testbed by combining data from several different sources. The credit rating values are the Standard & Poor's long-term issuer credit ratings obtained from Compustat. Accounting variables are also downloaded from the Compustat database. Quarterly accounting variables are used to calculate 22 financial ratios adopted in this study. The stock returns data required to compute Merton's distance to default (π_{Merton}) are obtained from the Centers for Research in Security Prices (CRSP).

We adopted the Wall Street Journal (WSJ), a popular business newspaper in the U.S., as the representative source of financial news. The WSJ was one of the highly circulated newspapers in our sample period according to the Audit Bureau of Circulations. Our collection consists of 278,824 WSJ news articles published from October 1999 to March 2007.

Our news analysis system processed the collection of WSJ news articles and computed the overall news coverage, overall sentiment, topic-specific coverage, and topic-specific sentiment by quarter. The news-based variables are then merged with the financial ratios. Since there is a lag between the end of a quarter and the announcement of quarterly financial reports, we check the announcement date of each quarterly report and only include the most recently available financial ratios in a given quarter. To ensure that our results are not driven by outliers, we Winsorize covariates at 1% on tails when necessary.

The financial service companies with an SIC code between 6000–6999 were excluded since firms in the financial sector were subject to different regulations and accounting conventions. While firms traded in NYSE, AMEX, and NASDAQ were all included in our initial data collection effort, firms need to have nonmissing book value, market value, and sales, in order to be included in our testbed. Table II lists the number of firm-quarters by industry (as determined by the first digit of the SIC code).

Table III. Credit Rating Changes by Quarter

Year-Quarter	Num. of Firm	Upgrade[†]	Downgrade[†]	Unchanged[†]
1999-4	924	2.2%	5.4%	92.4%
2000-1	999	3.1%	5.3%	91.6%
2000-2	955	2.0%	5.2%	92.8%
2000-3	976	2.9%	7.8%	89.3%
2000-4	962	3.1%	6.5%	90.3%
2001-1	1008	2.4%	6.7%	90.9%
2001-2	941	1.8%	8.2%	90.0%
2001-3	956	1.7%	8.1%	90.3%
2001-4	940	2.0%	9.0%	88.9%
2002-1	1022	1.5%	6.0%	92.6%
2002-2	961	1.1%	6.1%	92.7%
2002-3	958	1.7%	9.1%	89.2%
2002-4	935	1.3%	6.4%	92.3%
2003-1	1007	3.6%	7.1%	89.4%
2003-2	940	2.6%	4.6%	92.9%
2003-3	964	2.9%	4.5%	92.6%
2003-4	962	2.4%	3.4%	94.2%
2004-1	1014	2.9%	2.2%	95.0%
2004-2	960	1.6%	2.3%	96.1%
2004-3	956	2.7%	3.3%	93.9%
2004-4	929	3.8%	2.9%	93.3%
2005-1	974	3.8%	4.4%	91.8%
2005-2	938	3.4%	4.1%	92.5%
2005-3	923	2.2%	4.2%	93.6%
2005-4	928	2.6%	4.0%	93.4%
2006-1	965	3.1%	3.6%	93.3%
2006-2	947	2.3%	4.0%	93.7%
2006-3	930	2.6%	4.4%	93.0%
2006-4	923	3.3%	4.1%	92.6%
2007-1	946	5.5%	3.3%	91.2%
Average	958.1	2.6%	5.2%	92.2%

[†] Credit rating change by the end of the next quarter.

The dependent variable at a given quarter is how the credit rating value changes by the end of the next quarter. The outcome is 1, 0, and −1 for upgrade, unchanged, and downgrade, respectively. Our sample consists of 28,743 observations, spanning 30 quarters. Table III summarizes the distribution of future rating changes by quarter. On average, our testbed contains 958.1 firms each quarter. Only a small portion of credit ratings changed in the following quarter: 2.6% upgrades and 5.2% downgrades. Most ratings (92.2%) remain unchanged.

5.1.1 Topic Clustering. We set the number of clusters to 83 according to a grid search that maximized the testing posterior likelihood of the LDA model. Based on the top 15 keywords selected by LDA, we assigned a name for each topic. Table IV lists 20 LDA clustering topics with the highest coverage (defined below). A complete list of all 83 topics can be found in online Appendix B. Earnings related topic (Earnings (1)) have

Table IV. Topic Clustering Results

ID	Topic	Included	Coverage[†]	Prevalence[‡]	Keywords
22	Earnings (3)	V	25.7%	0.81%	stock stocks cents company nasdaq quarter market earnings shares index fell rose share year dow small analysts investors technology gained
11	Earnings (1)	V	15.0%	1.37%	million quarter share year company cents earnings net earlier revenue billion sales loss income profit rose analysts stock results fourth
36	Earnings (4)	V	12.6%	1.05%	million year quarter share company cents billion net earlier earnings revenue sales profit rose income loss reported results fourth compared fell
82	Stock Market Update (3)	V	10.6%	0.55%	index stock stocks market shares rose fell points investors prices trading dow markets year jones issues yen dollar day volume
31	Investment (2)	V	10.6%	1.09%	mr stock investors market companies year fund investment funds money company capital stocks shares firm price street years wall management
4	Equity Market (1)	V	10.3%	2.67%	shares stock company million trading corp jones dow share exchange market common york offering nasdaq cents companies price based symbol
35	Corp. Governance (2)	V	8.4%	1.58%	mr company business executive years chief board executives year management time people chairman job employees corporate president top firm companies
52	Corp. Governance (3)	V	8.0%	0.64%	president company mr chief executive vice officer named chairman years corp director board financial senior operating succeeds group unit post
55	Stock Market Update (2)	V	7.4%	1.32%	market stock stocks investors year index prices trading bond interest fell rose short rates dollar points average york shares price
19	Deal	V	6.6%	1.28%	billion company deal euros european group companies stake merger market bid french europe mr france shares sa million offer shareholders
17	Info. Technology (1)		5.9%	1.54%	company computer technology corp companies software market business intel systems chip computers ibm products chips calif year data mr million
57	Sales (2)		5.9%	0.96%	sales million year company stores quarter share cents store earnings earlier rose mart retailer billion wal net fiscal analysts retail
16	Advertisement		5.8%	1.40%	company advertising ad tobacco agency marketing ads million group year products brands coke brand cola industry market sales coca business
13	Earnings (2)		5.6%	1.18%	million company year quarter share sales revenue computer cents ibm billion earnings market corp analysts business net earlier mr profit
78	Fixed Income (2)		5.6%	1.49%	million bonds priced due basis points yield issue price notes moody coupon bank securities lead terms fees maturity spread par
73	Banking Industry (2)		5.4%	1.99%	bank billion banks financial investment credit debt capital investors firm morgan market deal business equity companies mr securities group street
23	Internet		5.4%	1.25%	microsoft software internet computer web company online companies technology apple corp site windows users mr system business computers yahoo google
66	Economy (2)		5.4%	0.53%	mr companies year market years business growth economy prices industry time price big economic people high costs past long money
8	Ecommerce		5.3%	1.36%	card information credit service web online site cards mail customers internet company companies data services business consumers security article people
14	Energy (1)		5.3%	1.54%	power energy plant water environmental company steel utility plants nuclear waste million utilities electric gas state industry electricity year years

[†] Number of news articles in a topic / total number of news articles.
[‡] The portion of firm-quarters that were mentioned in at least one news articles from the topic.

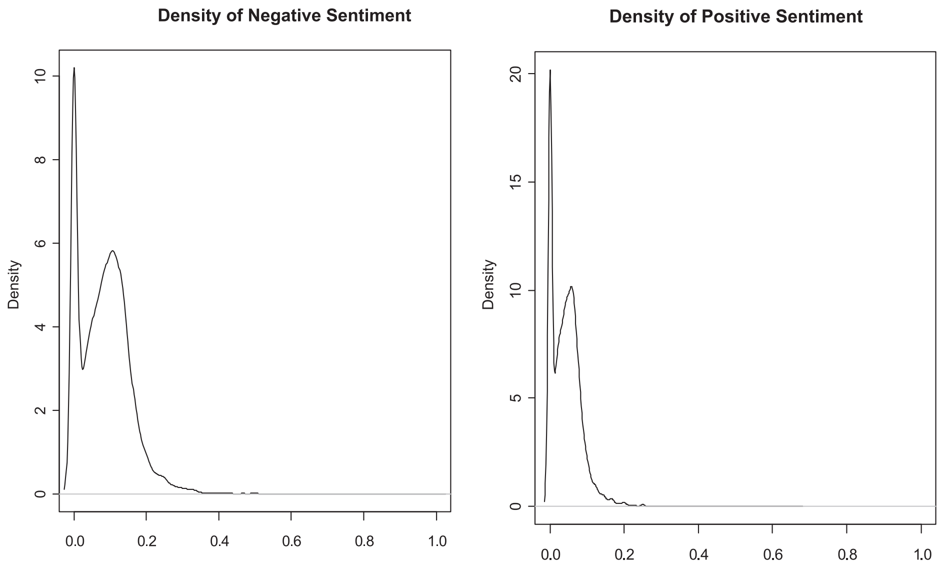


Fig. 2. Kernel density of negative and positive sentiment.

the highest prevalence (2.81%; number of articles in a topic divided by total number of articles) in our testbed. Some of the non-business related topics such as Culture (1) and Book also have relatively high prevalence (1.99% and 1.76%). The coverage of a news topic is the portion of firm-quarters that were mentioned in at least one news article from the topic. Some high prevalence topics such as Book usually do not talk about firm-specific events and thus contribute less to the coverage. Other business-related topics such as investments and corporate governance can be linked to publicly traded firms and thus contribute more to the news coverage.

It is clear from Table IV that the 20th highest coverage (Energy (1)) is 5.3%. The sentiment variables of this topic contain 94.7% (100%-5.3%) of missing values. The marginal contribution of news-based variables from this topic, as a result, is quite low. To mitigate the sparseness problem, we only include news topics with at least 6% of coverage. Ten news topics meet this criterion.

5.1.2 Sentiment Analysis. We computed the overall sentiment values ($pos\_score_{i,t}$ and $neg\_score_{i,t}$) for each firm-quarter when the news data is available. There are 73.81% of firm-quarters containing nonmissing overall sentiment values, which are much higher compared to the coverage of topic-specific variables. Figure 2 plots the kernel density [Sheather and Jones 1991] of negative sentiment ($pos\_score_{i,t}$ and positive sentiment ($neg\_score_{i,t}$). Kernel density is a nonparametric method for estimating the density function of a random variable. It is similar to a histogram but often provides more detailed information.

The kernel densities in Figure 2 show that both negative sentiment and positive sentiment are bimodal. There is a peak close to zero, representing the cluster of firm-quarters that have news coverage but close to neutral sentiment. The other peak shows the clusters with more salient sentiment information. We note that both the negative sentiment and positive sentiment have a value range of zero to one. The negative sentiment, nonetheless, spreads wider compared to the positive sentiment. Both random variables contain enough variation so that they can potentially be used as additional covariates for credit rating modeling.

5.1.3 Summary Statistics of Key Variables. The mean of dependent variable ($y_{i,t}$) is slightly negative (−0.026), consistent with the observation that there are more downgrades than upgrades. We take the logarithm of the asset accounting value (z1), total liabilities (z2), asset market value (z3), and book-to-market value (z4), to correct the skewed distribution. The Merton distance to default (π_{Merton}) has an average of 7.088% and a median of 0%, suggesting that the expected default probability of most firms is close to zero.

The mean of news coverage ($article_{i,t}$) is 7.518, with a standard deviation about 2 times larger (17.119). Both inflated zeros (26% of zero observations) and firm-quarters with high news coverage (maximum = 119) contribute to the large standard deviation. The mean and standard deviation of overall negative sentiment ($neg\_score_{i,t}$) are larger than that of the overall positive sentiment ($pos\_score_{i,t}$), consistent with the observation that the distribution of overall negative sentiment is more dispersed compared to overall positive sentiment. Similar patterns can also be found in topic-specific sentiment variables. A complete list of summary statistics can be found in online Appendix C.

Correlation coefficients of the 22 financial ratios (not shown) indicate that some of the financial ratios are highly correlated. The variance inflation factors (VIF) of a simple linear model using credit rating changes as the dependent variables and financial ratios and π_{Merton} as the independent variables confirms the problem of multicollinearity. The largest VIF (409.1 for logz1) is far bigger than the commonly used cutoff value 10 [Kutner et al. 2004]. To mitigate the problem, we selected financial ratios to cover firm characteristics known to interact with credit ratings: size, financial leverage, profitability, interest coverage, and liquidity. Our eleven control variables including π_{Merton}, size (z1 and z4), financial leverage (z5 and z7), profitability (z11, z13, z15, and z17), interest coverage (z19), and liquidity (z22). We recomputed the VIF using these financial ratios. All VIFs are less than 3.8, which indicates the selected financial ratios are free of the multicollinearity problem. Our subsequent discussion considers the selected financial ratios only.

We report the correlation coefficients between future rating changes and other covariates in Table V. It is interesting to note that the correlation coefficient of news coverage ($article_{i,t}$) is significantly negative, indicating higher news coverage is associated with future downgrades. The overall negative sentiment ($neg\_score_{i,t}$) also has a significantly negative coefficient, suggesting the potential value of the sentiment variables. The overall positive sentiment ($pos\_score_{i,t}$), on the other hand, is not significant. Previous studies [Tetlock 2007] on news sentiment have also found that negative sentiment is more relevant to future market returns compared to positive sentiment.

When divided into different news topics, the correlation coefficients of topic-specific news coverage are still significantly negative for 7 out of 11 included topics. The sign is in general consistent with the overall news coverage variable. Five out of eleven topic-specific negative sentiment variables are significantly negative, while all of them have negative signs. It should be noted that topic-specific sentiment variables have a large portion of missing values and the correlation coefficients are computed based on pairwise availability. Ten out of eleven topic-specific positive sentiment variables are not significant, consistent with the pattern in overall positive sentiment variables.

The Merton distance to default (π_{Merton}) variable is significantly negative, suggesting a higher expected default probability leads to downgrades. All selected financial ratios have correlation coefficients significantly different than zero, suggesting the usefulness of the controlled financial ratios. Both profitability variables (z11, z13, z15, and z17) and liquidity variable (z22) have positive signs, suggesting higher profits and better liquidity may improve future ratings. Firm size, on the other hand, has a negative sign. To understand this counter-intuitive result, consider a firm with a perfect

Table V. Correlation Coefficients Between Future Rating Changes and Other Covariates

Variable	Correlation Coef.	Variable	Correlation Coef.
logz1 (Log Asset Accounting Value)	−0.012**	$article_{i,t,22}$	−0.023***
logz4 (Log Book to Market Value)	−0.11***	$neg\_score_{i,t,22}$	−0.043***
z5 (Long-Term Debts / Total Invested Capital)	−0.054***	$pos\_score_{i,t,22}$	−0.013
z7 (Debt / Equity)	−0.026***	$article_{i,t,31}$	−0.016*
z11 (Operating Income / Received Capitals)	0.119***	$neg\_score_{i,t,31}$	−0.035*
z13 (Net Income Before Tax / Received Capitals)	0.128***	$pos\_score_{i,t,31}$	−0.026
z15 (Gross Profit Margin)	0.06***	$article_{i,t,35}$	−0.008
z17 (Earnings Per Share)	0.016***	$neg\_score_{i,t,35}$	−0.011
z19 (EBIT Interest Coverage)	0.114***	$pos\_score_{i,t,35}$	0.012
z22 (Quick Ratio)	0.116***	$article_{i,t,22}$	−0.023***
π_{Merton} (Merton's Distance to Default; %)	−0.1539***	$neg\_score_{i,t,36}$	−0.083***
$article_{i,t}$	−0.03***	$pos\_score_{i,t,36}$	−0.014
$neg\_score_{i,t,22}$	−0.027***	$article_{i,t,52}$	−0.017***
$pos\_score_{i,t,22}$	−0.006	$neg\_score_{i,t,52}$	0.001
$article_{i,t,4}$	−0.006	$pos\_score_{i,t,52}$	−0.022
$neg\_score_{i,t,4}$	−0.03	$article_{i,t,55}$	−0.011
$pos\_score_{i,t,4}$	−0.036***	$neg\_score_{i,t,55}$	−0.08***
$article_{i,t,11}$	−0.03***	$pos\_score_{i,t,55}$	0.012
$neg\_score_{i,t,11}$	−0.089***	$article_{i,t,82}$	−0.025***
$pos\_score_{i,t,11}$	0.003	$neg\_score_{i,t,82}$	−0.001
$article_{i,t,19}$	−0.024***	$pos\_score_{i,t,22}$	0.015
$neg\_score_{i,t,19}$	0.031		
$pos\_score_{i,t,19}$	0.026		

Correlation coefficients were computed based on pairwise availability. *, **, and *** indicate 1%, 5%, and 10% significance levels.

AAA rating. The rating may be decreased or remain unchanged in the future. The overall future trend for this firm, as a result, is downward. Following this line of argument, firms with good ratings have a larger chance to be downgraded in the future.

The simple correlation analysis suggests that news-based variables are indeed correlated with future credit rating changes. The next question is whether these variables are still valuable after controlling for the financial ratios. We explore this question in the following section.

5.2 Modeling Credit Rating Changes Using Missing-Tolerant Multinomial Probit Model (MT-MNP)

As discussed in the preceding, credit rating changes have three outcomes: −1 (downgrade), 0 (unchanged), and 1 (upgrade). The outcome 0 is used as the base class against the other two possibilities. Thus for a case k that corresponds to firm i in quarter t, two latent variables, $W_k = (W_{k,-1}, W_{k,1})$, determine its future rating changes. Each latent variable is modeled by

$$\begin{aligned} W_{k,j} = \beta_{j,0} &+ article_{i,t}\,\beta_{j,1} + neg\_score_{i,t}\beta_{j,2} + pos\_score_{i,t}\beta_{j,3} \\ &+ \sum_{g\in T} article_{i,t,g}\,\beta_{j,1,g} + neg\_score_{i,t,g}\beta_{j,2,g} + pos\_score_{i,t,g}\beta_{j,3,g} \\ &+ \sum_{h\in S} zh_{i,t}\,\gamma_{j,h} + \pi_{Merton} d_j + e_{k,j};\;\; j = (-1, 1). \end{aligned}$$

The future rating is unchanged ($Y_k = 0$) if $\max(W_{k,-1}, W_{k,1}) \leq 0$, downgrade ($Y_k = -1$) if $W_{k,-1} > W_{k,1} > 0$, or upgrade ($Y_k = 1$) if $W_{k,1} > W_{k,-1} > 0$.

The Gibbs sampling algorithm presented here was used for model estimation. We set the variance of the conjugate prior of the beta coefficients to 1000 (mean = 0). Larger variance, such as 5000 and 10,000 were also considered to check the robustness of the estimation results. We run 10,200 iterations and exclude the first 200 burn-in iterations. Previous studies adopted a similar number of iterations when estimating binary probit models [Albert and Chib 1993].

5.3 Performance Evaluation: Explanatory Power

Table VI summarizes the model estimation results of the MT-MNP model using all observations. The overall news coverage ($article_{i,t}$) has a significantly positive impact (0.003; $p < 0.1$) on future downgrades but not upgrades. Overall negative sentiment and positive sentiment has no explanatory power on future credit rating changes. Compared to the result of the correlation coefficient, the effect of overall negative sentiment disappeared after controlling for financial ratios and π_{Merton}.

When divided into topics, two (Topics 4 and 35) out of the then topic-specific news coverage variables have significant impacts on future upgrades. Topic-specific news coverage, on the other hand, does not explain future downgrades. Topic-specific sentiment variables present a different pattern compared to the overall news sentiment. Three topic-specific negative sentiment (Topics 11, 19, and 36) and two topic-specific positive sentiment variables (Topics 11 and 52) have significant impacts on future downgrades. The negative sentiment of Topics 22 and 55 have a significant impact on future upgrades. These results clearly suggest that fine-grained sentiment analysis can provide incremental value to explain future credit rating changes.

To understand the marginal effect of news-based variables, we compute $Prob(downgrade|X_i + 1.5\sigma_j)/Prob(downgrade|X_i)$ for selected variable j. The ratio reflects the change of future downgrade probability when a variable j is increased by 1.5 standard deviations. We average across all observations so that the result is representative for our sample. Table VII reports the results of selected variables. The probability of downgrade is on-average 1.36 times larger if the negative sentiment of Topic 36 is increased by 1.5 standard deviations. The effects on overall news coverage is smaller (1.17). The effect of z13 and z22 is 0.82 and 0.69, respectively. It is clear that news-based variables have a similar level of impact on the probability of downgrades compared to those of financial ratios.

Figure 3 plots the trace and density of selected variables based on 10,000 draws generated by the Gibbs sampling algorithm. Panel (a) includes the aggregated news coverage and sentiment variables and Panel (b) plots the coverage and sentiment of Topic 36. The traces show little impact of initial values. The results of Geweke tests [Geweke 1992] indicate that the chain indeed converged. The density plots can be used to verify the estimation results listed in Table VII. For example, the top density plot ($article_{i,t,36}$) in Panel (b) covers two sides around zero, suggesting an insignificant estimation, while the middle density plot ($neg_score_{i,t,36}$) resides mostly on the positive side, suggesting a significantly positive effect.

5.4 Prediction Performance

We evaluated the prediction performance of MT-MNP by conducting out-sample prediction quarter-by-quarter. Specifically, at quarter t, the historical data from quarters t-4 to t-1 were combined; two-thirds of them were randomly assigned as the training data while the rest were the tuning data. The training-tuning split was used to select the prior parameters of MT-MNP as well as the hyper-parameter (C) for the baseline

Table VI. MT-MNP Model Estimation Results

Covariate: -1 (Downgrade)	Estimate	Covariate: 1 (Upgrade)	Estimate
Intercept	−1.907***	Intercept	−1.232***
logz1 (Log Asset Accounting Value)	0.001	logz1 (Log Asset Accounting Value)	−0.028**
logz4 (Log Book to Market Value)	0.175***	logz4 (Log Book to Market Value)	−0.043*
z5 (Long-Term Debts/Total Invested Capital)	0.418***	z5 (Long-Term Debts/Total Invested Capital)	−0.043
z7 (Debt/Equity)	0.013*	z7 (Debt/Equity)	0.004
z11 (Operating Income/Received Capitals)	−1.223*	z11 (Operating Income/Received Capitals)	1.43*
z13 (Net Income Before Tax/Received Capitals)	−1.266***	z13 (Net Income Before Tax/Received Capitals)	−0.108
z15 (Gross Profit Margin)	−0.018	z15 (Gross Profit Margin)	−0.189**
z17 (Earnings Per Share)	0.386***	z17 (Earnings Per Share)	−0.682***
z19 (EBIT Interest Coverage)	−1.754***	z19 (EBIT Interest Coverage)	2.458***
z22 (Quick Ratio)	−0.126***	z22 (Quick Ratio)	0.069**
π_{Merton} (Merton's Distance to Default)	0.0078***	π_{Merton} (Merton's Distance to Default)	−0.001
$article_{i,t}$	0.003**	$article_{i,t}$	0.003
$neg\_score_{i,t}$	0.166	$neg\_score_{i,t}$	−0.085
$pos\_score_{i,t}$	0.134	$pos\_score_{i,t}$	0.297
$article_{i,t,4}$	0.026	$article_{i,t,4}$	0.071**
$neg\_score_{i,t,4}$	0.590	$neg\_score_{i,t,4}$	−0.984
$pos\_score_{i,t,4}$	1.463*	$pos\_score_{i,t,4}$	−0.922
$article_{i,t,11}$	0.037	$article_{i,t,11}$	−0.035
$neg\_score_{i,t,11}$	1.508***	$neg\_score_{i,t,11}$	−0.699
$pos\_score_{i,t,11}$	−1.312*	$pos\_score_{i,t,11}$	−0.538
$article_{i,t,19}$	0.017	$article_{i,t,19}$	0.002
$neg\_score_{i,t,19}$	−1.441**	$neg\_score_{i,t,19}$	−0.428
$pos\_score_{i,t,19}$	−1.046	$pos\_score_{i,t,19}$	−0.072
$article_{i,t,32}$	0.010	$article_{i,t,22}$	−0.020
$neg\_score_{i,t,22}$	0.619	$neg\_score_{i,t,22}$	−0.862*
$pos\_score_{i,t,22}$	0.044	$pos\_score_{i,t,22}$	−0.176
$article_{i,t,31}$	0.028	$article_{i,t,19}$	0.006
$neg\_score_{i,t,31}$	0.154	$neg\_score_{i,t,31}$	0.448
$pos\_score_{i,t,31}$	1.376	$pos\_score_{i,t,31}$	−0.575
$article_{i,t,19}$	−0.028	$article_{i,t,19}$	0.031*
$neg\_score_{i,t,35}$	−0.091	$neg\_score_{i,t,35}$	−0.442
$pos\_score_{i,t,35}$	−1.180	$pos\_score_{i,t,19}$	−1.734
$article_{i,t,36}$	0.029	$article_{i,t,36}$	−0.044
$neg\_score_{i,t,36}$	1.212***	$neg\_score_{i,t,36}$	−0.353
$pos\_score_{i,t,36}$	−0.757	$pos\_score_{i,t,36}$	−1.641
$article_{i,t,52}$	0.053	$article_{i,t,52}$	−0.054
$neg\_score_{i,t,52}$	−0.486	$neg\_score_{i,t,52}$	0.443
$pos\_score_{i,t,52}$	1.929*	$pos\_score_{i,t,52}$	0.686
$article_{i,t,55}$	−0.037	$article_{i,t,55}$	−0.025
$neg\_score_{i,t,55}$	0.960	$neg\_score_{i,t,55}$	−2.285***
$pos\_score_{i,t,55}$	−1.797	$pos\_score_{i,t,55}$	0.832
$article_{i,t,82}$	0.023	$article_{i,t,82}$	0.017
$neg\_score_{i,t,82}$	−0.057	$neg\_score_{i,t,82}$	1.103
$pos\_score_{i,t,82}$	−1.393	$pos\_score_{i,t,82}$	1.433

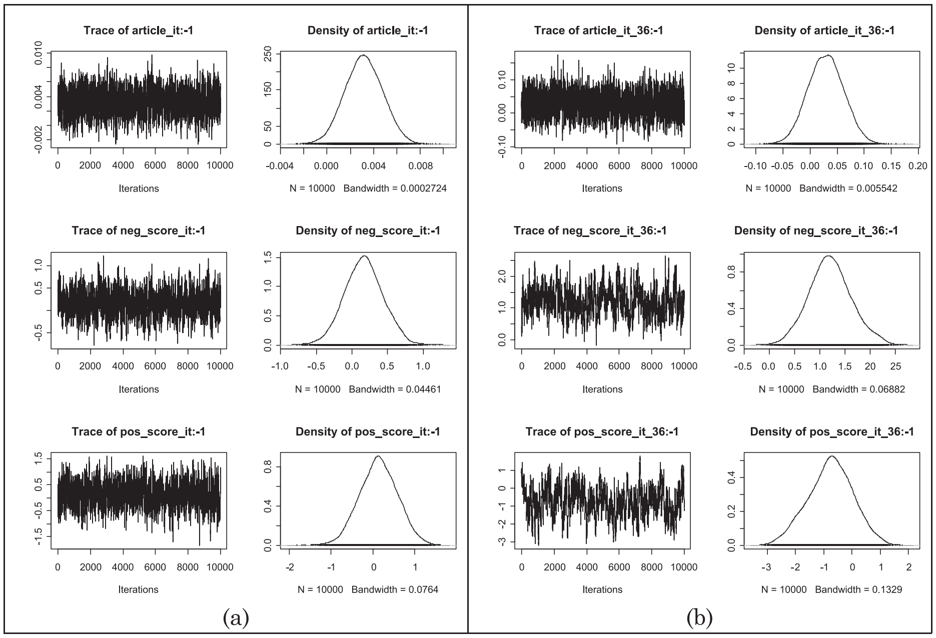
Covariance Matrix	
−1:−1	1.107***
−1:1	0.000
1:1	0.893***

The MT-MNP estimation results were computed based on 10,000 iterations of Gibbs sampling. The prior variance for the beat coefficients was 1000. *, **, and *** indicate 1%, 5%, and 10% significance levels.

Table VII. Marginal Contribution to Downgrade Probability

Variable	ΔP	Variable	ΔP
$neg\_score_{i,t,36}$	1.36^{***}	$article_{i,t}$	1.18^{***}
z13	0.82^{***}	z22	0.69^{***}

*, **, and *** indicate 1%, 5%, and 10% significance levels.



(a) (b)

Fig. 3. Trace and density of selected parameters.[†]

[†] The trace and density plot of selected variables in MT-MNP were based on 10,000 iterations of Gibbs sampling. The prior variance for coefficients was 1000.

SVM model via grid search. We considered the range [0.01, 100] for the variance of beta coefficients. The SVM C parameter is also tuned on [0.01, 100].

For each tuning parameter, the corresponding macroaveraging F measures are computed. The tuning parameters associated with the best macroaveraging F-measure are then recorded. The final models are trained using the selected parameter on the dataset with training and tuning combined. The prediction then can be made based on the data available at quarter t. Realized future credit rating changes are then used to compute the prediction performance.

We provided the prediction models at least three quarters of the historical data. The prediction started from the third quarter of 2000 until the first quarter of 2007, spanning 27 quarters. The 2000Q3 prediction was made based on only 3 quarters of historical data while the remaining 26 quarters have a full year of historical data available to the prediction models.

Table VIII summarizes the prediction performance of the MT-MNP and SVM models. Both news-based variables and financial ratios (including π_{Merton}) were made available to these models. It is clear from the results that both MT-MNP and SVM performed better in predicting unchanged ratings compared to predicting upgrades or downgrades. The recall, precision, and F-measure for unchanged ratings are close

Table VIII. Prediction Performance of MT-MNP and SVM Using News-Based Variables and Financial Ratios[‡]

		−1 (Downgrade)	0 (Unchanged)	1 (Upgrade)	Macro-Averaging
MT-MNP	Recall[†§]	0.1827 (0.0206)	0.9329 (0.0041)	0.0466 (0.0098)	
	Precision[†§]	0.1954 (0.0135)	0.9297 (0.0031)	0.0525 (0.0101)	
	F-Measure[†§]	0.1787 (0.0162)	0.9311 (0.0030)	0.0470 (0.0094)	0.3856 (0.0049)
SVM	Recall[†§]	0.0000 (0.0000)	1.0000 (0.0000)	0.0000 (0.0000)	
	Precision[†§]	0.0000 (0.0000)	0.9219 (0.0036)	0.0000 (0.0000)	
	F-Measure[†§]	0.0000 (0.0000)	0.9593 (0.0019)	0.0000 (0.0000)	0.3198 (0.0006)
Macro-Averaging F-Measure Difference (MT-MNP - SVM)[†]					0.0659 (0.0010) t = 65.91; $p < 0.01$

[†] Standard error is in the parentheses.
[‡] The result is computed from one-quarter ahead prediction starting from 2000Q3 to 2007Q1. Both news-based variables and financial ratios were included for model training. SVM prediction was made based on three one-against-one binary classifiers.
[§] Let TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative of classification results. Recall, precision, and F-measure are computed as:
Recall = TP / (TP + FN)
Precision = TP / (TP + FP)
F-measure = 2 × Precision × Recall / (Precision + Recall)

to one for both MT-MNP and SVM. The tendency for predicting unchanged rating is stronger for the SVM model. The performance for predicting downgrades and upgrades, on the other hand, is much weaker. Among 25,865 firm-quarters in our testing set, the MT-MNP predicted 1327 (5.1%) downgrades and 612 (2.3%) upgrades. For the predicted downgrades, only 19.54% are correct. The precision of predicted upgrades is even worse (5.25%). Among all observed downgrades, only 18.27% of them are correctly predicted (recall). The recall for upgrades is worse (4.66%). Combined with the results of predicting unchanged ratings, the macro-averaging F-measure is 38.56%, with a small standard deviation (0.49%).

The SVM performed even worse. It made no downgrade or upgrade predictions. The recall, precision and F-measure for the downgrades and upgrades, as a result, are all zero. In other words, the macroaveraging F-measure of the SVM model (31.98%) is the same as that of a classifier that predicts the majority class (unchanged ratings). We note that the hyper-parameter (C) was tuned to maximized macroaveraging F-measure during model training. The tuning routine, however, failed to overcome the tendency for SVM to favor the majority class.

Pairwise t-test shows that the difference (0.0659) is statistically significant with a t-value of 65.91. The performance gap is small but consistent across different quarters. The superior performance of MT-MNP mainly comes from predicting credit rating downgrades. The other two categories (unchanged and upgrade) have similar performance.

To check the robustness of the result, we widened the hyper-parameter tuning range for SVM from [0.01, 100] to [0.01, 200] and tested nonlinear kernel functions such as the radial basis and sigmoid. The result remained qualitatively unchanged. It is clear that a standard SVM implementation is not a good choice for predicting credit

Table IX. Prediction Performance of MT-MNP and SVM Using Financial Ratios

		−1 (Downgraded)	0 (Unchanged)	1 (Upgraded)	Macro-Averaging
MT-MNP	Recall[†]	0.0040 (0.0019)	0.9993 (0.0002)	0.0000 (0.0000)	
	Precision[†]	0.0785 (0.0333)	0.9222 (0.0036)	0.0000 (0.0000)	
	F-Measure[†]	0.0076 (0.0034)	0.9591 (0.0019)	0.0000 (0.0000)	0.3222 (0.0012)
SVM	Recall[†]	0.0000 (0.0000)	1.0000 (0.0000)	0.0000 (0.0000)	
	Precision[†]	0.0000 (0.0000)	0.9219 (0.0036)	0.0000 (0.0000)	
	F-Measure[†]	0.0000 (0.0000)	0.9593 (0.0019)	0.0000 (0.0000)	0.3198 (0.0006)
Macro-Averaging F-Measure Difference (MT-MNP - SVM)[†]					0.0025 (0.0002) t = 10.84; p < 0.01

[†] Standard error is in the parentheses.
The result is computed from one-quarter ahead prediction starting from 2000Q3 to 2007Q1. Financial ratios were included for model training. SVM prediction was made based on three one-against-one binary classifiers.

rating changes. Our MT-MNP has a small but consistent margin over the baseline SVM model.

One interesting question is the contribution of news-based variables to prediction performance. We ran through the same process using financial ratios (including π_{Merton}) as the predictors. Table IX summarizes the results. The MT-MNP model has a slightly lower macroaveraging F measure (0.3222) compared to the same model with additional news based variables. The difference (0.0634) is statistically significant ($p < 0.01$). The significant difference suggests that the news-based variables indeed help predicting future credit rating changes. The confusion matrix of out-sampling predictions can be found in online Appendix D.

One pattern that can be observed from the results is the poor performance of upgrade and downgrade predictions. A natural question to ask is whether the poor performance is caused by the specific learning approaches selected in this study. To investigate this question, we conducted out-sample prediction using two additional learning algorithms: Adaboost [Freund and Schapire 1996] and Random Forest [Breiman 2001]. Both learning algorithms are known to perform well on a wide range of tasks (see, e.g., Opitz and Maclin [1999]). We briefly discuss the two algorithms.

Adaboost [Freund and Schapire 1997] is a boosting algorithm, which improves the accuracy of a given base learner. The algorithm iteratively reweights the training dataset so that harder training examples are given higher weights. Subsequent base learners then can focus on classifying these examples. One reason that we selected Adaboost is the connection between Adaboost and SVM [Freund and Schapire 1999]. Adaboost can be seen as a method of solving a margin maximization problem. Both learning approaches put more emphasis on harder examples. Adaboost, however, adopted the L1 norm (as opposed to the L2 norm used by SVM) in the model. The other important difference is the approach used to search solutions. Adaboost employs greedy search while SVM uses kernel methods. These differences may lead to different performance when applied on real-world datasets. The Adaboost.M1 algorithm implemented in the adabag R package was used for the underlying task.

Random forests [Breiman 2001] combines predictions from a large number of decision trees. Each tree is grown using a bootstrap sample from the original dataset.

An unpruned classification tree is then grown. Since the upper bound of the generalization error is associated with the correlation between individual trees, the classification tree algorithm is modified so that a node is spitted by choosing the best predictor among p ($p = \sqrt{\# of\ predictors}$) randomly selected ones. This approach often givens lower correlation among trees and may lead to better overall performance. We adopted the original implementation by Breiman and Cutler[6] for our task.

The prediction performance of the Adaboost and Random Forest algorithms are qualitatively similar to that of the SVM algorithm reported in the preceding. Both approaches made no upgrade predictions. The performance for downgrade prediction is bad. The F-measure for downgrade prediction is 0.0317 for Random Forest and 0.0212 for Adaboost. The macro-averaging F-measure is 0.3302 for Random Forest and 0.3267 for Adaboost. Both approaches perform better than SVM but worse than MT-MNP. Detailed prediction performance together with confusion matrices can be found in the online Appendix D.

It is clear from the additional prediction results that predicting credit rating changes is a difficult problem. While MT-MNP performs better compared to other approaches considered, there is significant room for improvement in the future. The other concern is whether the result is caused by overfitting. All learning approaches considered in this study are designed to be resistant to overfitting. MT-MNP and Random Forest achieved the goal by reducing variance. SVM and Adaboost are resilient to overfitting because of the large-margin solutions achieved. While it is possible that all methods that prevent overfitting failed, it is more likely that the underlying characteristics of the problem on-hand is responsible for the reported results.

6. CONCLUSIONS

Issuer long-term credit ratings provide important information for credit risk management. To improve the ability to model and predict future credit rating changes, we proposed a novel design framework that extracts overall news coverage, overall news sentiment, topic-specific news coverage, and topic-specific news sentiment from news articles. Our missing-tolerant multinomial probit (MT-MNP) model handles missing value by conducting multiple imputation based on the Bayesian framework. The MT-MNP model, as a result, can make use of all nonmissing variables for estimation and prediction.

Using 278,824 WSJ news articles covering 30 quarters, our experiments show that overall news coverage has a statistically significant impact on credit rating downgrades while the effect of overall sentiment on future credit rating changes is not significant. Topic-specific news coverage and sentiment, on the other hand, are useful in explaining future credit rating changes. There is no overlap between topics that impact credit upgrade and those that impact credit downgrade. Topic-specific negative sentiment has a larger impact compared to topic-specific positive sentiment. Compared to the insignificant result of overall news sentiment, the fine-grained topic-specific sentiment variables have incremental value for credit rating change modeling.

Using SVM as a baseline, our MT-MNP model performs slightly better in terms of macro-averaging F-measure. The better performance comes from the better ability to predict future downgrades and upgrades.

Credit rating upgrades have the lowest frequency among the three possible outcomes, which may be one possible reason for the bad performance of upgrade prediction. We are working on developing novel classification approaches that can better handle the imbalanced data. In addition we plan to conduct industry- and market-level

[6] http://www.stat.berkeley.edu/~breiman/RandomForests/

analysis that will reveal the effects of competition and contagion that were found to be important in bankruptcy.

REFERENCES

ABBASI, A. AND CHEN, H. 2008. Cybergate: A design framework and system for text analysis of computer-mediated communication. *MIS Quart. 32*, 811–837.

ABBASI, A., CHEN, H., AND SALEM, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inform. Syst. 26*, 1–34.

AFONSO, A., GOMES, P., AND ROTHER, P. 2009. Ordered response models for sovereign debt ratings. *Appl. Econ. Lett. 16*, 769–773.

AGARWAL, V. AND TAFFLER, R. 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *J. Bank. Finance 32*, 1541–1551.

ALBERT, J. H. AND CHIB, S. 1993. Bayesian analysis of binary and polychotomous response data. *J. Amer. Stat. Assoc. 88*, 669–679.

ALTMAN, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance 23*, 589–609.

ALTMAN, E. I. AND HOTCHKISS, E. 2005. *Corporate Financial Distress and Bankruptcy: Predict and Avoid Bankruptcy, Analyze and Invest in Distressed Debt*. John Wiley & Sons, New York, NY.

ANTWEILER, W. AND FRANK, M. Z. 2004, Is all that talk just noise? The information content of internet stock message boards. *J. Finance 59*, 1259–1294.

BAI, X. 2011. Predicting consumer sentiments from online text. *Dec. Supp. Syst. 50*, 732–742.

BEAVER, W. H. 1966. Financial ratios as predictors of failure. *Account. Res. 4*, 71–111.

BELLOTTI, T., MATOUSEK, R., AND STEWART, C. 2011. A note comparing support vector machines and ordered choice models' predictions of international banks' ratings. *Decision Supp. Syst. 51*, 682–687.

BHARATH, S. T. AND SHUMWAY, T. 2008. Forecasting default with the Merton distance to default model. *Rev. Financial Studies 21*, 1339–1369.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

BREIMAN, L. 2001. Random forests. *Mach. Learn. 45*, 5–32.

CECCHINI, M., AYTUG, H., KOEHLER, G. J., AND PATHAK, P. 2010. Making words work: Using financial text as a predictor of financial events. *Decision Supp. Syst. 50*, 164–175.

CHAN, W. S. 2003. Stock price reaction to news and no-news: Drift and reversal after headlines. *J. Financ. Econ. 70*, 223–260.

CHANG, C.-C. AND LIN, C.-J. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. 2*, 1–27.

CHEN, K.-T., LU, H.-M., CHEN, T.-J., LI, S.-H, LIAN, J.-S., AND CHEN, H. 2011. Giving context to accounting numbers: The role of news coverage. *Decision Supp. Syst. 50*, 673–679.

DAS, S. R. AND CHEN, M. Y. 2007. Yahoo! For amazon: Sentiment extraction from small talk on the Web. *Manage. Sci. 53*, 1375–1388.

DAS, S. R., P. HANOUNA, P., AND SARIN, A. 2009. Accounting-based versus market-based cross-sectional models of CD spreads. *J. Bank. Finance 33*, 719–730.

DUFFIE, D. AND SINGLETON, K. J. 2003. *Credit Risk: Pricing, Measurement, and Management*. Princeton University Press.

FANG, L. AND PERESS, J. 2009. Media coverage and the cross-section of stock returns. *J. Finance 64*, 2023–2052.

FINKEL, J. R., GRENAGER, T., AND MANNING, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.*

FREUND, Y. AND SCHAPIRE, R. E. 1996. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning.*

FREUND, Y. AND SCHAPIRE, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci. 55*, 119–139.

FREUND, Y., AND SCHAPIRE, R. E. 1999. A short introduction to boosting, *J. Japanese Society Artif. Intell. 14*, 771–780.

GEMAN, S. AND GEMAN, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell. 6*, 721–741.

GEWEKE, J. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4,* J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith Eds., Clarendon Press, Oxford, UK.

GREENE, W. 2008. *Econometric Analysis*. Prentice Hall.

GRIFFITHS, T. AND STEYVERS, M. 2004, Finding scientific topics. *Proc. Natl. Acad. Sci. 101*, 5228–5235.

HILLEGEIST, S. A., KEATING, E. K., CRAM, D. P., AND LUNDSTEDT, K. G. 2004. Assessing the probability of bankruptcy. *Rev. Account. Stud. 9*, 5–34.

HUANG, Z., CHEN, H., HSU, C.-J., CHEN, W.-H., AND WU, S. 2004. Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Supp. Syst. 37*.

HWANG, R.-C., CHENG, K. F., AND LEE, C.-F. 2009. On multiple-class prediction of issuer credit ratings. *Appl. Stoch. Models Bus. Indus. 25*, 535–550.

HWANG, R.-C., CHUNG, H., AND CHU, C. K. 2010. Predicting issuer credit ratings using a semiparametric method. *J. Empir. Finance 17*, 120–137.

IBRAHIM, J. G., CHEN, M.-H., AND LIPSITZ, S. R. 2002. Bayesian methods for generalized linear models with covariates missing at random. *Canadian J. Statist. 30*, 55–78.

IMAI, K. AND VAN DYK, D. A. 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econometrics 124*, 311–334.

JOACHIMS, T. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola Eds., MIT Press.

JONES, S. AND HENSHER, D. A. 2004. Predicting firm financial distress: A mixed logit model. *Account. Rev. 79*, 1011–1038.

KOTHARI, S. P., LI, X., AND SHORT, J. E. 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *Account. Rev. 84*, 1639–1670.

KUTNER, M., NACHTSHEIM, C., AND NETER, J. 2004. *Applied Linear Regression Models*. McGraw-Hill.

LAU, R. Y. K., LIAO, S. Y., KWOK, R. C.-W., XU, K., XIA, Y., AND LI, Y. 2011. Text mining and probabilistic lnaguage modeling for online review spam detection. *ACM Trans. Manage. Inform. Syst. 2*.

LIN, C. AND HE, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the ACM Conference on Information and Knowledge Management.*

LOUGHRAN, T. I. M. AND MCDONALD, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *J. Finance 66*, 35–65.

MANNING, C. D., RAGHAVAN, P., AND SCHUTZE, H. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

MCCALLUM, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

MENG, X.-L. AND VAN DYK, D. A. 1999. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika 86*, 301–320.

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

OHLSON, J. A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res. 18*, 109–131.

OPITZ, D. AND MACLIN, R. 1999. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res. 11*.

PANG, B., LEE, L., AND VAITHYANATHAN, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

RUBIN, D. B. 1976. Inference and missing data. *Biometrika 63*, 581–592.

SHEATHER, S. J. AND JONES, M. C. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Statist. Soc. Series B (Methodological) 53*, 683–690.

SHUMWAY, T. 2001. Forecasting bankruptcy more accurately: A simple hazard model. *J. Bus. 74*, 101–124.

STANDARD & POOR'S. 2003. Ratings and ratios: Corporate ratings criteria, Standard & Poor's Report.

TANNER, M. A. 1996. *Tools for Statistical Inference.* Springer-Verlag, New York.

TETLOCK, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. *J. Finance 62*, 1139–1168.

TETLOCK, P. C., SAAR-TSECHANSKY, M., AND MACSKASSY, S. 2008. More than words: Quantifying language to measure firms' fundamentals. *J. Finance 63*, 1437–1467.

UHL, M. W. 2011. Explaining U.S. Consumer behavior with news sentiment, *ACM Trans. Manage. Inform. Syst. 2*, 9.

VAPNIK, V. 1995. *The Nature Of Statistical Learning Theory.* Springer-Verlag.

WIEBE, J., WILSON, T., AND CARDIE, C. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resources Eval. 39*, 165–210.

WILSON, T., WIEBE, J., AND HOFFMANN, P. 2005a. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT EMNLP)*.

WILSON, T., WIEBE, J., AND HOFFMANN, P. 2005b. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Comput. Ling. 35*, 399–433.

ZHANG, Z., LI, X., AND CHEN, Y. 2012. Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews. *ACM Trans. Manage. Inform. Syst. 3*.

Received August 2011; revised January 2012, June 2012; accepted July 2012