Sanjiv R. Das is the Terry Professor of Finance and Data Science at Santa Clara University in Santa Clara, CA
srdas@scu.edu

Seoyoung Kim is Associate Professor of Finance at Santa Clara University in Santa Clara, CA
srkim@scu.edu

Bhushan Kothari is Analyst, Trust and Safety at Google in Mountain View, CA
bhushankothari1992@gmail.com

Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News

Sanjiv R. Das Seoyoung Kim Bhushan Kothari

Abstract

Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News

Natural language processing (NLP) is a fast-growing area of data science for the finance industry. The authors demonstrate how an applied linguistics expert system may be used to parse corporate email content and news to assess factors predicting escalating risk or the gradual shifting of other critical characteristics within the firm before they are eventually manifested in observable data and financial outcomes. They find that email content and news articles meaningfully predict increased risk and potential malaise. We also find that other structural characteristics, such as average email length, are strong predictors of risk and subsequent performance. Implementations of three spatial analyses of internal corporate communication, (i.e., email networks, vocabulary trends, and topic analysis) are presented. The authors propose a RegTech solution by which to systematically and effectively detect escalating risk or potential malaise without the need to manually read individual employee emails.

JEL classification: G00; G01; G28; G38.
Keywords: Expert systems; Fintech; Regtech; Corporate governance; Text mining; Email analysis; Email networks; Mood and net sentiment.

The plethora and continuous flow of text makes it a far more prevalent and timely source of information than tabulated numbers and figures. Yet, difficulties lie in systematically and efficiently drawing appropriate inferences from large bodies of unstructured text. On one hand, manual parsing ensures proper comprehension within the relevant context and dialect; however, this method is not only inefficient, but also raises substantial privacy concerns. Thus, a platform built to detect escalating risk or malaise, without actually reading individual employee emails, is highly valuable to academics, regulators, and practitioners along both practical and ethical dimensions. Furthermore, network changes evidenced by shifting email clusters can also indicate sources ripe for investigation, as fraudulent activity among corporate employees tends to occur in social-network clusters.[1]

In this paper we use a fast-growing area of data science in finance, natural language processing (NLP), to develop an early-warning system for detecting corporate failure. Our purpose is to explore which indicators within employees' sent email content and sender/recipient networks can effectively predict changes in risk and subsequent performance. We examine not only sentiment-based indicators contained in the message bodies, but also non-textual structural characteristics such as the number of emails sent, the average length per email, and the shifting sender/recipient networks within the company over time. Of particular value are the non-content-based structural/network indicators of potential trouble, since email content may be easier to control or manipulate than the connectivity of a network.

Specifically, we parse a set of over 100,000 emails sent by 144 Enron employees, mostly senior management, spanning the critical period from January 2000 through December 2001, ending in Enron's demise. This data was first made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation into Enron's practices, and was subsequently collected and distributed by the Carnegie Mellon CALO project.[2] We focus on sent emails rather than all emails within the Enron corpus to: (i) analyze content specifically written by Enron employees during the time in question, and (ii) avoid processing the same content more than once.

We develop a mood/net sentiment score over time based on various positive/negative context-dependent dictionaries which we match to the email content. We then plot these sentiment scores against moving average stock returns and moving average stock prices over time. We also study other structural characteristics, such as the total emails sent, the number of recipients, and the average email length over time.

In additional exploratory analyses, we present implementations of three spatial analyses of internal corporate communication, i.e., email networks, vocabulary trends, and topic analysis. We construct a network graph of employees over time based on emails sent/received among employees within the firm, which we may use to explore how network connectedness changes over time as well as how pivotal employees (i.e., nodes) in the network shift. Ultimately, we plan to explore the information contained in major network shifts as to changes in firm risk and subsequent performance. We also employ Latent Dirichlet Allocation (LDA) to perform topic analyses on the sent email content to explore trends and shifts in main topics or concerns over time, which we can compare to simple word counts over time to explore how the use certain keywords fluctuates along the Enron timeline.

In summary, email and news content meaningfully predicts risk and potential malaise, and other structural characteristics, such as average email length, are predictors of risk and

[1] See, for instance, coverage regarding the recent fake-accounts scandal at Wells Fargo: http://www.thedailybeast.com/articles/2016/10/12/how-wells-fargo-made-5-000-criminals.html.

[2] See http://www.cs.cmu.edu/~enron/.

subsequent performance. Overall, this paper suggests the efficacy of a RegTech solution by which to systematically and effectively detect escalating risk or the potential demise of a firm without the need to manually read individual employee emails.

LITERATURE AND THEORETICAL DEVELOPMENT

In this section, we connect our model to the theoretical and empirical literature in four areas. And we present the research questions surrounding our proposed RegTech solution.

Relation to the Extant Literature

In this subsection we discuss the existing literature as it relates to our work, highlighting the main differences, with the view to describe why our approach has some advantages over existing methodologies. Specifically, this paper relates to four distinct literatures in the areas of text analytics, applied linguistics, risk management, and regulatory technology (RegTech).

In the text-analytics arena, there is extensive research on stock-return prediction using sentiment extracted from internet postings on message boards such as Yahoo!, Motley Fool, Silicon Investor, Raging Bull, etc.[3] The information content of other more official news sources, such as Lexis-Nexis, Factiva, Dow Jones News, etc., have also been explored (see Das et al. (2005), Boudoukh et al. (2013), Boudoukh et al. (2018)), and the "Heard on the Street" column in the *The Wall Street Journal* has been text-mined by Tetlock (2007) and Tetlock et al. (2008), and text from *The Wall Street Journal* news articles is used by Lu et al. (2010). Overall, these studies are based on public data, and we expect that using internal data (e.g., emails) offers substantially greater access to timely and relevant insights about a firm's impending risks. This distinguishes our paper from this extant literature, though we also use news in our empirical analyses for comparison to the insights extracted from emails.

In addition to the literature on stock message boards and news articles, there is an extensive and growing literature on extracting sentiment from Twitter feeds. For instance, the seminal paper by Bollen et al. (2011) provides evidence suggesting that tweets can predict the direction of the Dow Jones Index with 87% accuracy. The correlation between tweets and the market is further explored and found to be significant in several papers.[4] This literature uses short-run tweets data, and is unlikely to pick up longer term trends in corporate fortunes as we aim to do in our own analyses using email sentiment over longer periods of time. And finally, there is an extensive literature on using financial reports such as 10-Ks.[5] These reports occur at a low frequency (i.e., on a quarterly or annual basis) and are unlikely to be as timely as emails in detecting emerging problems. Therefore, emails have two advantages – they are likely to look farther ahead than tweets, and they are more timely than 10-Ks.

Unlike numerical data, textual data may be able to extricate corporate risk better than hard numbers, as it contains richer characteristics and additional nuances such as sentiment, word length, the readability of text, size of the text file, etc. The research record shows that readability matters, and Loughran and McDonald (2014) find that poor readability of

[3] See for example Tumarkin and Whitelaw (2001), Antweiler and Frank (2004), Das et al. (2005), Das and Chen (2007).

[4] See Sprenger et al. (2013); Sprenger (2011); Bar-Haim et al. (2011); Brown (2012); and Rao and Srivastava (2014).

[5] See Loughran and McDonald (2011); Burdick et al. (2011); Bodnaruk et al. (2013); Jegadeesh and Wu (2013); and Loughran and McDonald (2014).

10-Ks indicates risk in the future, partly because companies tend to report bad news in an obfuscated manner. However, the size of the 10-K filing seems to also matter, as bad news tends to be wrapped in a greater degree of verbiage than good news. Given this antecedent, it is not surprising that simply looking at the file size in bytes, uploaded to the SEC server, delivers a good predictor of future malaise of a firm. In this paper, we develop a range of metrics from the text of internal emails and examine these same issues, thereby leveraging existing ideas in the finance literature.

Regarding the literature within the risk-management arena, risk management in banks embodies many areas of risk, such as market risk, credit risk, liquidity risk, and operations risk.[6] Market risk refers to the changes in portfolio value arising from changes in market values such as stock prices, interest rates, exchange rates, commodities, etc. Credit risk arises from the risk of default in a bank's loan portfolio and its holdings of bonds. Assets that become illiquid (defined as a reduced ability to sell off a position at current prices) usually decline in price, and we may imagine liquidity risk to be a form of market risk, though it is treated as a distinct category by finance practitioners. Finally, operations risk arises from the detrimental failure of a financial institution to implement its operating processes, be they risk management methods or otherwise (e.g., being hacked). All of these risk management failures, if unchecked, can result eventually in bankruptcy. There are several applications in the expert systems field that develop bankruptcy prediction algorithms (see Shin and Lee (2002), Back et al. (1996) for the use of genetic algorithms; Min and Lee (2005) for support vector machines; Kim and Kang (2010), Tsai and Wu (2008) for ensemble neural networks; Wang et al. (2012) for credit scoring using ensemble trees). These papers use machine learning and advanced econometrics on financial data from firms to predict bankruptcy. Our paper complements these articles in showing how textual information in emails may be used as an early indicator of failure as well.

While there are well-established approaches for managing market, credit, and liquidity risk, analyzing text from communications within the firm offers a unique and new approach to detecting escalations in operations risk. Furthermore, it also offers a way to assess or predict the impact of the first three risks on the firm's financial condition. For instance, by looking at the change in the frequency of risk words over time or by analyzing shifting email network metrics, one may be able to detect the sudden emergence of any of these forms of risk that may otherwise remain undetected for a longer period of time. These word frequency plots, which we present later in the paper, resemble graphs generated using Google Trends and have the same functionality. We believe that this style of RegTech, i.e., the ongoing use of textual and internal network-based information at higher frequencies, provides another pillar of risk control measures consistent with the ideas in Basel III.[7]

Overall, RegTech is a new concept in the application of data science to financial risk management. Investopedia defines RegTech as a portmanteau of "regulatory technology" that was created to address regulatory challenges in the financial services sector through innovative technology. It is one part of the growing area known as FinTech. The importance of RegTech has grown rapidly since the financial crisis, as more than $160 billion has been paid in fines by various financial institutions, and the global demand for RegTech services

[6]Forecasting and optimizing of these risks is an active area of research in expert systems: Rada (2008); Boyacioglu and Avci (2010); Huang and Jane (2009); Lee et al. (2009); Atsalakis and Valavanis (2009); Wang (2002); Berutich et al. (2016); Guresen et al. (2011).

[7]For Basel III, see http://www.bis.org/bcbs/basel3.htm.

and products is expected to reach about $120 billion by 2020.[8] About 10-15% of the staff in financial institutions is dedicated to compliance,[9] suggesting that a RegTech solution would create a first-order reduction of costs. Arner et al. (2017) argue that RegTech is much more than a mere digitization of manual processes for reporting and compliance. It is indeed a paradigm shift in the way of thinking about risk management. It entails a move from know-your-customer (KYC) to know-your-data (KYD). Whereas much of RegTech deals with ex-post compliance, the approach in our paper is predictive, and uses emails to provide an early warning of impending risks, implementing a philosophy wherein early detection and prevention is better than a cure. In sum, RegTech is a significant arena within the burgeoning field of FinTech.

This paper also has many connections to the computational linguistics literature, i.e., in natural language processing (NLP), and to the more recent arena of statistical language processing (SLP), which resides in machine-learning. The emails we use are scored for sentiment using traditional context-dependent lexicons, such as those developed by Loughran and McDonald (2011). The widely known "syuzhet" algorithm is used in part, see Nielsen (2011). This sentiment scoring approach is based on the theoretical idea that scoring text is based on both the narrative story ("fabula") and its organization ("syuzhet"), an idea promoted by Russian linguists Propp (1928) and Shklovsky (1917). This lies within the broader field of linguistic annotation.

Our approach here also relates to the area of discourse processing, in which topic modeling is now an important area that is seeing tremendous growth. Whereas sentiment extraction operates at the word and sentence level, topic modeling operates at higher level, such as at the paragraph or document level in its entirety, referred to as the 'episode' or 'topic unit', see Halliday and Hasan (1976). Topic modeling is now implemented through Markov Chain Monte Carlo methods using the seminal work of Blei et al. (2003), of which we provide examples later in the paper as well.

An interesting contribution of our paper also lies in its *spatial* text analytics, using email networks, network degree distributions, and word frequencies over time. These ideas come from network modeling (Das (2016)), and also from ideas in Google Trends, as well as the treatise on spatial text over time, see Aiden and Michel (2013). These spatial methods facilitate an iterative design science research approach through visualizations, enabling regulation via an amalgam of text analytics, spatial methods, and linguistics.

Research Questions and Design Science Guidelines

Given the extant literature with regard to information contained in the textual portions of public dialogue and formally issued statements, a natural question arises as to whether the informal, internal dialogue among employees can be used to predict escalating risk or malaise in a timely manner. That is, using the context-dependent sentiment dictionaries from prior literature, we quantify a net sentiment score based on employee email content over time, which we can relate to subsequent stock returns of the corporation.

We have also seen in prior work that the information contained in the net sentiment score of public filings is enhanced, or perhaps even encompassed, by the information contained in objective, nonverbal, structural characteristics such as document size or length. Another

[8] See: https://letstalkpayments.com/a-report-on-global-regtech-a-100-billion-opportunity-\market-overview-analysis-of-incumbents-and-startups/.

[9] See: Martin Arnold, "Market grows for regtech, or AI for regulation," Financial Times, 13 Oct 2016.

natural question then arises as to whether internal email content itself is sufficiently informative, or whether other nonverbal structural characteristics of employee emails provide incremental information as to subsequent stock returns of the corporation.

Thus, in summary, we seek to address the following research questions:

1. Does the sentiment conveyed by employee communications (in our case, internal employee emails) contain value-relevant information?

2. Is this information conveyed in a timely manner (i.e., does email sentiment lead subsequent stock returns)?

3. Do other structural characteristics of internal employee emails (e.g., email length, email volume, or email-network characteristics) also contain value-relevant information?

4. Between the actual verbal content versus structural characteristics of employee emails, which tends to contain more value-relevant information?

A novel contribution of this paper is that it intersects the field of finance with the paradigm of *design science research* (see Hevner et al. (2004)) in information systems. Since design science research is aimed at practical applications of information systems to improving performance, the interdisciplinary ideas in this paper dovetail well with this notion of design science. Van Aken (2005) suggests that the main goal of the design science research is to develop methods that professionals in the field may then apply, a characterization that directly applies to the work in this paper. As with design science research, our approach is practically motivated, and falls under the purview of Mode 2 research, i.e., work that iterates between design and implementation, with a practical and not merely academic goal (the latter being Mode 1; e.g., see Van Aken (2005); Markus et al. (2002)). Design science research is an evolution from pure theory-based measures, to a more practical one, as we undertake in this paper (Iivari (2007)).

Following the Design-Science paradigm outlined by Hevner et al. (2004), we provide a basic mapping of our approach to the Design-Science research guidelines as follows:

1. *Design as an Artifact.* Our purpose is to introduce a prescriptive manual for textual, structural, and network analysis of internal employee communications with the view to employing this platform for ongoing assessments as to potential problems within the firm in a timely manner.

2. *Problem Relevance.* Given the rise of financial technology (FinTech) in general, and regulatory technology (RegTech) in particular, and the growing costs associated with regulatory compliance, a technology-based solution to assess internal communications is increasingly attractive. Furthermore, this technological method is also highly valuable due to the privacy concerns raised in manually parsing internal employee communication.

3. *Design Evaluation.* Data-based caveats aside (which we broach in the data section of this paper), we find a strong predictive association between the sentiment conveyed by employee email content and subsequent stock returns. We also find a strong predictive association between email length and subsequent stock performance. We present the results in detail further below, after we describe our data and parsing process.

4. *Research Contributions.* Overall, our paper adds to extant work in the areas of text analytics, applied linguistics, risk management, and regulation technology (RegTech). Sitting at this multidisciplinary intersection, we seek to expand these works by designing a platform by which to parse email content and to extract other useful metrics encompassing email usage among employees, and to demonstrate the efficacy of this information in providing timely and relevant insights about a firm's impending risks.

We now proceed to discuss our dataset and process flow.

DATA SOURCES, METHODOLOGY, MEASURES

Email Corpus

Our sample spans the two-year period from January 2000 through December 2001. We begin with the entire set of close to 500,000 emails during this time frame, which we obtain from a dedicated site for the Enron corpus hosted by Carnegie Mellon's computer science department, largely to provide a test sample for understanding the preferred structural organization of emails across individual accounts and to develop improved email tools along this regard.

This dataset was first made publicly available by FERC during its investigation of Enron, and was subsequently collected and distributed by the Carnegie Mellon CALO project. Over time, the Enron corpus has undergone further cleaning for legal reasons and as part of a redaction effort due to requests from affected employees. Details regarding exclusion criteria have not been made public, though some obvious redactions are readily identifiable. For instance, user "fastow-a" is notably missing,[10] and much of the email chatter relating to Mr. Skilling's sudden resignation on August 14, 2001 has been expunged. As such, our analyses should be viewed as a prescriptive manual, with implications for analyses and methodology in textual parsing, rather than as a positive study as to the exact workings and trends within the Enron employee network.

In contrast to financial statements and stock charts, a set of emails lacks the structure required for immediate content analysis. Thus, we begin the process by cleaning and organizing the Enron email corpus to arrive at a suitable structure for our analyses. That is, we first parse email content to extract the sender, recipient(s), date of email, email length, and actual message body itself, which we store as separate elements of a data frame. In extracting the message body, we remove redundant factors such as forwarded content as well as repeated content from previous threads in a reply email.[11]

Finally, we focus on *sent* emails rather than all emails within the Enron corpus to: (i) analyze content specifically written by Enron employees during the time in question, and (ii) avoid processing the same content more than once. That is, if user "lay-k" sends an email to "skilling-j", then the same content would be double counted if we dually analyzed both *sent* as well as *received* messages. Throughout this process, we also filter along email length and the total number of recipients on each email in an effort to remove noisy (i.e. junk) mail. Specifically, we remove emails greater than 3,000 characters in length, and we also

[10] Andrew Fastow was the CFO of Enron and a key person in the investigations; therefore, redaction of his emails is material.

[11] We employ R throughout this process. See, for instance, Mizumoto and Plonsky (2015) for an introduction to using R in applied linguistics analyses. Feature selection for text classification is discussed in Uysal (2016).

Exhibit 1: *Summary statistics.* This exhibit presents descriptive statistics of various email characteristics. Our sample encompasses the 113,266 sent emails of 144 employees of Enron Corporation over the period spanning January 2000 through December 2001.

Panel A. Characteristics by Employee ($N = 144$)						
Variable	Mean	Min	P25	Median	P75	Max
Emails per Person	787	2	105	349	891	8,793
Average "Connectedness"	1.62	1	1.21	1.44	1.76	4.47
Average Length per Person	279.92	19.15	160.45	227.90	338.07	944.23
Panel B. Email Characteristics ($N = 113,266$)						
Variable	Mean	Min	P25	Median	P75	Max
Length of Email (# of characters)	362	0	46	163	466	2,998
Direct Recipients per Email ("to")	1.44	0	1	1	1	20
Indirect Recipients per Email ("cc")	0.32	0	0	0	0	19
Total Recipients per Email	1.77	1	1	1	2	20

remove emails sent to more than 20 recipients, since such messages are unlikely to contain relevant content, and are more likely to be general information emails. These additional filters remove approximately 3,000 emails from our final sample.[12]

Overall, we are left with a set of 113,266 sent emails of 144 Enron employees, mostly senior management, over the period spanning January 2000 through December 2001.[13] Finally, we obtain stock prices and returns from the Center for Research in Securities Prices (CRSP), for additional analyses relating to the risk and returns of the corporation during this time, and we obtain sentiment dictionaries for word classification from the Harvard Inquirer and the Loughran and McDonald sentiment word lists, developed in their various papers.[14] [15] The system flow chart for the entire expert system is shown in Appendix .

In **Exhibit 1**, we present summary statistics on our sample of 113,266 sent emails of 144 Enron employees over the period spanning January 2000 through December 2001. The average email in our sample is sent to 1.77 recipients and is 362 characters in length, with a median of 163 characters.[16] We also observe that a sizable minority of sent emails are simply forwarded without added text, as indicated by the 12,222 emails with a character count of zero (not tabulated).

We observe a noticeable trend in email length (number of characters) over time, as shown in **Exhibit 2**. Throughout the year 2000 as well as the earlier portion of 2001, the average email length is relatively stable, straddling approximately 400 characters per email. We

[12]We find that our results are not sensitive to variations in cutoff choices in the filtering process.

[13]There were a few emails before and after this period, but we restricted the sample to these two specific years as they are the bookends to the critical period in the Enron end game.

[14]We obtain the sentiment dictionaries from: (i) `http://www.wjh.harvard.edu/~inquirer/homecat.htm` and (ii) `https://www3.nd.edu/~mcdonald/Word_Lists.html`, respectively.

[15]See Loughran and McDonald (2011); Loughran and McDonald (2014); and Jegadeesh and Wu (2013) for examples of prior studies employing these sentiment dictionaries to gauge the overall mood of financial statements. See also Das (2014) and Loughran and McDonald (2016) for recent surveys regarding the use of textual analysis in accounting and finance.

[16]This median character count comports with a recent study of two million Yahoo Mail users, where the median reply length was 31 words for users in the 36-to-50 age bracket. See Greenwood (2016).

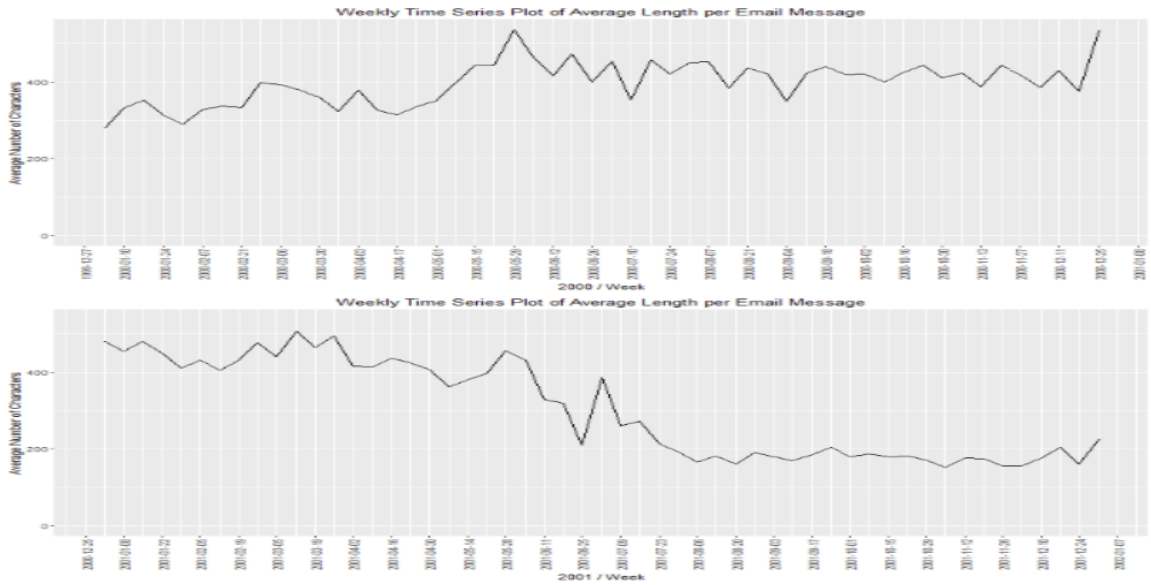

Exhibit 2: *Email length.* This exhibit plots the average number of characters per email on a weekly basis, in the period spanning January 2000 through December 2001.

notice a marked decline in average email length as we progress further into 2001, with the average character count declining by approximately 50%.

We define sentiment in a simplistic way by counting positive and negative words based on the aforementioned Harvard inquirer dictionary.

- Each word is classified as negative, positive, or uncertain based on word classifications provided by a compilation of the Harvard Inquirer. See Stone and Hunt (1963), and the Loughran and McDonald (2011) for sentiment word lists.
- Net sentiment is the difference between the number of positive and negative words, scaled by the sum of positive and negative words; i.e., $\frac{Pos-Neg}{Pos+Neg}$. This metric is often denoted as "polarity".
- Disagreement is defined as one minus the absolute difference between the number of positive and negative words, scaled by the sum of positive and negative words; i.e., $1 - \frac{|Pos-Neg|}{Pos+Neg}$.

We also observe substantial trending in net sentiment and disagreement in sent email content over time, which we present in **Exhibit 3**. Overall, we observe a marked decline in net sentiment and a marked increase in disagreement as we progress further into 2001.

Factiva News Data

In order to examine multiple sources of textual data we also extract news articles from `Factiva`, and we undertake similar text mining of these articles to generate a time series of news sentiment on a weekly basis. We mine only the articles regarding Enron that appear on PR Newswire (US) over the two-year period spanning January 2000 through December 2001, which delivers a total of 1,302 articles. The plot of the article frequency over time is

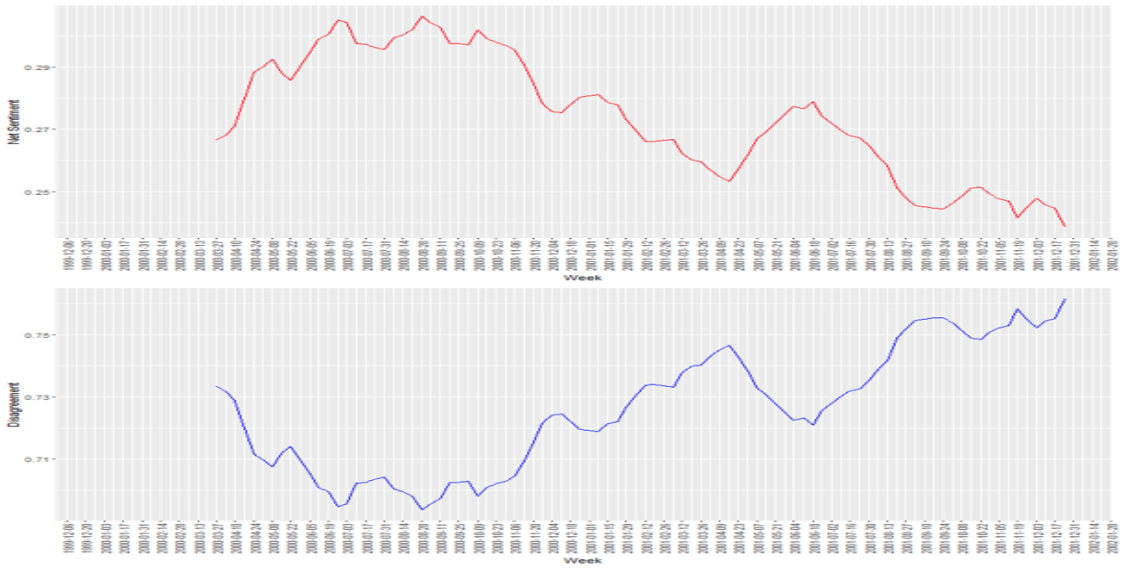

Exhibit 3: *Email sentiment and disagreement.* This exhibit plots the 13-week (i.e., quarterly) moving-average net sentiment index (red line) along with the 13-week moving average disagreement (blue line) on a weekly basis, for the period spanning January 2000 through December 2001. We aggregate emails on a weekly basis and parse the content to determine the extent of positive, negative, and ambiguous sentiment. Net Sentiment and Disagreement are then calculated as $\frac{Pos-Neg}{Pos+Neg}$ and $1 - \frac{|Pos-Neg|}{Pos+Neg}$, respectively.

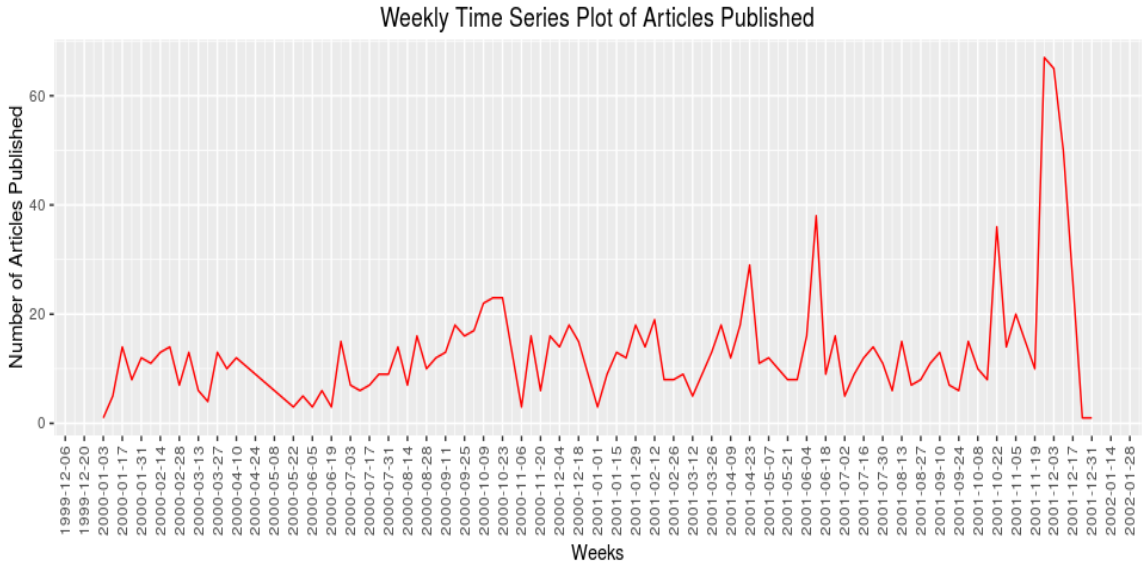

Exhibit 4: *Factiva News Articles.* This exhibit plots the number of news articles from PR Newswire (US) in Factiva on a weekly basis, for the period spanning January 2000 through December 2001. The spike in news articles in the last quarter of 2001 coincides with the period when Enron filed for bankruptcy.

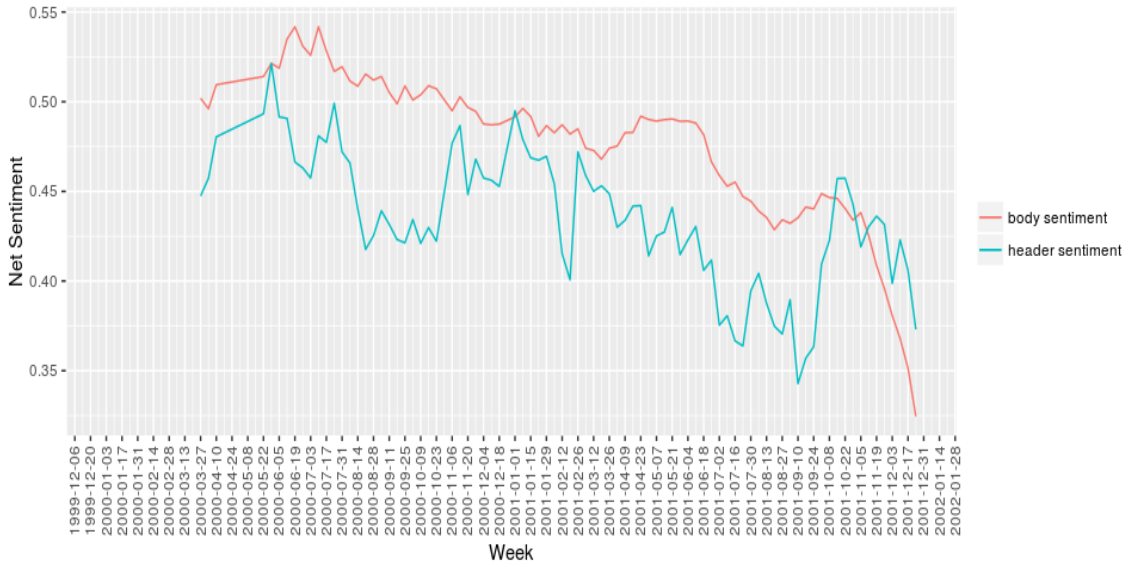

Exhibit 5: *Factiva News Sentiment.* This exhibit plots the 13-week (i.e., quarterly) moving-average net sentiment index, derived from news articles on Factiva PR Newswire, on a weekly basis for the period spanning January 2000 through December 2001. The net sentiment index based on text gleaned from the bodies of the news articles is displayed in red, and the net sentiment index based on text gleaned from the headers of the articles is displayed in blue.

shown in **Exhibit 4**. As expected, we observe that news coverage spikes as we approach Enron's demise.

As with email content, we extract the text from these articles, which after suitable cleaning, we bifurcate into two separate categories: one for the article header, and one for the article body itself. We then apply the same sentiment analysis here as was employed on the Enron email corpus. As with email sentiment over time, we observe a marked decline in news-based sentiment over time, as shown in **Exhibit 5**. That is, we observe a downward trend in the moving average net sentiment conveyed by the news body content (red line) as well as for the sentiment conveyed by the news header alone (blue line).

EMAIL CONTENT AND STOCK RETURNS OVER TIME

We begin by visually examining moving average net sentiment in relation to moving average stock returns and stock prices over time, where moving averages are calculated based on a 13-week (i.e., quarterly) look back period. The results, which we plot in **Exhibit 6** (moving average stock returns) and **Exhibit 7** (moving average stock prices), indicate that stock-return and stock-price movements generally follow the net sentiment of email content over time, with a marked decline in all measures toward the end of the Enron lifecycle.

To explore the role of non-content-based structural characteristics, we also plot the moving average stock returns alongside the moving average email length over time (**Exhibit 8**). Likewise, we plot the moving average stock prices alongside moving average email length over time (**Exhibit 9**). Interestingly, we observe that email length is a potentially powerful indicator of corporate malaise, with email length trending downward along with stock returns and stock prices as we approach Enron's demise.

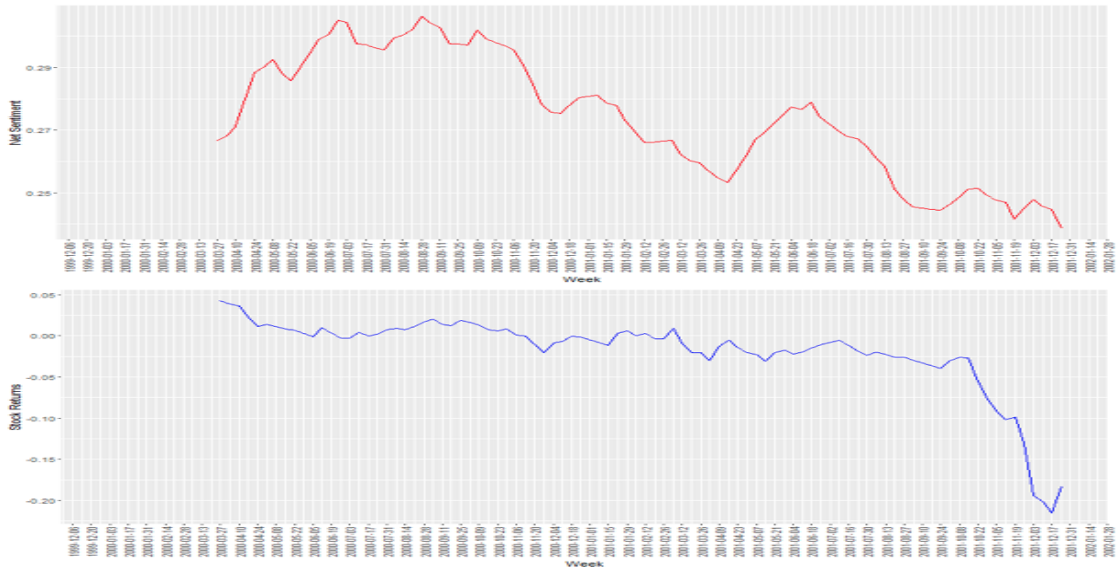

Exhibit 6: *Returns and sentiment.* This exhibit plots the 13-week (i.e., quarterly) moving-average net sentiment index (red line) along with the 13-week moving average stock returns (blue line) on a weekly basis, for the period spanning January 2000 through December 2001. We aggregate emails on a weekly basis and parse the content to determine the extent of positive, negative, and ambiguous sentiment. Net Sentiment is then calculated as $\frac{Pos-Neg}{Pos+Neg}$ and $1-\frac{|Pos-Neg|}{Pos+Neg}$.

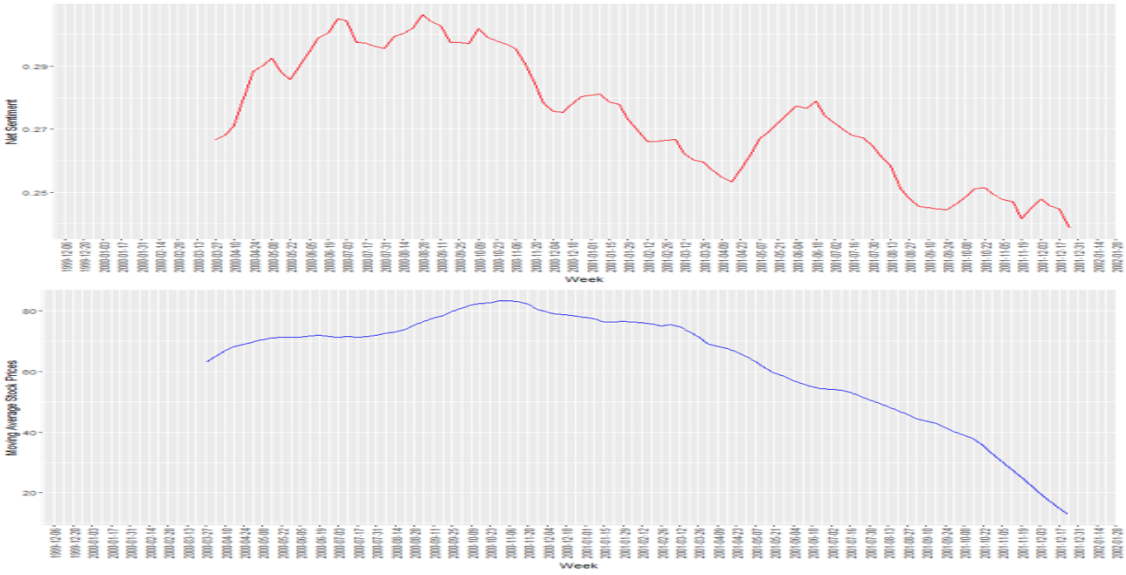

Exhibit 7: *Prices and sentiment.* This exhibit plots the 13-week (i.e., quarterly) moving-average net sentiment index (red line) along with the 13-week moving average stock prices (blue line) on a weekly basis, for the period spanning January 2000 through December 2001. We aggregate emails on a weekly basis and parse the content to determine the extent of positive, negative, and ambiguous sentiment. Net Sentiment is then calculated as $\frac{Pos-Neg}{Pos+Neg}$ and $1-\frac{|Pos-Neg|}{Pos+Neg}$.

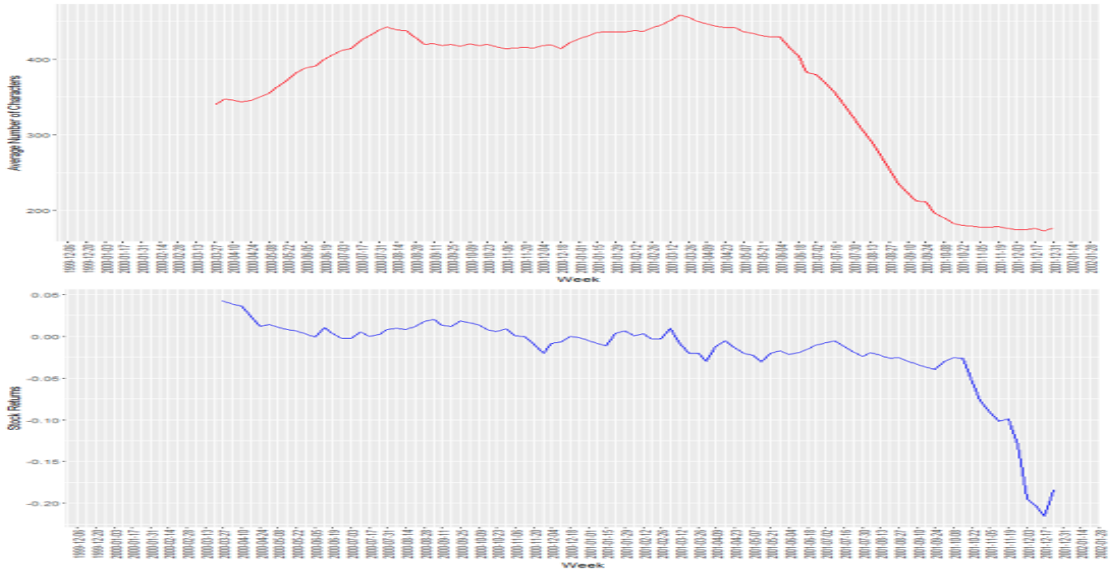

Exhibit 8: *Returns and email length.* This exhibit plots the 13-week (i.e., quarterly) moving-average stock returns (blue line) along with the 13-week moving average email length (red line) on a weekly basis, for the period spanning January 2000 through December 2001. We aggregate emails on a weekly basis and parse the content to determine the average number of characters over time.

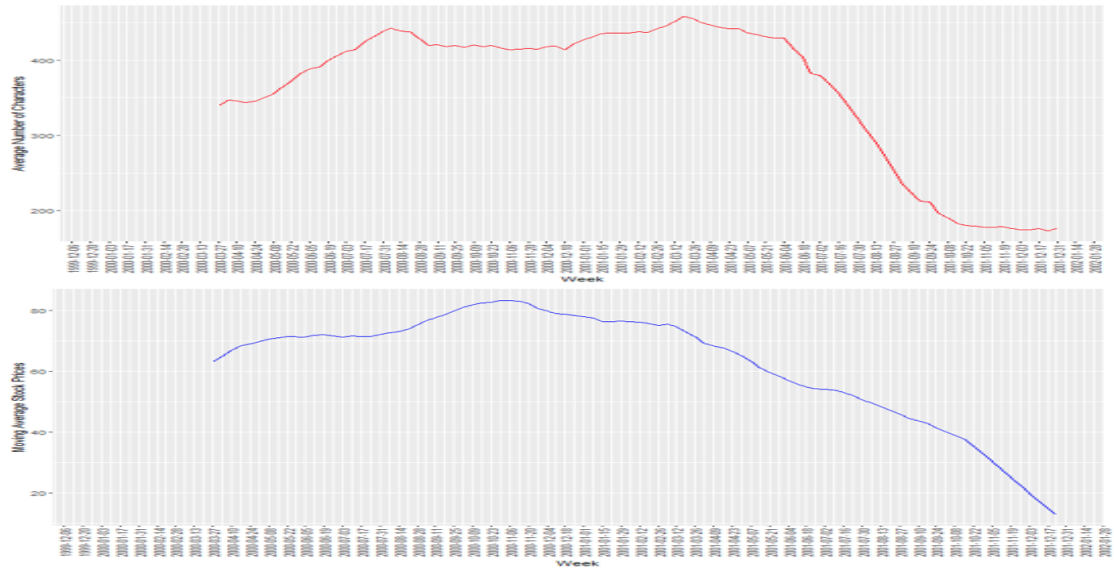

Exhibit 9: *Moving average prices and email length.* This exhibit plots the 13-week (i.e., quarterly) moving-average stock prices (blue line) along with the 13-week moving average email length (red line) on a weekly basis, for the period spanning January 2000 through December 2001. We aggregate emails on a weekly basis and parse the content to determine the average number of characters over time.

Exhibit 10: *Sentiment and Returns: Parsing Email Content.* This exhibit presents estimates from the following time-series regression (see equation 1): $\text{Returns}_t = \alpha + \beta_1 \cdot \text{MA Email Sentiment}_t + \beta_2 \cdot \text{MA Email Length}_t + \beta_3 \cdot \text{MA Total Emails}_t + \epsilon_t$. $\texttt{Returns}_t$, the dependent variable, is the stock return for Enron Corporation in week t. $\texttt{Email Sentiment}_t$ is the average net sentiment across all emails sent in week t; $\texttt{Email Length}_t$ is the average number of characters per email across all emails sent in week t (scaled by 1,000); and $\texttt{Total Emails}_t$ is the total number of emails sent in week t (scaled by 1,000). For all independent variables, we take the 13-week (i.e., quarterly) moving average on a weekly rolling basis. That is, for $\texttt{MA Email Sentiment}_t$, we calculate the average $\texttt{Email Sentiment}_t$ from week $t-12$ through week t, inclusive. T-statistics are presented below in parentheses, and significance levels at 10%, 5%, and 1% are denoted by *, **, and ***, respectively.

Variable	Coefficient Estimate (t-statistic)			
	(1)	(2)	(3)	(4)
$\texttt{MA Email Sentiment}_t$	2.347***	0.575	2.330***	-1.397
	(3.27)	(0.63)	(3.14)	(-1.25)
$\texttt{MA Email Length}_t$		0.584***		1.046***
		(2.97)		(4.19)
$\texttt{MA Total Emails}_t$			-0.004	-0.131***
			(-0.10)	(-2.83)
Intercept	-0.680***	-0.406*	-0.671***	0.117
	(-3.45)	(-1.93)	(-3.08)	(0.43)
Adjusted R^2	0.10	0.18	0.09	0.24
No. of observations	88	88	88	88

Note: the results reflect the same significant variables when all RHS variables are used at one lag, i.e., they predict returns.

To explore the predictability and significance of these factors in a multivariate setting, we estimate the following time-series regression:

$$\begin{aligned} \text{Returns}_t &= \alpha + \beta_1 \cdot \text{MA Email Sentiment}_t + \beta_2 \cdot \text{MA Email Length}_t \\ &\quad + \beta_3 \cdot \text{MA Total Emails}_t + \epsilon_t \end{aligned} \tag{1}$$

$\texttt{Returns}_t$, the dependent variable, is the stock return for Enron in week t. $\texttt{Email Sentiment}_t$ is the average net sentiment across all emails sent in week t, where net sentiment is calculated as the difference between the number of positive and negative words scaled by the sum of positive and negative words. $\texttt{Email Length}_t$ is the average number of characters per email across all emails sent in week t. $\texttt{Total Emails}_t$ is the total number of emails sent in week t. For each of the independent variables, we take the 13-week (i.e., quarterly) moving average on a rolling weekly basis. That is, for $\texttt{MA Email Sentiment}_t$, we calculate the average $\texttt{Email Sentiment}_t$ from week $t-12$ through week t, inclusive. For ease of interpretation and aesthetic appeal, we scale $\texttt{Email Length}_t$ and $\texttt{Total Emails}_t$ by $1,000$.

The results, which we present in **Exhibit 10**, show that the net sentiment of email content is a meaningful predictor of subsequent stock returns (Column 1). Specifically, a one standard-deviation (i.e., 0.019) decrease in the net sentiment gleaned from emails is

associated with a 4.5% decline in stock returns (coefficient estimate = 2.347, t-statistic = 3.27). However, this positive relation is eroded with the inclusion of the moving average email length (Column 2). Specifically, the coefficient estimate on $\texttt{MA Email Sentiment}_t$ decreases to 0.575 (t-statistic = 0.63), and is no longer statistically significant. However, $\texttt{MA Email Length}_t$ remains a statistically significant determinant of stock returns, with a coefficient estimate of 0.584 (t-statistic = 2.97), which translates to a 1.17% decline in stock returns for a 20-character decline in the 13-week moving average email length. Intuitively, as corporate risk and malfeasance escalate, emails become increasingly shorter, as employees are less likely to include specific details in emails sent via the corporate server.

We also re-estimate the regression in **Exhibit 10** with all independent variables at one lag (i.e., time $t-1$) to investigate whether they form a predictive relationship for returns. We find that exactly the same variables are significant in each of the four regressions, suggesting that email characteristics lead and predict returns.

NEWS CONTENT AND STOCK RETURNS OVER TIME

We now proceed to explore predictive content contained in the weekly net sentiment gleaned from Factiva news articles relative to the net sentiment conveyed by employee email structure and content over time. As before, we calculate moving average based on a 13-week (i.e., quarterly) look back period, and we re-estimate regression equation (1), this time adding the moving average net sentiment of news content as an additional regressor. We explore two types of news-based sentiment: (i) the net sentiment gleaned from the bodies of the articles, and (ii) the net sentiment based on the article headers.

The results, which we present in **Exhibit 11**, show that the net sentiment of news body content is also a meaningful predictor of subsequent stock returns (Panel A, Column 1), though the net sentiment contained in the news header alone is not (Panel B, Column 1). Furthermore, the net sentiment conveyed by news body content has greater predictive power than the net sentiment contained in employee emails (Panel A, Column 2); though, this may at least in part be caused by the redactions applied to the Enron corpus, which we broached in an earlier section. Irrespective, we see that email length continues to dominate with regard to the information it conveys about subsequent returns (Panel A, Columns 3 and 5).

With regard to the relative timeliness of information content in employee emails as opposed to news articles, we examine the correlations between email-based sentiment and news-based sentiment at various lags. With respect to the net sentiment gleaned from the bodies of news articles, which we present in **Exhibit 12**, we observe that contemporaneous correlation is highest, with the decline in correlation tapering off at approximately a 10-week lag (i.e., time $t-10$ email sentiment against time t news sentiment). Thus, email content does not appear to meaningfully lead the news, though this result may be due to the aforementioned redactions of the Enron corpus (whereas news archives do not undergo such expunging). We do, however, observe that the net sentiment from emails has a slight lead on the net sentiment from news headers, presented in **Exhibit 13**, whereby the correlation initially increases with the numbers of lags before tapering off and declining as time goes on.

The fact that email length is the best predictor of poor performance is interesting in reflecting that email size matters more than email content, and more than news content. This suggests that simple quantification of email traffic may be useful and more complicated metrics are not needed. How much people talk is less likely to be manipulated, and is more

Exhibit 11: *Sentiment and Returns: Parsing News Content.* This exhibit estimates the same regression equation as in Exhibit 10 (see 1)), this time adding the moving average net sentiment of news content as an additional regressor. As before, $\texttt{Returns}_t$, the dependent variable, is the stock return for Enron Corporation in week t. $\texttt{Body Sentiment}_t$ (Panel A) is the average net sentiment across all Factiva news articles released in week t, and $\texttt{Header Sentiment}_t$ (Panel B) is the average net sentiment across the headers of all Factiva news articles released in week t. All other variables are as specified in Exhibit 10.

Variable	Coefficient Estimate (t-statistic)				
	(1)	(2)	(3)	(4)	(5)
Panel A. News Body Sentiment and Returns					
$\texttt{MA Body Sentiment}_t$	1.410***	1.501**	0.657	1.505**	-0.827
	(3.95)	(2.49)	(0.87)	(2.48)	(-0.92)
$\texttt{MA Email Sentiment}_t$		-0.245	0.377	-0.284	-1.293
		(-0.19)	(-0.29)	(-0.22)	(-1.02)
$\texttt{MA Email Length}_t$			0.486*		1.380***
			(1.81)		(3.34)
$\texttt{MA Total Emails}_t$				-0.009	-0.164***
				(-0.24)	(-2.77)
Intercept	-0.711***	-0.688***	-0.426*	-0.668***	0.399
	(-4.18)	(-3.27)	(-1.69)	(-2.94)	(1.04)
Adjusted R^2	0.15	0.14	0.17	0.13	0.23
No. of observations	81	81	81	81	81
Panel B. News Header Sentiment and Returns					
$\texttt{MA Header Sentiment}_t$	-0.795	-1.136*	-0.772	-1.210**	-0.893
	(-1.31)	(-1.96)	(-1.34)	(-2.03)	(-1.61)
$\texttt{MA Email Sentiment}_t$		2.628***	0.705	2.566***	-1.254
		(3.30)	(0.66)	(3.18)	(-1.03)
$\texttt{MA Email Length}_t$			0.560**		1.026***
			(2.59)		(3.93)
$\texttt{MA Total Emails}_t$				-0.024	-0.138***
				(-0.59)	(-2.91)
Intercept	0.307	-0.256	-0.096	-0.178	0.485
	(1.15)	(-0.84)	(0.75)	(-0.54)	(1.39)
Adjusted R^2	0.01	0.12	0.18	0.11	0.25
No. of observations	81	81	81	81	81

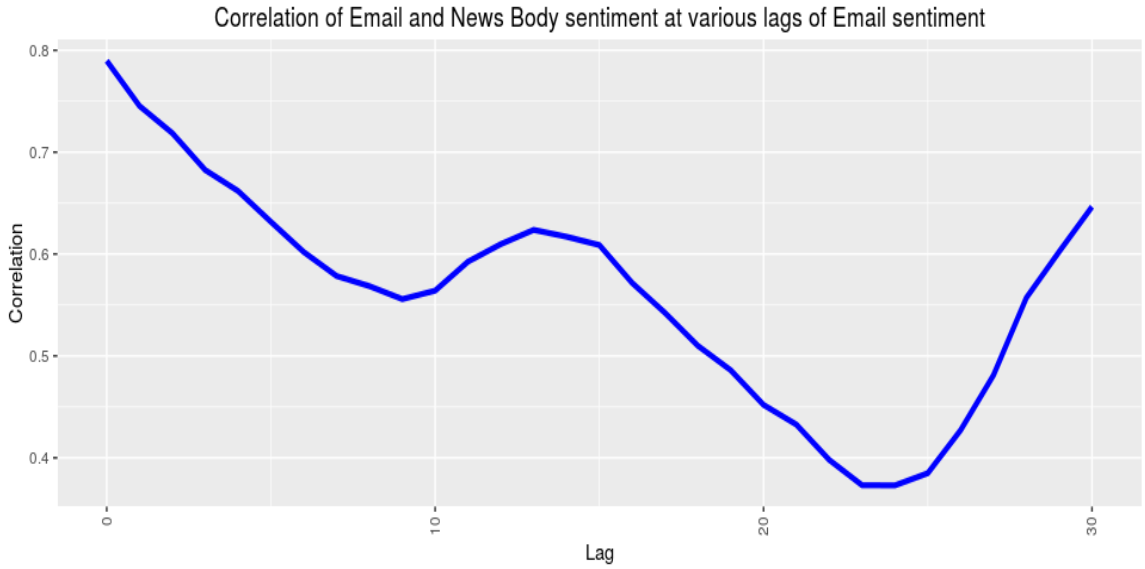

Exhibit 12: *Correlation between Email Sentiment and News Body Sentiment.* This exhibit plots the correlation between net sentiment from emails and net sentiment from the bodies of news articles at various lags of email net sentiment.

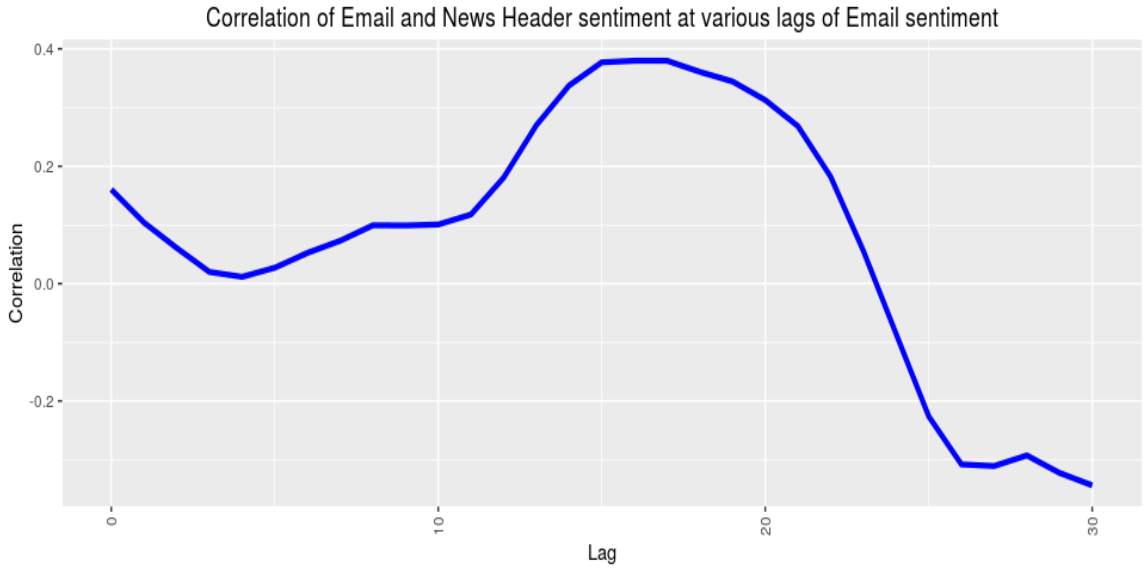

Exhibit 13: *Correlation between Email Sentiment and News Header Sentiment.* This exhibit plots the correlation between net sentiment from emails and net sentiment from the headers of news articles at various lags of email net sentiment.

important than what they say.

SPATIAL ANALYSIS EXTENSIONS

Thus far, we have observed that even a simple construct such as sentiment, based on positive and negative word dictionaries, has predictive implications for the viability of a corporate firm such as Enron, which suggests that the language in emails and news offer important insights into the fortunes or escalating risk of a corporation. As we have seen, email length is even better at forecasting stock returns. In addition to the quantity of text, and tone conveyed by language, other spatial features of communication may be useful to explore as complementary analyses. We now proceed to illustrate a few examples as possible future methods for consideration in RegTech systems.

In this regard, we explore three ideas. First, we consider whether the structure of communication among top officers of Enron offers revealing insights into management behavior in times of impending crisis. To explore this aspect, we examine at the network of email communication. Second, we develop an interactive system to drill down into vocabulary trends. Third, we offer topic analysis as a possible means to map evolving corporate fortunes.

Email Network and Degree Distribution over Time

Here, we construct a network graph of employees' email interactions over time based on sent emails. We observe a marked shift in connectedness from the fourth quarter of 2000 to the fourth quarter of 2001 (**Exhibit 14**). Accordingly, we see a much flatter and more evenly apportioned degree distribution from Q4 of 2000 to Q4 of 2001 (**Exhibit 15**). That is, each node (i.e., employee) has more degrees (i.e., direct email connections) in the latter period. As Enron proceeded to its demise, the connectivity through emails of top officers in the firm became more intense and widespread. Combining this information with the results of an earlier section, we note that this greater interaction is coupled with shorter emails.

Vocabulary Trends

Words used in good times may be different from those used in bad times. To explore these shifts, we build an interactive word analyzer, similar to that provided by Google Trends,[17] to plot the relative frequency of a particular word so as to examine how vocabulary usage trends over time. To illustrate, we present the trend plots for two words, "profits" and "losses", normalized by the total number of emails over time. See **Exhibit 16**. The word "profits" declines steadily as Enron slipped into failure in the last quarter of 2001, and the word "losses" spikes just before it declares bankruptcy in October 2001. Sometimes, the variation in the usage of a key word over time may be extremely revealing especially if it tends to spike just prior to an impending crisis.

Topics

Topic analysis was developed in a seminal paper by Blei et al. (2003), who proposed a technique known as Latent Dirichlet Allocation (LDA) to decompose word correlations and

[17] https://www.google.com/trends/

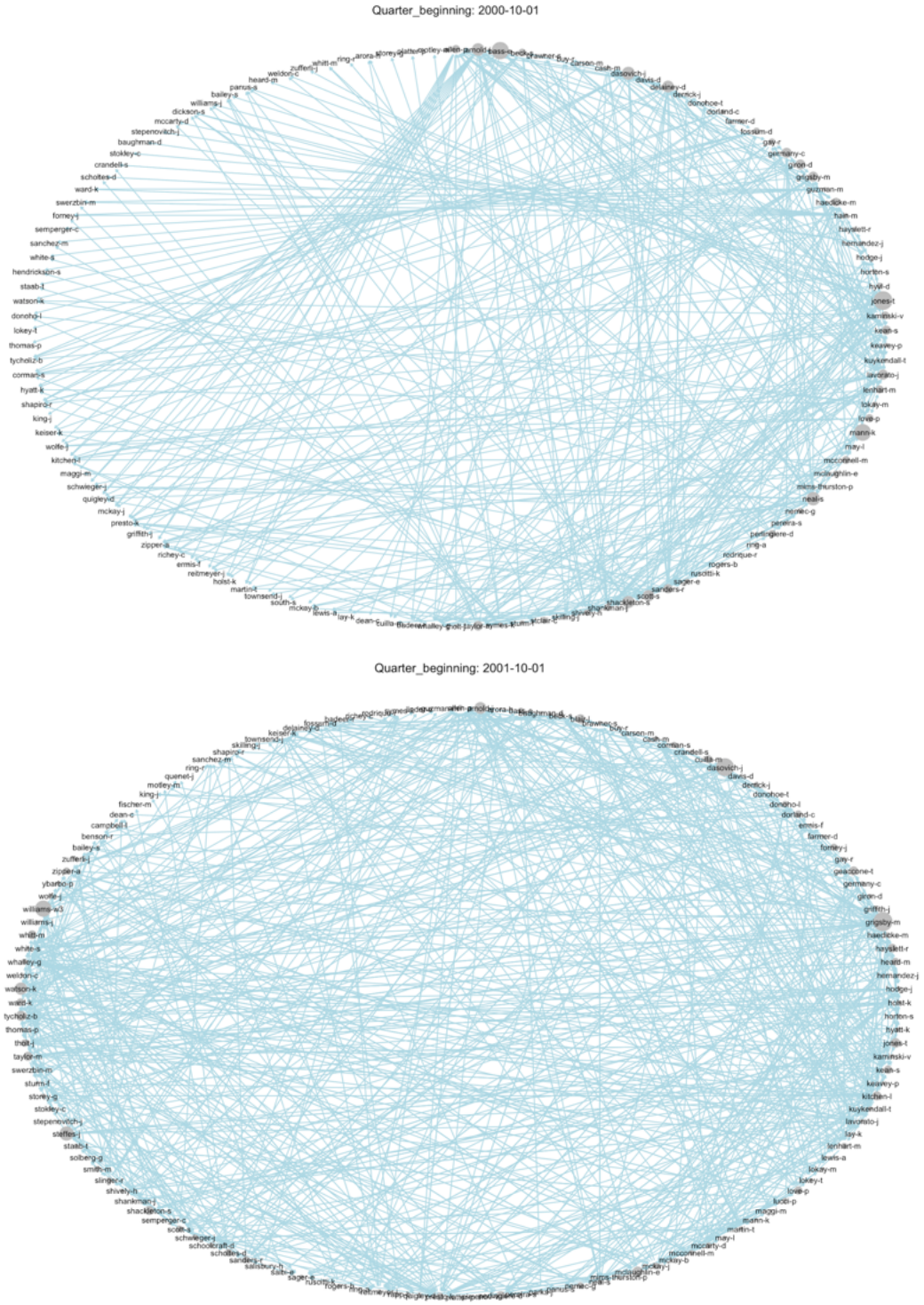

Exhibit 14: *Email networks.* This exhibit plots the email networks for Q4, in years 2000 and 2001. The size of the circle plotted for each node (i.e., employee) is proportional to the number of connections it has in the network, i.e., its "degree".

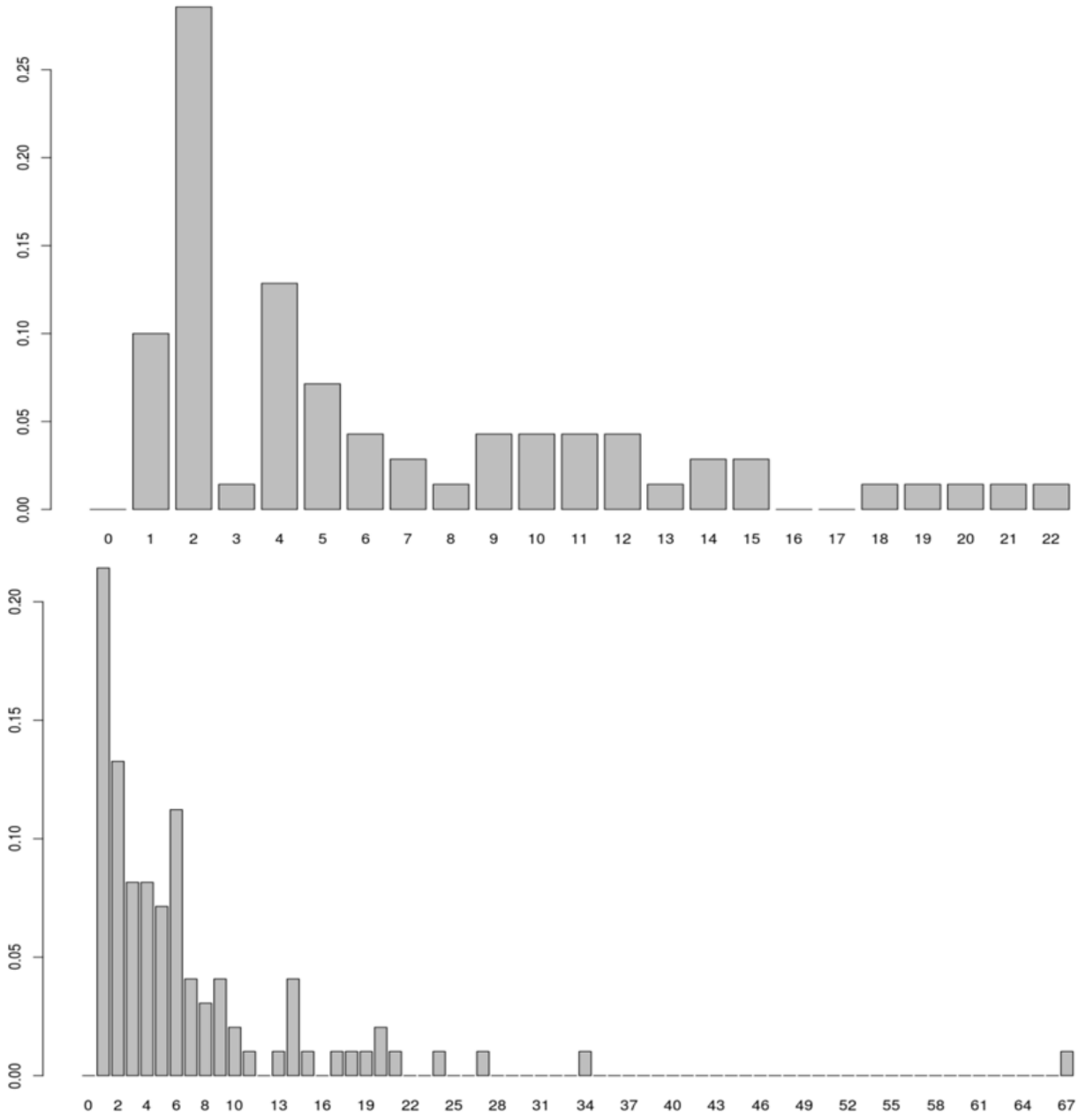

Exhibit 15: *Degree distribution.* This exhibit shows the email networks' degree distribution for Q4, in years 2000 and 2001. The number of connections a node (i.e., employee) has in the network is its "degree".

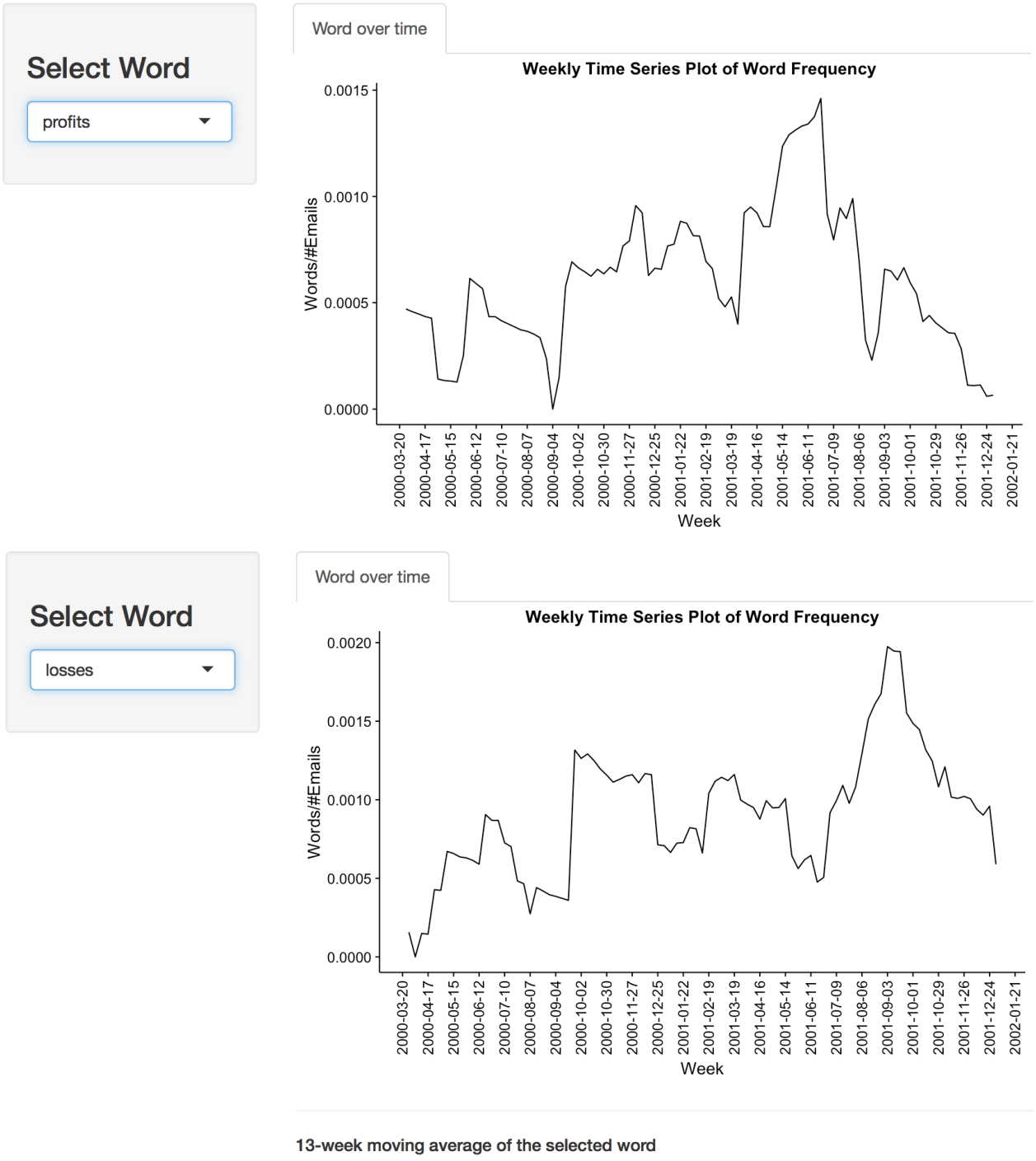

Exhibit 16: *Vocabulary trend plots.* We plot the word frequency of two words to illustrate an interactive system that allows the user to explore the weekly usage of a word over time. The word frequency is normalized by the number of emails each week. The system allows entry of a few letters and then shows the choice words so as to allow the user to select a word for analysis. The screen shots are displayed here.

document correlations into topics. Earlier antecedents on model-based clustering were developed by Pritchard et al. (2000). This decomposition may then be used to relate topics to words and documents. There is now a large literature on using topic analysis in natural language applications.[18]

We may think of LDA as a probability model that connects documents with words and topics. The topics are latent. Assume a vocabulary of V words $\{w_1, ..., w_i, ..., w_V\}$, indexed by i, and in this framework a document is a vector of N words. At a higher level, a corpus is a collection of M documents, $\{d_1, ..., d_j, ..., d_M\}$, indexed by j. These constructs have to be connected to a preset number of K topics, $\{t_1, ..., t_l, ..., t_K\}$, indexed by l.

The starting point for topic analysis is a matrix representation of the documents as word vectors. Any document is a collection of words and a collection of documents is denoted as a corpus. A document may be represented by a vector of size equal to the number of unique words in the corpus. The vector representation of a document is captured by the frequency count of each word in the document. These document vectors are sparse vectors as only a small number of the total number of unique words in the corpus appears in each document. A collection of all documents is represented by what is known as a Term Document Matrix or TDM for short. This matrix has all unique words (terms) on the rows and documents on the columns. The transpose of this matrix is known as the Document Term matrix (DTM).

Topics introduce a third dimension to documents and terms. Our exposition here of topc modeling using LDA parallels Blei et al. (2003). We may include topics through a decomposition of the DTM into two matrices: (i) a matrix $A \in \mathcal{R}^{M \times K}$, with documents on the rows and topics on the columns, and (ii) a matrix $B \in \mathcal{R}^{K \times V}$, with topics on the rows and terms on the columns. For matrix A, we would require that the row sums equal 1, i.e., that the probability of a given document being assigned to any topic adds up to 1, where all elements of the matrix are probabilities and are therefore greater than 0 but less than 1. For matrix B, we also require that all elements are probabilities, lie between zero and one, and the row sums equal 1, i.e., for a given topic, the sums of probabilities of all words in that topic adds up to 1. The probabilities in matrices A and B need to be calibrated to data in the DTM. This means decomposing the DTM into matrice A and B, along with making sure the entries in these matrices are probabilities, which requires specifying a functional form for these latent probabilities.

Topics are latent and the probability of a topic mixture $\theta \in R^K$ is governed by a Dirichlet distribution, hence the method is called LDA. Actual topics x are drawn from the mixture distribution, i.e., $p(x_t|\theta)$. The Dirichlet density function for the topic mixture θ is

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{t=1}^{K} \alpha_t)}{\prod_{t=1}^{K} \Gamma(\alpha_t)} \prod_{t=1}^{K} \theta_t^{\alpha_t - 1}$$

where $\alpha_t, t = 1...K$ are parameters to be calibrated and $\Gamma(\cdot)$ is the gamma function. Since matrix A contains the conditional probability of a topic for a document, i.e., $p(\theta|d)$ and matrix B gives the conditional probability of a word in a given topic, $p(w|\theta)$, we can combine these probabilities to write out the likelihood function for the entire corpus of documents, denoted D.

$$p(D) = \prod_{j=1}^{M} \int p(\theta_j|\alpha) \left(\prod_{t=1}^{K} \sum_{x_{jt}} p(x_t|\theta_j) p(w_t|x_t) \right) d\theta_j$$

[18] See Kar et al. (2015); Pavlinek and Podgorelec (2017); Yeh et al. (2016); Zhang et al. (2016).

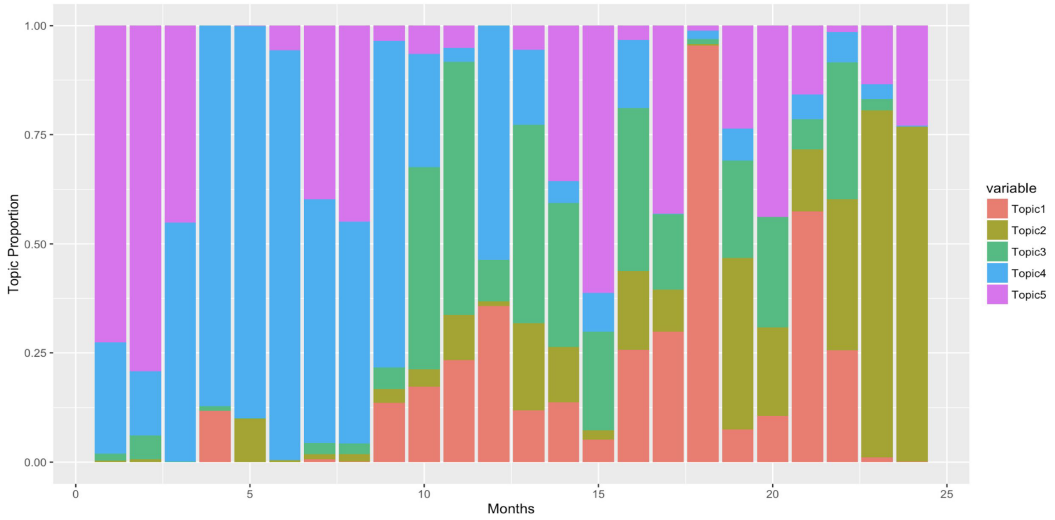

Exhibit 17: *Topic contribution by month.* We display the time distribution of the top five topics from a distillation of 1,302 news articles on Factiva, from PR Newswire (US). The topic contribution is shown for each of the 24 months in the period spanning January 2000 to December 2001. We also calculate the normalized sentiment from the words in each topic. The sentiment scores for the five topics are: $\{0.40, -0.47, 0.93, 0.66, -1.52\}$, i.e., topic 3 has the highest (most positive) sentiment, and topic 5 has the lowest (most negative) sentiment.

This likelihood is the product of likelihoods of all M documents. The likelihood of each document is an integral of document probabilities for each possible topic, coming from matrix A. Within this, The probability of a word coming from a topic is $p(w_t|x_t)$ and comes from matrix B. These are all combined in the equation above. Calibration is undertaken using Markov Chain Monte Carlo (MCMC) methods to maximize likelihood $p(D)$. There are many efficient methods for this, and we implemented this using the `gensim` package in the Python programming language.

Using the two years of news articles we extracted from Factiva, we extract the top five topics, and in **Exhibit 17**, we plot the topic contribution for each of 24 months from 2000 through 2001. We also use the words in each topic to compute the normalized sentiment for each topic. We find that topics 1, 3, and 4 have positive connotations, whereas topics 2 and 5 have negative connotations. Toward the end of the sample period, as Enron slipped into bankruptcy, topics 2 and 5 become the salient contributors. Overall, trends in topic contribution provide a visually appealing rendition of the evolving fortunes of the firm.

CONCLUSION

In this paper, we develop a RegTech expert system solution to parse corporate email content to detect shifts in critical characteristics in a timely, efficient, and non-invasive manner. We find that the net sentiment extracted from sent email content substantially predicts increased risk and malaise, and we find that other structural characteristics are also strong predictors of the declining health of a firm. We also explore spatial structure/network-based indicators of shifting risk, vocabulary trends, and topic analysis, as complementary analyses to our time-series analysis of email- and news-based frequency, length, and sentiment. Such avenues are particularly ripe for exploration, since the sentiment conveyed directly in email

content itself is much easier to suppress than email length or the connectivity or shifting sub-connectivity within an entire network. This type of textual RegTech analysis may be used by corporate management as a means to detect risk in a timelier fashion. It may also be used by regulators in their audit process, as they can requisition such analyses from firms, without intrusively reading emails. All told, RegTech appears to be a promising avenue for applied linguistics.

Appendix: Analytics processing pipeline

Our entire data analytics pipeline for this expert system is portrayed in **Exhibit** 18. A description of this is as follows.

1. `Enron_Email_Analysis_S1.Rmd` : reads in the data and provides some basic descriptive statistics. Output is stored in `AllMails.RData` (1.3GB). Contains a single data frame EnMails with user name, mail folder, full message. (517,395 x 3).

2. `Enron_Email_Analysis_S2.Rmd` : Reads in `AllMails.RData` and then extracts the text (message body) of all emails and stores them in a new data frame. Creates columns for the from, to, cc, bcc, etc. Creates a file `AllMails2.Rdata` (550MB). (517,394 x 18). Has the following columns in the EnMails data frame, which is now overwritten from the previous version: {`User, MailBox, RawMessage, Date, Year, Month, Week, Quarter, MessageBody, CharCnt, from, toList, toCount, ccList, ccCount, bccList, bccCount, totalCount`}. The column `RawMessage` is converted to `MessageBody` which is clean text. At this stage we still have all emails. No de-duplication has been done so far.

3. `Enron_Email_Analysis_S3.Rmd` : Reads in `AllMails2.RData`. Does further clean up for users who have more than one email account. (Lots of exception handling here, hope this is not the case with better email data.) Creates summary statistics. Then only accesses the Sent mails to prevent duplication of emails across email boxes. Creates plots of emails over time, mainly for the years 2000 and 2001 (Enron failed in Oct 2001). Does mood scoring of the message body. By week and user. Sentiment and Disagreement. Calculates weekly returns, uses `ENE_daily.csv`. Regressions of sentiment and returns. Wordclouds (though not very informative). Network analysis, creates list of users for network analysis. Saves EnMails DF and Weekly Data for Regressions to `AllMails3.RData`. (113,263 x 19, 57MB). Creates `MoodScoredDf.RData` which contains all the mood scoring statistics.

4. `Enron_Email_Analysis_S3_2.Rmd` : Does network analysis. Reads in `AllMails3.RData`. Creates adjacency matrix and plots by week. Creates weekly stats. And plots degree distribution.

5. `Enron_Email_Analysis_S4.Rmd` : Reads in `AllMails3.RData`. Creates list of 1000 most important words over time, for further use in Shiny app. Intersects with the word lists so we only get dictionary words and not nonsense compound words. Creates term-document matrix: `save(tdm2, freq1000, poswords, negwords, uncwords, weekNames, weeklyMailCount, file="AllMails4_tdm.RData")`. $\dim(\text{tdm2}) = 76518 \times 105$, words x weeks. 1353 negwords, 288 poswords, 206 uncwords, 105 weeks.

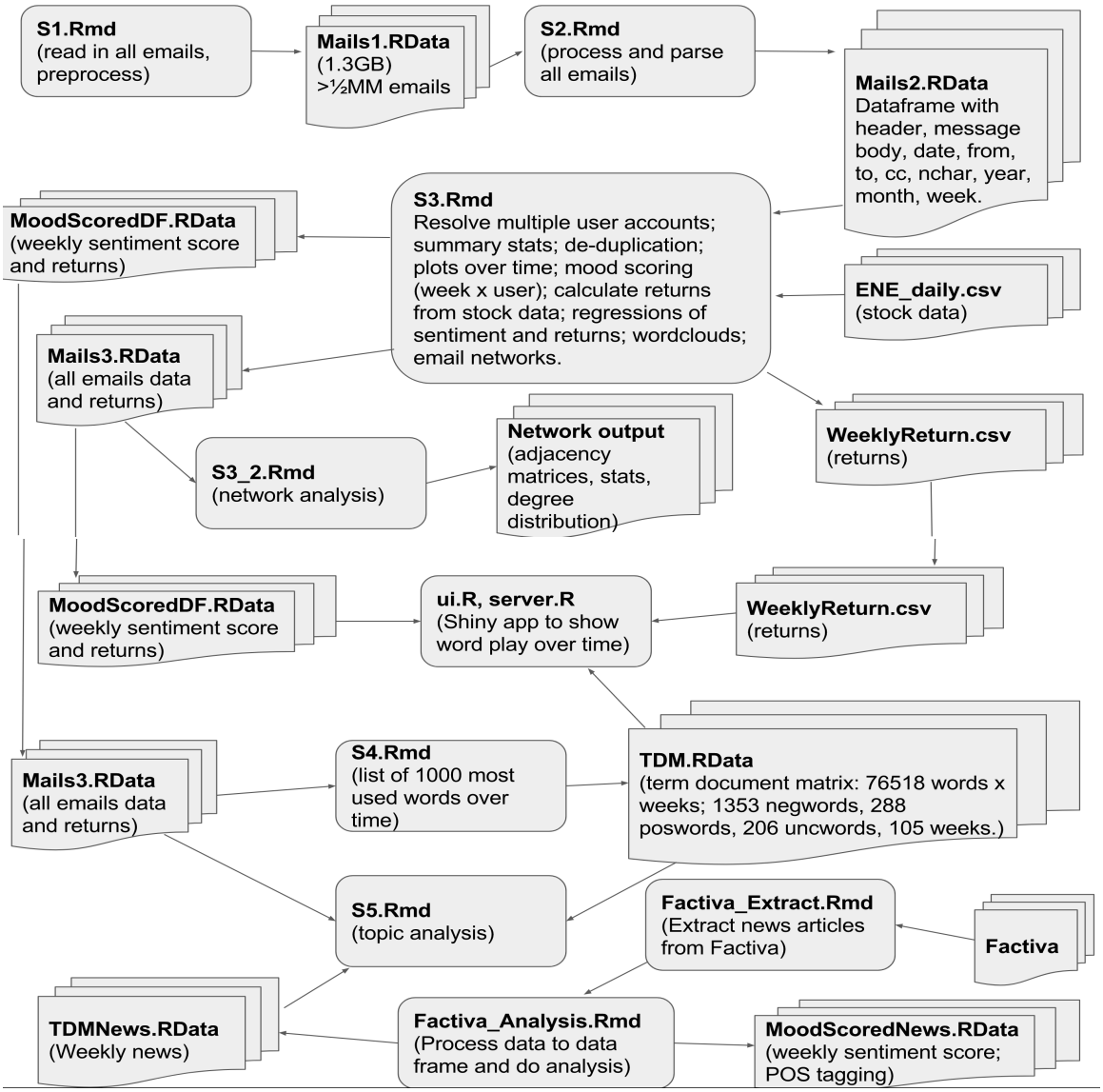

Exhibit 18: *Expert System Pipeline.* This graphic shows the data sources, program flow, and output at each stage of the process of mining emails.

6. `ui.R` and `server.R` : `RShiny` app to display words over time. And some plots with correspondence of words with sentiment and returns. Uses `AllMails4_tdm.Rdata` and `MoodScoredDf.Rdata`. Also `WeeklyRet.csv`.

7. `Enron_Email_Analysis_S5.Rmd` : Topic analysis. Uses `AllMails3.Rdata` and `AllMails4_tdm.Rdata`. Plots topic share, and also sentiment scores the top 20 words in each topic.

8. `Enron_Factiva_Analysis.Rmd` : Reads in `EnNews.RData` (created by `Enron_Factiva_Extract.Rmd` – this gives a data frame with {`Date, Author, Header, Body`}). Then the Analysis program proceeds to work on this data frame and adds more columns. Creates `MoodScoredDF_News.RData`. Creates `TDMWklyArt.RData`. The file is an entire program to do all aspects of text analytics: creating the text Body, sentiment extraction, wordclouds, topic modeling, POS tagging. POS tagging is a new analysis that is being tried out. See "The Secret Life of Pronouns: What Our Words Say About Us" by Pennebaker (2011).

Acknowledgements

We thank Claire Brennecke, Magnus Mähring, Manju Puri, Jack Reidhill, and seminar participants at Korea Capital Market Institute (KCMI) and the R/Finance 2017 Meeting for discussions on the topic of email analysis.

References

Aiden, E. and Michel, J.-B. (2013). *Uncharted: Big Data as a Lens on Human Culture*. Penguin, New York.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance*, 59(3):1259–1294.

Arner, D. W., Barberis, J. N., and Buckley, R. P. (2017). Fintech, regtech and the reconceptualization of financial regulation. *Northwestern Journal of International Law & Business*, 37(3):371–413.

Atsalakis, G. S. and Valavanis, K. P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications*, 36(7):10696 – 10707.

Back, B., Laitinen, T., and Sere, K. (1996). Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications*, 11(4):407 – 413. The Third World Congress on Expert Systems.

Bar-Haim, R., Dinur, E., Feldman, R., M, F., and Goldstein, G. (2011). Identifying and following experts in stock microblogs. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1310–1319.

Berutich, J. M., López, F., Luna, F., and Quintana, D. (2016). Robust technical trading strategies using gp for algorithmic portfolio selection. *Expert Systems with Applications*, 46(Supplement C):307 – 315.

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

Bodnaruk, A., Loughran, T., and McDonald, B. (2013). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2013). Which news moves stock prices? a textual analysis. *NBER Working Paper 18725*.

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2018). Information, trading, and volatility: Evidence from firm-specific news. *Review of Financial Studies*, 32(3):992–1033.

Boyacioglu, M. A. and Avci, D. (2010). An adaptive network-based fuzzy inference system (anfis) for the prediction of stock market return: The case of the istanbul stock exchange. *Expert Systems with Applications*, 37(12):7908–7912.

Brown, E. (2012). Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. *Proceedings of the Southern Association for Information Systems Conference*.

Burdick, D., Das, S., Hernandez, M. A., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I., and Vaithyanathan, S. (2011). Extracting, linking and integrating data from public sources: A financial case study. *IEEE Data Engineering Bulletin*, 34(3):60–67.

Das, S. (2016). Matrix metrics: Network-based systemic risk scoring. *Journal of Alternative Investments, Special Issue on Systemic Risk*, 18(4):33–51.

Das, S., Martinez-Jerez, A., and Tufano, P. (2005). einformation: A clinical study of investor discussion and sentiment. *Financial Management*, 34(3):103–137.

Das, S. R. (2014). Text and context: Language analytics in finance. *Foundations and Trends in Finance*, 8(3):145–261.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388.

Greenwood, V. (2016). Still waiting for that email? *Scientific American*, March 1, online.

Guresen, E., Kayakutlu, G., and Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, 38(8):10389 – 10397.

Halliday, M. and Hasan, R. (1976)). *Cohesion in English*. Longman, London.

Hevner, A., March, S., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems*, 28(1):75–105.

Huang, K. Y. and Jane, C.-J. (2009). A hybrid model for stock market forecasting and portfolio selection based on arx, grey system and rs theories. *Expert Systems with Applications*, 36(3):5387–5392.

Iivari, J. (2007). A paradigmatic analysis of information systems as a design science. *Scandinavian Journal of Information Systems*, 19(2):39–64.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Kar, M., Nunes, S., and Ribeiro, C. (2015). Summarization of changes in dynamic text collections using latent dirichlet allocation model. *Information Processing & Management*, 51(6):809 – 833.

Kim, M.-J. and Kang, D.-K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications*, 37(4):3373 – 3379.

Lee, W.-S., Tzeng, G.-H., Guan, J.-L., Chien, K.-T., and Huang, J.-M. (2009). Combined mcdm techniques for exploring stock selection based on gordon model. *Expert Systems with Applications*, 36(3):6421–6430.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69(4):1643–1671.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Lu, H., Chen, H., Chen, T., Hung, M., and Li, S. (2010). Financial text mining: Supporting decision making using web 2.0 content. *IEEE Intelligent Systems*, 25(2):78–82.

Markus, M. L., Majchrzak, A., and Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. *MIS Quarterly*, 26(3):179–212.

Min, J. H. and Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28(4):603–614.

Mizumoto, A. and Plonsky, l. (2015). R as a lingua franca: Advantages of using r for quantitative research in applied linguistics. *Applied Linguistics*, 37(2):284–291.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, abs/1103.2903:93–98.

Pavlinek, M. and Podgorelec, V. (2017). Text classification method based on self-training and lda topic models. *Expert Systems with Applications*, 80(Supplement C):83 – 93.

Pennebaker, J. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, New York.

Pritchard, J. K., Stephens, M., and Donnely, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Propp, V. (1928). *Morphology of the Folk Tale. English trans. Laurence Scott.* University of Texas Press (English translation in 1968).

Rada, R. (2008). Expert systems and evolutionary computing for financial investing: A review. *Expert Systems with Applications*, 34(4):2232 – 2240.

Rao, T. and Srivastava, S. (2014). *Twitter Sentiment Analysis: How to Hedge Your Bets in the Stock Markets*, pages 227–247. Springer International Publishing, Cham.

Shin, K.-s. and Lee, Y.-J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, 23:321–328.

Shklovsky, V. (1917). *Art as Technique.* University of Nebraska Press.

Sprenger, T. (2011). Tweettrader.net: Leveraging crowd wisdom in a stock micro blogging forum. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 663–664.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2013). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

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

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

Tsai, C.-F. and Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4):2639 – 2649.

Tumarkin, R. and Whitelaw, R. F. (2001). News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51.

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications*, 43(C):82–92.

Van Aken, J. (2005). Management research as a design science: Articulating the research products of mode 2 knowledge production in management. *British Journal of Management*, 16(1):19–36.

Wang, G., Ma, J., Huang, L., and Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, 26(Supplement C):61 – 68.

Wang, Y.-F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications*, 22(1):33 – 38.

Yeh, J.-F., Tan, Y.-S., and Lee, C.-H. (2016). Topic detection and tracking for conversational content by using conceptual dynamic latent dirichlet allocation. *Neurocomputing*, 216(Supplement C):310 – 318.

Zhang, W., Clark, R. A., Wang, Y., and Li, W. (2016). Unsupervised language identification based on latent dirichlet allocation. *Computer Speech & Language*, 39(Supplement C):47 – 66.