What Matters Most? How Tone in Initial Public Offering Filings and Pre-IPO News Influences Stock Market Returns

April 8, 2015

Feuerriegel, Stefan, University of Freiburg, Platz der alten Synagoge, 79098 Freiburg, Germany, stefan.feuerriegel@is.uni-freiburg.de

Schmitz, Jonas Tobias, University of Freiburg, Platz der alten Synagoge, 79098 Freiburg, Germany, jonastobiasschmitz@gmx.de

Pröllochs, Nicolas, University of Freiburg, Platz der alten Synagoge, 79098 Freiburg, Germany, nicolas.proellochs@is.uni-freiburg.de

Neumann, Dirk, University of Freiburg, Platz der alten Synagoge, 79098 Freiburg, Germany, dirk.neumann@is.uni-freiburg.de

Abstract

Before shares of a company are sold to the general public on a security exchange for the first time, regulatory publication requirements force U. S. firms to file an initial public offering prospectus. While accounting information in IPO filings are closely studied by investors and analysts, research must also examine the textual content of these filings. Thus, we measure the proportion of uncertain language. As a result, this paper provides empirical evidence that soft content is a major driver of first-day stock market returns; however, a much stronger impact on stock market prices comes from the pre-IPO news tone. Interestingly, the more uncertainty words appear in pre-IPO news, the higher the following first-day stock market return. It seems that investors mainly focus on the chances of a company rather than on the risks.

Keywords: Initial public offering, filings, information processing, sentiment analysis, text mining.
1 Introduction

Companies are increasingly turning to stock markets to raise expansion capital to monetize the investments of early private investors. An initial public offering (IPO) refers to a type of public offering where shares of stock in a company are sold to the general public, for the first time, on a securities exchange. As a result of regulatory publication requirements, firms must file a prospectus with, e.g., the Securities and Exchange Commission (SEC) in the U.S. prior to its initial public stock offering (c.f. Kenney & Patton 2013a; Ritter & Welch 2002). Details of the proposed offering are disclosed to potential purchasers in the form of a detailed document known as a prospectus. This prospectus contains key figures from financial accounting, as well as information on current and future business models, potential opportunities and risks threatening the business. The prospectus is used as an instrument to remove or at least alleviate disadvantages stemming from information asymmetries among different investor groups. Obviously, the prospectus plays a major role in the formation of IPO prices, which ultimately determines the amount of capital that accrues from the IPO. Thus, all involved stakeholders in the IPO may need to understand how the prospectus affects the final IPO prices. While the underwriting company has a natural incentive to cast a positive light on their company, investors have a need to decipher the text messages for relevant signals on the future performance of the company.

The theory of market microstructure may fill the bill in explaining the relationship between prospectus and final IPO price as it refers to the branch of finance that is concerned with the details of how exchange occurs in markets. More specifically, market microstructure theory epitomizes “the study of the process and outcomes of exchanging assets under a specific set of rules” (O’Hara 1995). One crucial aspect refers to the role of information and how prices reflect them. Interestingly, market microstructure theory is still in its infancy, dealing with information which consists of text messages, such as the prospectus. The theory is, thus, nearly silent to explain how text messages are processed into prices (c.f. Jegadeesh & Wu 2013; Loughran & McDonald 2011; Tetlock et al. 2008).

Altogether, information from an IPO prospectus provides the basis for the decision making process for potential investors. Thus, it is logical consequence to study the information processing of investors facing IPO filings in the context of efficient electronic markets. Though IPO filings consist of a few quantitative facts, a large proportion accounts for qualitative information. While prior research focuses on quantitative facts, such as the number of offered shares, knowledge of the textual components is rare. Hence, this paper examines the statistical relationship between tone in IPO filings and the subsequent stock market performance.

As a main contribution to finance research, this paper sheds some light into information processing along IPO filings. The soft data contained in these filings conveys insights and potentially valuable information that is absent in traditional quantitative numbers. A recent study by Ferris et al. (2013) points out that “the extent to which management is confident about the success of its issue and the implications that such beliefs have on IPO pricing are largely ignored in the literature”. To close this gap, individual research questions addressing the content of IPO prospectuses, as well as contributions, are as follows.

Confirmatory Analysis: To what extent are stock market returns of firms driven by their IPO filing tone?

As a confirmatory analysis of previous research (Ferris et al. 2013; Loughran & McDonald 2013), we investigate how investors react to IPO prospectuses and analyze the information processing empirically by measuring its tone. We find that tone in IPO prospectuses correlates with first-day return. In fact, one standard deviation increase in the uncertainty tone measure is negatively linked to an increase in first-day returns by an economically significant 10.56%.

Research Question 1: How is the IPO performance affected by pre-IPO news tone?

The confirmatory analysis proves that effect IPO filings have on the subsequent stock market returns. However, the media coverage of firms going public occurs in more depth, since journalists also report on the progress of the IPO, the previous performance of the firm and the general outlook. Because of these reasons, we
investigate the impact of pre-IPO news announcements on first-day returns and, finally, compare the influence of tone in IPO filings and news announcement in terms of strength.

Research Question 2: Which positive and negative words account for the tone in IPO filings?

In addition to the previous research question, we aim at identifying the driving forces behind the decision-making when investing in initial public offerings. While previous research looks merely in the words appearing most commonly, we detect words that usually serve as an indicator for underpricing.

The remainder of this paper is structured as follows. In Section 2, we combine previous research on both initial public offerings and approaches for measuring tone in order to study literature on their intersection, which deals with the information processing theory of IPO filings. To gauge the tone of these IPO filings, we detail our research model in Section 3. Using this research model, Section 4 analyzes how tone in an IPO prospectus is linked with the corresponding stock market performance and how information processing is influenced by pre-IPO news tone. Section 5 concludes the paper with a summary and an outlook on future research.

2 Related Work

In this section, we present related literature grouped into three categories. First, we revisit literature covering different types of IPO filings. Second, we compare approaches for financial documents to measure their tone. Third, we review previous work on information processing of investors when facing IPO filings. All in all, the following references provide evidence that investigating the effect of IPO filing tone across different time spans, as well as the impact of pre-IPO news tone, is both a novel and relevant research question to the finance community.

To a large extent, market efficiency relies upon the availability of information (Fama 1965). Access to market information is promoted with ease in electronic markets and, because of the straightforward access, decision makers (i.e. consumers, suppliers and intermediaries) can use more detailed information to make purchases and sales more beneficial (e.g. Granados et al. 2010). In fact, it is both native and crucial to finance research regarding how decision makers process and act upon (qualitative) information in IPO filings. Though information processing has been extensively studied in capital markets, this is not the case when it comes to literature focusing on IPO filings: “an examination of the textual or soft information contained in prospectuses is less common” (Ferris et al. 2013).

2.1 IPO Filings

This section introduces the basic filing types required by U.S. firms going public. For U.S. firms, one of the first steps in going public is filing a Form S-1 on the Securities and Exchange Commission’s (SEC) Electronic Data Gathering, Analysis and Retrieval (EDGAR) system. Typically, the S-1 offers investors their first detailed glimpse of a firm’s business model and financial statements. It is a pre-effective registration statement submitted when a company decides to go public. Hence, an investor can use its content to evaluate potential investment opportunity. If the S-1 is followed by a successful initial public offering, firms file a final prospectus (so-called Form 424), which is published via EDGAR on the day of or a few days following the IPO. Loughran and McDonald (2013) report a median time lag of 88 calendar days between S-1 and Form 424. According to the authors, more calendar time between the S-1 and 424 filings (the final IPO prospectus) is associated with significantly lower first-day returns. Companies with more problematic S-1 filings or facing adverse market conditions are often delayed in issuing stock. Some of the delay in going public could be due to the many days management needed to properly respond to SEC concerns about the IPO document or a sharp decline of IPO market conditions.
In recent years, the median number of words in both S-1 and Form 424 filings has significantly increased (Loughran & McDonald 2013). The typical 1997 IPO contained less than 40,000 words in either the Form S-1 or 424, while the documents for the typical 2010 IPO had about 80,000 words (see Figure 1). The median Form 424 has about 8% more words than the median Form S-1. However, the overall tone of the S-1 is typically very similar to the tone of the final IPO prospectus. In fact, changes in tone between the S-1 and 424 filings seem to be unrelated to positive revisions in the offer price (Loughran & McDonald 2013).

2.2 Analyzing News Content in the Finance Discipline

Methods that use the textual representation of documents to measure the positivity and negativity of the content are referred to as opinion mining or sentiment analysis. In fact, sentiment analysis can be utilized to extract subjective information from text sources. A more general approach is content analysis that, besides others, aims at measuring different variants of tone. For instance, one can study how market participants perceive and react upon financial materials. Here, one uses the observed price reactions following the financial text to validate the accuracy of the analysis routines. Based upon measures, one can study the relationship between financial documents and their effect on markets. For example, empirical evidence shows that a discernible relation between news content and its stock market reaction exists (e.g. Antweiler & Frank 2004; Tetlock 2007, for the first).

As methods that study tone are applied to a broad variety of domains and text sources, research has devised various approaches. For example, Pang and Lee (2008) provide a comprehensive domain-independent survey. Within finance, recent literature surveys (Minev et al. 2012; Mittermayer & Knolmayer 2006b) compare studies aimed at stock market prediction. For example, dictionary-based approaches are very frequently used in recent financial text mining research (cp. Demers & Vega 2010; Henry 2008; Jegadeesh & Wu 2013; Loughran & McDonald 2011; Tetlock et al. 2008). These methods count the frequency of pre-defined positive and negative words from a given dictionary – producing results that are straightforward and reliable. Machine learning approaches (e.g. Antweiler & Frank 2004; Li 2010; Mittermayer & Knolmayer 2006a; Schumaker & Chen 2009) offer a broad range of methods, but may suffer from overfitting (Sharma & Dey 2012).

In our research framework, we have only a few hundred IPO filings, each linked to a large text basis (i.e. a length of several thousand words) along with a continuous return value. Thus, we have experienced difficulties in achieving robust results with machine learning approaches and focused, instead, on dictionary-based methods. In fact, we have tested all of the above dictionary-based metrics in combination with various dictionaries. In short, we find that dictionaries vary strongly in the strength of the link between news tone
and stock market reaction. Along with previous research on IPOs, the ratio of words labeled as *uncertainty* according to the Loughran and McDonald Financial Sentiment Dictionary (McDonald 2012) outperforms all other approaches.

### 2.3 Information Processing Theory of IPO Filings

This section provides an overview of related research investigating the information processing of IPO filings; see Table 1. First, we review other research studying the information processing of IPO, but without considering the polarity of the prospectus content.

- **Arnold et al. (2010)** examine the risk factors section of a prospectus. They not only count the number of risk factors disclosed in this section, but also measure the number of words used to explain each of the risk factors. They find that the soft information is significantly related to both the initial and subsequent IPO returns, but not to ex post measures of investor sentiment.

- **Bajo and Raimondo (2015)**

- **Hanley and Hoberg (2012)** find that issuers tradeoff first-day returns and disclosure in amended filings (S-1/A) as a hedge against the risk of future law suits. The two authors show that IPOs having a higher risk of material omissions in their filings are more likely to hedge litigation exposure with higher levels of underpricing. A greater level of disclosure by issuers, as proxied by more meaningful revisions in their S-1/A filings during the bookbuilding period, lowers the probability of being sued by investors. The key assumption of Hanley and Hoberg (2012) is that an issuer with a large pre-IPO price adjustment (but having only revised their filings slightly) “is likely to have a material omission in the prospectus as the issuer did not disclose the information underlying the price change”.

- To evaluate disclosure style, Loughran and McDonald (2014) create a standardized statistic that aggregates a series of writing components specifically identified by the SEC. The six components are average sentence length, average word length, passive voice, legalese, personal pronouns, and negative/superfluous phrases. In the context of Form 424 filings, the authors find substantial differences in the content of IPO filings. This is a result of the 1998 rule implemented by the SEC requiring firms to use plain English in their prospectus filings.

Closest to our research are the approaches by Ferris et al. (2013), Hanley and Hoberg (2010), as well as Loughran and McDonald (2013), in terms of both research objectives and methodology. We individually address each of these papers as follows:

- **Ferris et al. (2013)** examine the effect that conservative or cautionary language (measured using negative tone) in the prospectus might have on IPO performance. As a result, greater conservatism in the prospectus is related to increased underpricing. The authors observe that this relation is stronger for technology than non-technology IPOs. Besides that, they study how conservative language evolves across time. In addition, they find a significant relationship between conservative language from the Loughran-McDonald dictionary and an average industry-adjusted return on assets for the 3-years following the IPO, but give no insights how returns evolve throughout time.

- **Hanley and Hoberg (2010)** examine the information content of IPO initial prospectuses and its effect on pricing. The authors split the information contained in the S-1 into standard and informative components. Interestingly, they find that IPOs with more informative content in their S-1 have lower offer price revisions and first-day returns. They argue that IPOs with more standard S-1 content would be more likely to solicit information from investors during bookbuilding.

- **Liu et al. (2014)** explore the effects of pre-IPO media coverage (in the Factiva database) measured by the number of newspaper articles. According to the authors, this news count correlates positively with the stock’s long-term value, liquidity, analyst coverage, institutional investor ownership and first-day
returns. Overall, the paper provides evidence for a long-term role for media coverage, consistent with Merton’s attention or investor recognition hypothesis.

- Loughran and McDonald (2013) study the relationship between S-1, as well as Form 424 filings, and the IPO first trading day returns. As a result, IPOs with a more uncertain or weak modal tone experience higher first trading day returns. Moreover, the percentages of uncertain, weak modal, and negative words in the S-1 are much more powerful variables in explaining levels of underpricing than many commonly used IPO control variables, such as venture capital dummy, top-tier underwriter dummy, or trailing annual sales. In fact, the $t$-statistics are higher when using Form 424 instead of S-1. In addition, tone is positively linked with higher volatility in a 60-day period following the offering, but there is no similar evidence for stock market returns.

In fact, the $t$-statistics are higher when using Form 424 instead of S-1. They use their word list (Loughran & McDonald 2011) to classify S-1 words into uncertain, weak modal, negative, positive, legal, and strong modal categories. However, these are not selective as of the 291 uncertain words, 40 overlap with the negative list. To assist the reader in understanding which words have a high weight in their analysis, they present the ten most frequently occurring words by the three word categories and by document type, but this does not identify words responsible for positive or negative IPOs.

| Table 1. Literature related to the Information Processing of Initial Public Offerings |
|---------------------------------|----------|-----------------|---------------------------------|-----------------|-----------------|-----------------|
| Reference                  | Corpus          | Years        | Dependent Variables                        | Regression Features                  | Method                      |
| Arnold et al. (2010)        | S-1, NYT articles | 1995–2013   | Underpricing, price revision               | Influence of media coverage/tone      | Pre-IPO media article count, dictionaries (positive, negative, uncertainty, weak modal, litigious) |
| Bajo et al. (2015)          | S-1, NYT articles | 1999–2005   | Number of analysts following & institutional investors holding stock, long-run valuation, price revision, expected return |          |                  |
| Ferris et al. (2013)        | S-1, NYT articles | 1999–2005   | Underpricing, post-IPO industry-adjusted ROA | By prospectus section, by tech./non-tech. firms | 3 Dictionaries (negative) |
| Hanley et al. (2010)        | S(B)-1           | 1996–2005   | First-day return, up revision               | By prospectus section                 | Document similarity         |
| Liu et al. (2014)           | Factiva news count | 1980–2004   | First-day return, absolute revision, up revision, post-IPO return volatility | Withdrawn IPOs                        | Pre-IPO media article count |

Hence, information processing has been the subject of many finance research publications. However, the above research papers concentrate on a partial examination of prospectus tone and, thus, many questions are left unanswered. All listed publications lack (1) an in-depth evaluation of how persistent the influence of filing content on long-term returns is, and (2) a comparison between the reception of IPO filings and pre-IPO news tone. Consequently, an empirical study analyzing the effects of filing tone in the long-term
and in pre-IPO news seems to be an open research question. Consequently, we pursue a rigorous approach as follows: first, we measure uncertainty language in IPO filings. Second, we specify a model to measure the impact of prospectus tone empirically.

3 Research Methodology: Measuring Tone of Filings and News

This section introduces our research methodology as depicted in Figure 2. In a first step, only those IPO prospectuses and news announcements are filtered that fit our research focus. Then, each prospectus and each announcement is subject to preprocessing steps (Section 3.1) which transform the running text into machine-readable tokens. The frequencies of these tokens are aggregated to compute the corresponding news tone in Section 3.2. Then, we analyze the influence of news tone on returns.

3.1 Preprocessing IPO Filings

Before measuring the actual tone, several operations are involved in a preprocessing phase. The individual steps are as follows.

- **Cleaning.** We use the textual component of IPO filings only and, then, all HTML tags and XBRL syntax are removed. This is consistent with Ferris et al. (2013), as well as Loughran and McDonald (2013).

- **Tokenization.** Each announcement is split into sentences and single words named tokens (Grefenstette & Tapanainen 1994).

- **Negations.** Negations invert the meaning of words and sentences. We adopt the following approach (Dadvar et al. 2011; Pröllochs et al. 2015): when encountering the word no, each of the subsequent three words (i.e. the object) is counted as words from the opposite dictionary. When other negating terms are encountered (rather, hardly, couldn’t, wasn’t, didn’t, wouldn’t, shouldn’t, weren’t, don’t, doesn’t, haven’t, hasn’t, won’t, hadn’t, never), the meaning of all succeeding words is inverted.

- **Stop word removal.** Words without a deeper meaning, such as the, is, of, etc. are named stop words and, thus, can be removed. We use a list of 571 stop words (Lewis et al. 2004).

- **Stemming.** Stemming refers to the process for reducing inflected words to their stem (Manning & Schütze 1999). Here, we use the so-called Porter stemming algorithm.

- **Document-term matrix.** The frequencies of terms that occur in the corpus is stored in a document-term matrix. In addition, we remove sparse terms that occur in less than 10% of all documents.
• **Weighting.** The information retrieval approach *term frequency-inverse document frequency* (tf-idf) reflects the importance of a word to a document $d$ in a collection $D$ and allows for the identification of discriminative words (Salton, 1983). Thereby, the raw frequency $f_{t,d}$ of each term $t$ in $d$ is weighted by the ratio of the total number $N$ of documents divided by the number $n_t$ of documents that contain the term $t$, i.e.

$$
tf-idf(t,d,D) = tf(t,d) \cdot \text{idf}(t,D) = f_{t,d} \cdot \log \frac{N}{|\{d \in D \mid t \in d\}|} = f_{t,d} \cdot \log \frac{N}{n_t}.
$$

(1)

### 3.2 Measuring News Tone

As shown in a recent empirical evaluation (Feuerriegel & Neumann, 2013), the robustness of different methods to measure news tone varies strongly. Out of potential dictionaries, the ratio of words labeled as *uncertain*—according to the Loughran and McDonald Financial Sentiment Dictionary (McDonald, 2012)—achieves the highest robustness and, consequently, we rely upon this approach in the following evaluation.

Let us briefly recapitulate this method. Let $W_{\text{total}}(P)$ denote the total number of words in the prospectus $P$; $W_{\text{uncertain}}(P)$ denote the number of *uncertainty* words in the prospectus $P$. Then, the news tone is defined by

$$
\text{Tone}(P) = \frac{W_{\text{uncertain}}(P)}{W_{\text{total}}(P)} \in [0,1].
$$

(2)

Thus, the news tone variable $\text{Tone}(P)$ measures the number of *uncertainty* words normalized by the number of total words.

### 3.3 Spike and Slab Regression

As part of our subsequent evaluation, we aim at selecting decisive words statistically in financial documents that have an impact on stock market returns. Such a problem setting can be formulated mathematically in terms of a regression model. Regression analysis is widely applied in statistics to investigate how a response variable is influenced by potential explanatory variables. Hence, our previous goal translates into the decision which subset of words should serve as potential explanatory variables in a regression model. The simplest regression model, i.e. the standard linear regression model, explains a response $y = (y_1, \ldots, y_N)^T$ of subjects $i = 1, \ldots, N$ by a linear function

$$
y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),
$$

(3)

of the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$ with a Gaussian error term $\varepsilon$, an intercept $\mu$ and a vector $x_i = (x_{i1}, \ldots, x_{ip})$ of $p$ potentially explanatory variables. Selecting a relevant subset of all potential explanatory variables poses a crucial question to the appropriateness of the model: on the one hand, omitting regressors with non-zero effects leads to biased estimates. On the other hand, including non-relevant regressors leads to overfitting, which worsens the predictive performance and limits the interpretability of the model. Thus, a primary goal of statistical analysis is correct classification into relevant and non-relevant regressors. This classification is called *variable selection*.

We can achieve variable selection effectively by introducing so-called *indicator variables* that distinguish between regressors with zero and almost zero effects. Consequently, the indicator variable $\delta_j$ for each regression coefficient $\beta_j$ represents inclusion or exclusion of the regressor into the model. The indicator variable takes the values zero and one, if the regressor is excluded or included, i.e.

$$
\delta_j = \begin{cases} 
0, & \text{if } \beta_j = 0, \\
1, & \text{otherwise.}
\end{cases}
$$

(4)
With the help of the above indicator variables $\delta_j$, we can rewrite the linear regression model into

$$ y_i = \mu + \delta_1 \beta_1 x_{i1} + \delta_2 \beta_2 x_{i2} + \ldots + \delta_p \beta_p x_{ip} + \epsilon_i. \quad (5) $$

The vector $\delta = (\delta_1, \ldots, \delta_p)$ then contains information regarding which elements of $\beta$ are included in the final model or set to zero. Choosing a vector $\delta$ yields the following reduced normal regression model in matrix notation,

$$ y = \mu + X^\delta \beta^\delta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \quad (6) $$

where $\beta^\delta$ contains only the non-zero elements of $\beta$ and the matrix $X^\delta$ consists of only the columns in $(x_1, \ldots, x_p)$ corresponding to non-zero effects.

From a Bayesian viewpoint, introducing the indicator variable $\delta_j$ in the normal linear regression model results into a prior in the form of a mixture of two distributions for each regression coefficient, so-called spike and slab priors. The spike component is a distribution with its mass concentrated around zero and the slab component is a flat distribution spread over the parameter space. Related literature proposes two different types of prior distributions for the spike component (Malsiner-Walli & Wagner 2011): first, spikes specified by an absolutely continuous distribution and second, so-called Dirac spikes defined by a point mass at zero. The former option comes along with computational advantages and we thus utilize a prior with an absolutely continuous distribution for the spike component. In this case, the spike and slab components are specified with the same distribution family but with a variance ratio $r$ smaller than 1, given by

$$ r = \frac{\text{var}_{\text{spike}}(\beta_j)}{\text{var}_{\text{slab}}(\beta_j)} \ll 1. \quad (7) $$

The spikes and slabs can be represented as scale mixtures

$$ \beta_j | \delta_j, \psi_j \sim \mathcal{N}(0, r(\delta_j) \psi_j), \quad (8) $$

of normal distributions with zero mean and variance $\psi_j$ where

$$ r(\delta_j) = \begin{cases} r, & \text{if } \delta_j = 0, \\ 1, & \text{if } \delta_j = 1. \end{cases} \quad (9) $$

Furthermore, the prior inclusion probability $w_j$ of a regressor $\beta_j$, i.e. the prior probability that a regressor has a non-zero effect, follows a Beta distribution

$$ p(\delta_j = 1 | w_j) = w_j, \quad w_j \sim \mathcal{B}(a_{w_j}, b_{w_j}). \quad (10) $$

As proposed by Ishwaran and Rao (2005), we choose normal mixtures of inverse Gamma distributions (NMIG prior) as the distribution family for the spikes and slabs, where the variance $\sigma$ itself follows an inverse gamma distribution $\sigma \sim \mathcal{G}^{-1}(v, Q)$. It can be shown (Konrath et al. 2008) that the resulting priors for the spike and slab components are given by two student distributions, one with high variance, i.e. the spike, and one with small variance, i.e. the slab. This characteristic is visualized in Figure 3, while we can formally define it via

$$ p_{\text{spike}}(\beta_j) = t_{2v}(0, rQ/v) \quad \text{and} \quad p_{\text{slab}}(\beta_j) = t_{2v}(0, Q/v). \quad (11) $$

Consequently, variable selection relies on the posterior probability of assigning the corresponding regression effect to the slab component and the decision is thus given by the posterior inclusion probability $p(\delta_j = 1 | y)$. In practice, the posterior inclusion probability and the regression coefficients can be calculated by Markov Chain Monte Carlo (MCMC) methods. In our setting, the MCMC scheme integrates into spike and slab regression for dictionary generation as follows: we treat each row of the document-term matrix as an
observation, while we use each column, i.e. each word, as explanatory variables to explain first-day stock market returns. We run the MCMC scheme (Gibbs sampling) for \( M = 1000 \) iterations after a burn-in of 500 draws. Afterwards, we calculate the posterior inclusion probabilities and the regression coefficients for each draw. Finally, the magnitude of the posterior inclusion probability \( p(\delta = 1) \) measures for each regressor the corresponding variable importance.

4  Empirical Evaluation: Analyzing News Tone of IPO Filings

Having discussed the steps to compute news tone values, we apply the aforementioned method to investigate how stock market returns of initial public offerings are driven by their prospectus. Initially, we choose the IPO filing corpus and gain first insights by descriptive statistics. Then, we specify all control variables, which, ultimately, give the regression design that inspects information processing by linking returns with prospectus tone. In addition to that, we analyze the reception of news covering initial public offerings.

4.1  IPO Filing Corpus, News Corpus and Stock Market Data

A structured dataset of U.S. initial public offerings is found in the so-called Kenney-Patton IPO database (Kenney & Patton 2013a, 2013b). This database consists of all de novo firms going public on American exchanges and filed with the Securities and Exchange Commission (SEC), where we restrict the corpus to those initial public offerings between the years 2003 and 2010. This accounts for a total of 625 IPOs. These initial public offerings are not evenly distributed across years, in fact, the number of IPOs exceeds 100 from 2004 to 2006, but drops sharply below 20 in 2008 and 2009 during the worldwide financial crisis – see Figure 4.
In a next step, we collect the final prospectus (i.e. Form 424) for each of these initial public offerings from the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system\(^1\). Filings are available online for 621 out of 625 initial public offerings.

Afterwards, we combine the Form 424 filings with their corresponding stock market performance. Here, we use the first-day return, defined as

\[
R_1 = \frac{(\text{ShareClosingPrice}_{1} - \text{InitialSharePrice})}{\text{InitialSharePrice}}. \tag{12}
\]

Daily data from stock markets is (mostly) provided by Thomson Reuters Datastream where price data is available for 577 initial public offerings. As in (Ferris et al. 2013; Loughran & McDonald 2013), we require an initial offer price of at least $5. In order to adjust for sector-specific differences like previous research (Ferris et al. 2013; Loughran & McDonald 2013), we introduce dummy variables for each industry. We use the dummies provided by the Thomson Reuters Business Classification (TRBC)\(^2\), which are classified into 52 industry groups. Because of data availability, we use a total of 571 out of these 577 observations.

In Table 2 the distribution of the sectors in my data set is listed. The majority comes from sector Pharmaceuticals and Biotechnology with a portion of 15.53%. Sector Software and Computer Services is the second biggest item (11.87%) closely followed by sector Health Care Equipment and Services with a portion of 9.60%.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|}
\hline
\textbf{Sector} & \textbf{Abs. Freq.} & \textbf{Rel. Freq.} & \textbf{Accum. Freq.} \\
\hline
Pharmaceuticals and Biotechnology & 89 & 15.53\% & 15.53\% \\
Software and Computer Services & 68 & 11.87\% & 27.40\% \\
Health Care Equipment and Services & 55 & 9.60\% & 37.00\% \\
Technology Hardware and Equipment & 47 & 8.20\% & 45.20\% \\
Financial Services & 38 & 6.63\% & 51.83\% \\
General Retailers & 36 & 6.28\% & 58.12\% \\
Others & 240 & 41.89\% & 100.00\% \\
Sum & 573 & 100.00\% & \\
\hline
\end{tabular}
\caption{Descriptive statistics of the sectors.}
\end{table}


Our news corpus originates from the Thomson Reuters News Archive for Machine Readable News. We choose Reuters news deliberately because of three reasons (MacGregor 2013; Paterson 2007): (1) Reuters conveys, in particular, news about stock markets. (2) Reuters news is third-party content and, thus, gives a certain level of objectivity. (3) In contrast to newspapers, news agencies feature a shorter time lag and lack perturbations by edits. All announcements provided by Reuters arise from January 1, 2003 onwards. The announcements come along with stock symbols denoting the firm the content deals with. Based upon these labels, the news corpus is filtered such that we extract the announcements of firms going public\(^3\). In addition, we require announcements to be published during the 365 days prior to the IPO. All in all, this set of criteria filters a total of 1085 announcements covering initial public offerings. If no announcement on an IPO is available, we set the tone value to zero. If more than one announcement is available, we use the average tone value. Further, we incorporate the number of news announcement linked to each IPO into a separate control variable \#News.

Examples of headlines are given in Table 3. From this table, we can see that the news not only reports that a firm is going public, but states more insights of the pre-IPO activities.

<table>
<thead>
<tr>
<th>Date &amp; Time</th>
<th>Stock Symbol</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2, 2003 14:15</td>
<td>SCOR</td>
<td>Cardinal Health completes acquisition of Syncor</td>
</tr>
<tr>
<td>Jan 6, 2003 10:49</td>
<td>CE</td>
<td>Council of Europe plans $750 mln 2010 bond Monday</td>
</tr>
<tr>
<td>Jan 7, 2003 16:59</td>
<td>ONE</td>
<td>Bank One launches $1 billion five-year note sale</td>
</tr>
<tr>
<td>Jan 7, 2003 17:10</td>
<td>PATH</td>
<td>AmeriPath shareholder urges buyout rejection</td>
</tr>
<tr>
<td>Jan 7, 2003 18:38</td>
<td>ONE</td>
<td>New Issue - Bank One sells $1 bln five-yr notes</td>
</tr>
<tr>
<td>Jan 8, 2003 14:01</td>
<td>MGG</td>
<td>MGM Mirage says it will miss Wall St. profit estimates</td>
</tr>
<tr>
<td>Jan 8, 2003 23:45</td>
<td>MGG</td>
<td>Park Place has not changed Q4 guidance-spokesman</td>
</tr>
<tr>
<td>Jan 9, 2003 10:17</td>
<td>PRO</td>
<td>Provalis expects first annual pre-tax profit</td>
</tr>
</tbody>
</table>

### 4.2 Descriptive Statistics of Stock Market Returns

In our following evaluation, we vary the time span of returns to study the persistence of prospectus tone. All descriptive statistics of issued share prices, as well as first-day returns, are presented in Table 4. Accordingly, initial prices of issued prices range from $4.00 to $85.00, while the mean initial share price accounts for $14.2613 with a standard deviation of $6.5506. Further, the returns after 21 trading days span approximately from \(-0.8337\) to 30.8000, with a non-zero mean of approximately 0.4304. Interestingly, the kurtosis of 134.3806 on the first day of trading is substantially higher than 3 indicating the existence of heavy tails, which are probably caused by price spikes. Both characteristics of non-zero mean and a high kurtosis are consistent with related literature (cf. Jenkinson & Ljungqvist 2001; Ritter & Welch 2002, for recent reviews), where these effects are referred to as underpricing.

In addition to that, box plots visualize quartiles of first-day returns across different years in Figure 6. Here, all medians (except for the year 2007) show very similar characteristics. It also interesting to note that the years 2008–2010 show considerably fewer outliers than the previous years.

---

3 This is achieved by applying a set of filter criteria (Feuerriegel & Neumann 2013): (1) the language must be English. (2) The event type is Story Take Overwrite to guarantee that we not yield an alert but the actual message. (3) Special types of announcements, such as alerts or tabular data, might have limited relevance and we want to exclude these. Thus, we omit announcements that contain specific words (advisory, chronology, corrected, feature, diary, instant view, analysts view, newsmaker, corrected, refile, rpt, schedule, table, service, alert, wrapup, imbalance, update) in their headline. (4) In order to remove white noise, we require announcements to count at least 50 words.
Table 4. Descriptive Statistics of Tone, Issued Prices and First-Day Returns

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToneIPO</td>
<td>0.0173</td>
<td>0.0173</td>
<td>0.0125</td>
<td>0.0246</td>
<td>0.0017</td>
<td>0.2094</td>
<td>0.5365</td>
</tr>
<tr>
<td>ToneNews</td>
<td>0.0049</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0476</td>
<td>0.0074</td>
<td>1.6724</td>
<td>3.1994</td>
</tr>
<tr>
<td>#News</td>
<td>2.2945</td>
<td>1.0000</td>
<td>0.0000</td>
<td>108.0000</td>
<td>8.2986</td>
<td>8.8609</td>
<td>90.7178</td>
</tr>
<tr>
<td>Issued Price</td>
<td>14.2613</td>
<td>14.0000</td>
<td>4.0000</td>
<td>85.0000</td>
<td>6.5506</td>
<td>3.6347</td>
<td>29.3660</td>
</tr>
<tr>
<td>R(1)</td>
<td>0.4216</td>
<td>0.0842</td>
<td>−0.8558</td>
<td>35.0000</td>
<td>2.2441</td>
<td>10.5652</td>
<td>134.3806</td>
</tr>
</tbody>
</table>

Figure 5. Histograms of First-Day Returns

Figure 6. Box Plots comparing First-Day Log-Returns across Years 2003–2010
4.3 Control Variables

When isolating and extracting the effect of tone, we use a wide range of control variables. We control for the development of the stock market and, in addition to that, further control variables comprise the number of shares, the share overhang, as well as the underwriter discount. The latter three variables are retrieved from the Kenney-Patton IPO database (Kenney & Patton 2013a, 2013b), which we improved with additional corrections. Similar variables are used by a number of prior papers to explain first-day returns (see e.g. Ferris et al. 2013; Loughran & McDonald 2013). Individual control variables are as follows:

- **UpRevision** This control variable measures the percentage upward revision from the mid-point of the filing range if the offer price is greater than the mid-point, otherwise it is set to zero. As in (Ferris et al. 2013; Loughran & McDonald 2013), including this variable is crucial; when this variable is added last, many others are no longer significant.

- **ΔDaysS-1,IPO** To control for possible delays of IPOs, we include a variable measuring the logarithmized calendar days between S-1 and initial public offering.

- **NASDAQ_{−15d}** We include the general development of a stock market index to ensure that the stock market performance is driven by tone instead of the economic cycle or the general mood of investors. Thus, we choose the return of the NASDAQ-100 stock market index of the 15 days prior to the initial public offering. Here, values originate from Thomson Reuters Datastream.

- **lnSales** This variable is defined as the logarithm of the number of shares sold to the public in the offering multiplied by the initial share price. Thus, the logarithm of all sales controls for the volume of the initial public offering.

- **ShareOverhang** This variable is the number of shares retained divided by the number of shares in the initial offering. More precisely, it is defined as

  \[ ShareOverhang = \frac{S_{\text{outstand}}}{S_{\text{outstand}} + S_{\text{sold}}}, \]

  where \( S_{\text{outstand}} \) is the number of shares outstanding after the offer, and \( S_{\text{sold}} \) is the number of shares sold to the public in this offering.

- **UnderwriterDiscount** This variable includes the per share discounts and commissions taken by the underwriters.

By using these control variables (Ferris et al. 2013; Loughran & McDonald 2013), we want to assure that our results measure the impact that comes from tone only and avoid perturbations by other causes, such as changes in fundamental variables or external events. Next, descriptive statistics of all variables are presented in Table 5. In addition to that, we incorporate additional dummies into our model to adjust for both seasonal and sector-specific effects.

---

4 Although many papers control if the firm is backed by venture capital. However, Loughran and McDonald (2013) find no evidence that this variable has a significant impact and, thus, we omit it.
Table 5. Descriptive Statistics of Control Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta_{Days_{S-1,IPO}}$</td>
<td>3.1603</td>
<td>3.0201</td>
<td>0.0000</td>
<td>6.6201</td>
<td>0.8557</td>
<td>-0.0029</td>
<td>5.1859</td>
</tr>
<tr>
<td>$NASDAQ_{-15d}$</td>
<td>0.0065</td>
<td>0.0072</td>
<td>-0.1800</td>
<td>0.0979</td>
<td>0.0373</td>
<td>-0.4665</td>
<td>0.5743</td>
</tr>
<tr>
<td>$\ln Sales$</td>
<td>18.3719</td>
<td>18.2869</td>
<td>15.5690</td>
<td>21.5984</td>
<td>0.8806</td>
<td>0.4368</td>
<td>0.9694</td>
</tr>
<tr>
<td>$ShareOverhang$</td>
<td>0.7699</td>
<td>0.7826</td>
<td>0.0004</td>
<td>0.9879</td>
<td>0.0008</td>
<td>-2.7661</td>
<td>16.7556</td>
</tr>
<tr>
<td>$UnderwriterDiscount$</td>
<td>0.9538</td>
<td>0.9100</td>
<td>0.0000</td>
<td>3.8400</td>
<td>0.3752</td>
<td>1.8949</td>
<td>10.6112</td>
</tr>
</tbody>
</table>

4.4 Regression Design

This section presents the corresponding regression design, i.e.

$$R_I = \alpha + \beta_1 Tone_{IPO} + \beta_2 UpRevision + \beta_3 \Delta{Days_{S-1,IPO}} + \beta_4 NASDAQ_{-15d} + \beta_5 \ln Sales + \beta_6 ShareOverhang + \beta_7 UnderwriterDiscount + \sum_j \beta_j Dummy_j + \varepsilon,$$

where $\beta_i$ are coefficients to be estimated and $\varepsilon$ is the error term. Dummies are added for each sector, as well as year, to consider additional external events not covered by the control variables and to handle non-seasonally-adjusted time series. The correlation of filing tone and first-day returns is depicted in Figure 7. This diagram also features a so-called LOWESS trend line, which is a locally weighted scatterplot smoothing calculated via local regressions. From this LOWESS trend line (smoothing parameter $f$ set to 2/3), we can identify a visible relationship between filing tone and stock market returns. Finally, we give justice to extreme stock price effects and remove outliers at the 0.05% level at both ends.

Figure 7. Scatterplot of First-Day Returns and Filing Tone with LOWESS Trend Line

4.5 Impact of Filing Tone on Stock Market Performance

In this section, we investigate how investors react to IPO prospectuses and analyze the information processing empirically. To succeed in this goal, we measure the relationship between investor behavior and the content
of IPO filings.

**Confirmatory Analysis: To what extent are stock market returns of firms driven by their IPO filing tone?**

We use the above regression design from Equation (14) to measure the impact of prospectus tone on IPO underpricing. We tested for heteroskedasticity, constant variance, serial correlation and normally distributed residuals at the 0.01 % level to ensure that the results are not confounded. When checking Variance Inflation Factors, we also see no indication of multicollinearity. Independence across IPO filings is given as long as all prospectuses are entirely novel and not based on an interrelated course of events. Under the assumption that first-day returns are jointly multivariate normal, as well as independently and identically distributed through time, the model can be estimated using Ordinary Least Squares (OLS).

Regression results are given in Table 6. According to this table, we observe that prospectus tone influences first-day returns significantly. The corresponding t-value accounts for −2.505 with a p-value of 0.013. Further, we find that one standard deviation increase in the tone measure is negatively linked to an increase in first-day returns by an economically significant 10.56 %. When additionally comparing the coefficients of fundamental variables and filing tone, we find that the tone coefficient (accounting for −61.879 ) exceeds all other coefficients originating from fundamental variables greatly. In addition to that, none of the control variables are, except for UpRevision, significant at the 10 % level. Altogether, these results provide evidence that tone in IPO prospectuses correlates significantly with first-day returns.

<table>
<thead>
<tr>
<th>Table 6. OLS Regression linking Filing Tone and First-Day Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(a)</strong></td>
</tr>
<tr>
<td><strong>Tone</strong></td>
</tr>
<tr>
<td><strong>UpRevision</strong></td>
</tr>
<tr>
<td><strong>NASDAQ</strong></td>
</tr>
<tr>
<td><strong>NASDAQ</strong></td>
</tr>
<tr>
<td><strong>In Sales</strong></td>
</tr>
<tr>
<td><strong>Share Overhang</strong></td>
</tr>
<tr>
<td><strong>Underwriter Discount</strong></td>
</tr>
<tr>
<td><strong>Intercept</strong></td>
</tr>
<tr>
<td><strong>AIC</strong></td>
</tr>
<tr>
<td><strong>BIC</strong></td>
</tr>
<tr>
<td><strong>Multiple R</strong></td>
</tr>
</tbody>
</table>

Stated: Coef. and t-Stat. in Parenthesis. Dummies: Year, Sector. Obs.: 571. Signif.: ** 0.001, * 0.01, ' 0.05

### 4.6 Filing Tone versus News Tone

While the previous sections analyzed the influence of news, the question of the relevance of the general news coverage is left unanswered.

**Research Question 1: How is the IPO performance affected by pre-IPO news coverage?**

We extend the regression of Equation (14) by an additional variable measuring the tone of news announcements related to that initial public offering. Out of all 571 initial public offerings, we find at least one
This is interesting for two reasons. First, a higher t-value indicates a significant link between returns and news, even stronger than filing tone. Second, while more uncertainty words in the prospectus lower the first-day returns, this does not hold for the news announcements. The more uncertainty words in the news announcements, the higher the first-day return. While one standard deviation increase in the IPO filing tone measure is linked to a decrease in first-day returns by an economically significant 11.30%, the news tone behaves differently. Here, one standard deviation change in news tone correlates with an increase in first-day returns by 14.72%. Furthermore, the number of news announcements (given by \( \# \text{News} \)) has no impact on IPO prices at any common significance level. Astonishingly, the effect shows a larger magnitude and, at the same time, the more negative the pre-IPO news coverage, the higher the first-day returns.

| Table 7. OLS Regression comparing Influence of Tone in IPO Filings and News |
|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| (a)                     | (b)                     | (c)                     | (d)                     | (e)                     | (f)                     | (g)                     | (h)                     | (i)                     |
| \( \text{Tone}_{\text{IPO}} \) | -46.617                | -44.120                 | -44.066                 | -42.432                 | -39.287                 | -36.993                 | -56.361*                | -56.439*                | -66.045*                |
|                         | (-1.813)               | (-1.732)                | (-1.735)                | (-1.667)                | (-1.540)                | (-1.446)                | (-2.113)                | (-2.113)                | (-2.456)                |
|                         | (3.167)                | (3.385)                 | (3.381)                 | (3.294)                 | (3.238)                 | (3.503)                 | (3.489)                 | (3.518)                 |
| \( \#\text{News} \)       | -0.010                 | -0.010                  | -0.009                  | -0.009                  | -0.009                  | -0.01                   | -0.01                   | -0.01                   |
|                         | (-1.937)               | (-1.857)                | (-1.777)                | (-1.754)                | (-1.905)                | (-1.905)                | (-1.905)                | (-1.905)                |
| \( \text{Up} \text{Revision} \)  | 0.005                  | 0.005                   | 0.005                   | 0.008                   | 0.008                   | 0.010                   | 0.010                   |
|                         | (1.065)                | (1.086)                 | (1.222)                 | (1.825)                 | (1.781)                 | (1.969)                 |
| \( \Delta \text{Days}_{5,1,\text{IPO}} \)  | -0.074                 | -0.074                  | -0.086                  | -0.086                  | -0.086                  | -0.084                  |
|                         | (-1.452)               | (-1.455)                | (-1.680)                | (-1.681)                | (-1.681)                | (-1.631)                |
| \( \text{NASDAQ}_{-15d} \)     | -1.279                 | -1.372                  | -1.376                  | -1.356                  | -1.356                  | -1.173                  |
|                         | (-1.117)               | (-1.204)                | (-1.181)                | (-1.181)                | (-1.181)                | (-1.020)                |
| \( \text{In Sales} \)       | -0.130*                | -0.135*                 | -0.135*                 | -0.135*                 | -0.135*                 | -0.120*                 |
|                         | (-2.430)               | (-2.407)                | (-1.966)                | (-1.966)                | (-1.966)                |
| \( \text{ShareOverhang} \)     | 0.053                  | 0.070                   | 0.070                   |
|                         | (0.113)                | (0.146)                 |
| \( \text{UnderwriterDiscount} \)     | -0.130                 |
|                         | (-0.890)               |
| \( \text{Intercept} \)       | 0.723                  |
|                         | (1.263)                |
| \( \text{AIC} \)             | 1385.908               |
|                         | 1377.069               |
| \( \text{BIC} \)             | 1568.072               |
|                         | 1563.47                |
| \( \text{Multiple } R^2 \)    | 0.086                  |
|                         | 0.105                  |
| Stated: Coef. and t-Stat. in Parenthesis | Dummies: Year, Sector | Obs.: 523 | Signif.: \( * * * \) 0.001, \( * \) 0.01, \( \cdot \) 0.05

A possible explanation can be attributed to the investors’ psychological constraints in information processing – the processing depends on the attention different information receives. This phenomenon is already addressed in accounting literature (Hirshleifer et al. 2011): limited investor attention theory for earning announcements suggests that investors neglect information about future earnings in current-period announcements creating a post-earning announcement drift. Subsequently, this mispricing is corrected over time. The same explanation may apply to IPO prospectuses, where mainly the risks of the company are recognized, which may entail the negative coefficient of our filing tone. Underestimating the future opportunities may create a post-prospectus announcement drift that was visible in our previous analysis. News announcements prior to the IPO are obviously less suspect to the companies’ risks. On the contrary, the positive coefficient suggests that more
emphasis is put on the companies’ opportunities. Apparently, the degree of investor attention depends on the specific type of information.

### 4.7 Decisive Words for Stock Market Performance

This section aims at putting the results into context, but before that, we discuss the driving forces behind decision-making when investing in firms going public.

*Research Question 2: Which positive and negative words account for the tone in IPO filings?*

Analyzing how many times words are used in a positive or negative environment reveals interesting insights. For example, in previous research, Loughran and McDonald (2013) study the frequency of words from certain word sets, such as a negative list. Even though this outlines words most commonly used, a different measure is necessary to identify words that actually drive stock market returns. Thus, we visualize words by their appearances in filings linked with positive, as well as negative returns. Exploiting such a two-dimensional representation, one can easily spot words appearing significantly more often in underpriced than overpriced (or vice versa) initial public offerings. In fact, not the pure frequency of a word is responsible for stock market returns, but more unevenly distributed occurrences across IPOs with positive/negative first-day returns. In addition to that, we improve the metric, in contrast to (Ferris et al. 2013; Loughran & McDonald 2013), by using rules to negate phrases with inverted meaning (cf. Section 3). This allows us to boost accuracy further and overcome issues (see Loughran & McDonald 2013) to finally study also positive word lists.

Plots the frequency of words linked with positive/negative stock market returns. Here, we omit stop words, such as and, the or of to extract only words conveying a deeper meaning. The distance from the line through the origin is a measure on the strength of the link with the financial results. In a next step, we transform the number of occurrences into a so-called Bi-Normal Separation score used in term extraction (Forman 2002, 2003, 2008). The more positive the BNS score of a word, the stronger its link with a positive stock market reaction and vice versa. Examples of word stems – usually occurring in overpriced IPO prospectuses – with a BNS score of $-1.77$ and below are, for example, immunotherapi, forstmann, franchise and mortgagerel. Similarly, one considers the following word stems as extremely positive with a BNS score of $1.77$ and above: bank, complianc, expand, grow, govern, technolog and trust. Thus, these words usually serve as an indicator for underpricing.

**Results for:**

- 570 IPOs
- Sparsity 0.4
- Normalized term-frequencies
- 500 burn-in draws
- 1000 MCMC iterations
Tone in IPO Filings and Pre-IPO News

<table>
<thead>
<tr>
<th>Word</th>
<th>Coefficient</th>
<th>$P(\delta = 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>citizen</td>
<td>0.4030</td>
<td>1.0000</td>
</tr>
<tr>
<td>issuer</td>
<td>0.4116</td>
<td>1.0000</td>
</tr>
<tr>
<td>satisfactori</td>
<td>0.3984</td>
<td>1.0000</td>
</tr>
<tr>
<td>never</td>
<td>0.2420</td>
<td>0.8090</td>
</tr>
<tr>
<td>line</td>
<td>0.2316</td>
<td>0.7560</td>
</tr>
<tr>
<td>appear</td>
<td>−0.0908</td>
<td>0.4720</td>
</tr>
<tr>
<td>mark</td>
<td>0.0911</td>
<td>0.4650</td>
</tr>
<tr>
<td>ventur</td>
<td>0.0919</td>
<td>0.4400</td>
</tr>
<tr>
<td>overview</td>
<td>0.0551</td>
<td>0.2920</td>
</tr>
<tr>
<td>make</td>
<td>−0.0542</td>
<td>0.2660</td>
</tr>
<tr>
<td>consumm</td>
<td>0.0324</td>
<td>0.2070</td>
</tr>
<tr>
<td>six</td>
<td>0.0266</td>
<td>0.1720</td>
</tr>
<tr>
<td>neither</td>
<td>−0.0337</td>
<td>0.1690</td>
</tr>
<tr>
<td>less</td>
<td>−0.0361</td>
<td>0.1680</td>
</tr>
<tr>
<td>like</td>
<td>−0.0255</td>
<td>0.1680</td>
</tr>
<tr>
<td>communiti</td>
<td>0.0293</td>
<td>0.1620</td>
</tr>
<tr>
<td>experienc</td>
<td>−0.0242</td>
<td>0.1530</td>
</tr>
<tr>
<td>promulg</td>
<td>0.0181</td>
<td>0.1470</td>
</tr>
<tr>
<td>select</td>
<td>−0.0273</td>
<td>0.1470</td>
</tr>
<tr>
<td>estim</td>
<td>−0.0226</td>
<td>0.1410</td>
</tr>
</tbody>
</table>

### 4.8 Discussion

All in all, we can empirically establish a relationship between tone in IPO filings and stock market reaction. This dependency appears not only on the first day of trading, but also on longer phases of up to 10 days of trading. In fact, “soft information can offer context to financial numbers and share values, provide insight into managerial expectations, and identify important qualifiers or caveats that are absent from purely numerical data” (Ferris et al. 2013) As an explanation, Loughran and McDonald (2013) argue “we could expect the IPOs with substantial uncertain/negative language to have, on average, low preliminary offer prices, large upward price revisions, and high first-day returns due to the need of bankers to compensate investors for their information production. An added benefit for IPOs in having institutional investors spending more resources to correctly price the offering is to reduce the litigation risk faced by managers (cf. Hanley & Hoberg 2012)”. According to (Ferris et al. 2013), the findings “suggest that when hard information is noisier, textual information is seen as having greater usefulness and is more likely to be factored into IPO pricing”.

### 5 Conclusion

Before shares of a company are sold to the general public on a security exchange for the first time, regulatory publication requirements force U.S. firms to file an initial public offering prospectus. Studying the information processing of investors facing these filings is an active research question in the context of efficient electronic markets, though knowledge is rare (e.g. Liebmann et al. 2012): “while the accounting numbers in IPO prospectuses are closely studied by investors, analysts, and others involved in the equity issuance process, an examination of the textual or soft information contained in prospectuses is less common” (Ferris et al. 2013).
To close this gap, research must harness decision analytics in order to explain the relationship between the textual content of filings and final IPO prices.

In this paper, we address information processing in IPO filings by analyzing how tone in prospectuses influences stock market performances. Using final filings of 571 U.S. initial public offerings (Form 424) between 2003 and 2010, we can identify soft content as a major driver of stock returns. All in all, we can empirically establish a relationship between tone in IPO filings and stock market reaction. One standard deviation increase in the tone metric correlates negatively with a change in first-day returns by an economically significant 10.56%. This dependency appears not only on the first day of trading, but also for longer phases of up to 10 days of trading. Apparently, “soft information can offer context to financial numbers and share values, provide insight into managerial expectations, and identify important qualifiers or caveats that are absent from purely numerical data” (Ferris et al. 2013). As an explanation, we find supporting evidence that we can expect initial public offerings with substantial uncertain language to have, on average, lower preliminary offer prices. This effect occurs due to the need of bankers to compensate investors for their information production (Loughran & McDonald 2013). According to Ferris et al. (2013), the findings “suggest that when hard information is noisier, textual information is seen as having greater usefulness and is more likely to be factored into IPO pricing”. However, a much stronger impact on stock market prices happens not through the IPO prospectus, but comes from the pre-IPO news tone. Interestingly, the more uncertainty words that appear in pre-IPO news, the higher following first-day stock market return. Consequently, this uncertainty puzzle must be resolved. It seems that investors mainly focus on the chances of a company rather than on the risks. This stands in contrast to the prospectus where risk considerations play a major role.

The work presented in this paper opens several avenues for future research. First, further effort is needed to validate our approach in terms of robustness and, thus, we plan to extend our analysis further to filings, such as S-1 or, alternatively, transcripts of court decisions. Second, an obvious next step is to bridge the gap from explanatory regression models to a predictive model (Shmueli & Koppius 2011) that can estimate expected stock market returns from IPO filings.

References


