Text Mining for Studying Management's Confidence in IPO Prospectuses and IPO Valuations

Full Paper

Jie Tao
Fairfield University
jtao@fairfield.edu

Amit V. Deokar
Pennsylvania State University
avd108@psu.edu

Abstract

Understanding pricing strategies in the context of the Initial Public Offering (IPO) process has been receiving much attention. Most prior studies have however focused on information sources from post issuance periods, and understanding such strategies from the management's perspective during the IPO process is still an open research issue. Form 424 variants, as finalized IPO prospectus approved by *Security Exchange Committee* (SEC), contain rich and genuine information about the issuing firms. In this study, we analyze the inter-relationships between the management's confidence (through the proxy of sentiments expressed in textual contents in the *Management's Discussion & Analysis* (MD&A) sections in the prospectus) and the pre-/post-IPO valuations. We develop an analytical framework namely *FOCAS-IE* (*Feature-Oriented, Context-Aware, Systematic Information Extraction*) to derive sentiments from the MD&A sections. Further, we construct predictive models using information extracted using FOCAS-IE to predict IPO pricings. The results have shown to outperform results from prior related studies.

Keywords

Text Mining, Predictive Modeling, Initial Public Offering.

Introduction

Understanding the determinants and dynamics behind the pricing strategies in the Initial Public Offering (IPO) process has been of keen interest to researchers and practitioners in different management domains. Although there is a large body of work attempting to understand the "*underpricing*" phenomenon (e.g., (Ferris et al. (2013), Hanley and Hoberg (2010), Loughran and McDonald (2013), Loughran and Ritter (2004)), much less work has been done to explain the sentiments of the issuer/underwriters as it relates to the IPO price revisions. Most studies in this area focus on analyzing the post-IPO market/investor sentiments and their effects on the demand of IPOs, via sentiment analysis of post issuance financial news, reports, and social media (Geva and Zahavi 2014; Lin et al. 2011). Open research issues remain regarding how the management's confidence about the offering as well as the company's future performance outlook affects IPO price revisions. Management's confidence, or its opposite, issuer's conservatism, or tone, are used interchangeably within the finance literature, which depict the management's prospects toward the future performance of the organization, and are measured using sentiment indicators (Ferris et al. 2013; Li 2010). Researchers have argued that the pre-IPO pricing is of equal importance, if not more, compared to the post-IPO pricing, and that the key to understanding the pre-IPO pricing is through sentiment analysis of management's confidence about the company's future outlook (Cornelli et al. 2006).

Among the variety of public IPO-related reports filed with the Security Exchange Committee (SEC), Form 424 and its variants provide a detailed snapshot of the offerings. Form 424 is the final IPO prospectus, filed on or within a few days of the IPO date, and is indicative of successful completion of the IPO process. IPO prospectus, including both Form 424 and the initial prospectus (namely S-1 filings), has been recognized as the most reliable source because SEC's regulations of comprehensiveness and truthfulness. Further, prior studies have found out that the tone in Form 424 reflected through the textual content in

the filings is a direct proxy of the management's confidence (or conservativeness) (Ferris et al. 2013; Hanley and Hoberg 2010; Loughran and McDonald 2013). Within the Form 424 filings, Hanley and Hobery (2010) have identified that four sections are more informative than the others, namely: *Prospectus Summary*, *Risk Factors*, *Use of Proceeds*, *Management's Discussions and Analysis* (MD&A). Among these, generally MD&A sections disclose the management's prognosis of the firm's prospects in a richer and more direct manner through use of sentiment-oriented statements (Hanley and Hoberg 2010).

It may be noted that extant studies in this context are primarily explanatory, as opposed to predictive (Hanley and Hoberg 2010; Loughran and McDonald 2013). While explanatory studies have yielded fruitful results in testing existing theory in the finance domain, there is little focus on the studying the predictive ability of IPO characteristics as well as textual features on IPO pricing. As argued by Shmueli and Koppius (2009), predictive analysis is effective in assessing the predictability of features identified through explanatory studies, as well developing new features and measures from prediction standpoint, eventually leading to new theoretical propositions. In this study, we link the management confidence with the pre-IPO offering price adjustment and post-IPO initial returns for a sample of 513 completed U.S. IPOs with Form 424B4 filings during the recent decade 2003-13. The overall research problem of this study can be stated as "*How does the management's confidence regarding the firm's outlook (expressed through the sentiments in the Form 424 filings MD&A section) affect IPO valuations?*" In this paper, we develop an analytical framework, namely *FOCAS-IE* (*Feature Oriented, Context-Aware, Systematic Information Extraction*) to derive sentiments from Form 424 filings data and subsequently apply predictive modeling techniques to understand the ability of the derived sentiments to predict pre-IPO price revision and post-IPO initial returns. The contributions of this study also include a sentiment-based metric as a proxy for the management's confidence toward the firm's issuance and future performance. Further, we conduct a predictive modeling-based experiment that shows the value-added with the proposed approach by comparing the predictive ability of the domain-specific sentiment-oriented features in relation to only considering the IPO characteristics-based predictors used in extant studies.

The remainder of the article is organized as follows. Section 2 provides an overview of related work, conceptual foundations and research question of this study based on them. Section 3 describes the study data. The proposed *FOCAS-IE* framework is described in Section 4. Section 5 discusses the predictive modeling based on sentiment-oriented, context-aware features to analyze pre-IPO price revision and post-IPO initial return. Following this, Section 6 presents the results, validation, and discussion. Finally, Section 7 concludes the paper with summary of contributions, limitations and future directions.

Conceptual Foundations

IPO Valuation and Information Content of the Prospectus

Traditional IPO valuation and pricing strategy related studies have reported manual investigation of quantitative, financial facts disclosed in the prospectus, particularly in the use of proceeds section (e.g., Ritter and Welch (2002), Bhabra and Pettway (2003), Jenkinson and Jones (2009), Löffler et al. (2005), Lowry and Schwert (2004)). In the past decade, computer-mediated content analysis has become a key direction for this research. For example, Arnold et al. (2010) examine the informative contents within the *Risk Factor* sections of the prospectus, and make an argument that the ambiguity level of the disclosed information has a correlation with the pre- as well as post-IPO price volatilities. Their findings suggest that more ambiguous information in the prospectus leads to higher price adjustment. Hanley and Hoberg (2010) conduct content analysis on four important sections within the prospectus, and classify content into *standard* and *informative* content. Their findings indicate that IPOs with more informative content tend to have lower pre-IPO price revisions and initial returns. The sentiment-oriented contents in the prospectus have also been analyzed, and they argue that positive tone related to valuation and due diligence (accounting, corporate strategies, and products/revenues) links to lower price revision and initial return (i.e., market's price adjustment). In a follow-up study, Hanley and Hoberg (2012) further illustrate that the risks to future legal issues correlate to greater levels of underpricing of the IPO. Although prior related studies have generally argued that the sentiments are effective proxy of the management's confidence (or to the contrary, conservatism) with respect to the organization's prospect (Ferris et al. 2013; Loughran and McDonald 2013), measures for capturing the sentiments based in a

context-aware manner are not considered. Our study builds on this rationale while proposing measures for sentiment-oriented features that are based on the IPO context being analyzed.

Few recent studies have used sentiments and tones expressed in the IPO prospectus textual contents to analyze IPO valuation. Loughran and McDonald (2013) analyze the tone of the prospectus (mainly Form S-1 which is claimed to be as important as Form 424 filings) and the IPO valuation by considering the occurrences of sentiment words from six sentiment word lists (L&M word lists focused exclusively on the finance/IPO domain) developed by them in an earlier study (Loughran and Mcdonald 2011) as a proxy for the overall tone of the prospectus. The major finding in this study is that IPO prospectuses with more uncertain words (negative or weak modal) have higher pre-IPO price revisions and initial returns. In a similar study by Ferris et al. (2013), the authors proxy the management's confidence from the perspective of conservatism, which is calculated as the ratio between the negative words and total words in the prospectus. The study also utilizes the L&M word list for analyzing sentiments and the findings in their paper are consistent with the study by Loughran and McDonald (2013). Other knowledge resources have also been reported in examining IPO valuations or management's confidence. For instance, in studies such as (Li 2010; Lin et al. 2011), annual and quarterly corporate financial reports are used (i.e. 10-K, 10-Q) to conduct content analysis to understand the management's prospects and stock price movement, respectively. Compared to the annual reports, the prospectuses are better data sources for studying IPO valuation since they are more comprehensive (over several years before IPO) and targeted (Loughran and McDonald 2013). Other studies utilize the financial news and social media contents as the knowledge resources (Hagenau et al. 2013; Schumaker and Chen 2009). In comparison to the IPO prospectus, these sources lack comprehensiveness, authority, and compliance with respect to the IPO process.

While the aforementioned studies report substantial progress, following research gaps may be noted: (a) Arguably, without the consideration of the context, mere counts of the sentiment-oriented words are not precise proxies of the management's outlook. For instance, a simple negation in the context might change the direction of the sentiments. Similarly, phrases expressed in a certain structure may imply a different meaning. Thus, a context-aware sentiment analysis is deemed necessary in analyzing prospectus contents. Even from a generic sentiment analysis standpoint, this advancement can be further adapted to various domains. (b) As indicated in Hanley and Hoberg (2010), the topical contents disclosed in different sections are different – thus, their impacts on the IPO valuation might vary. On the other hand, even the same sentiment on different features might have different effects on the IPO pricing (e.g., a negative sentiment on marketing and operation strategies could have opposite effects on price revisions). Thus, a section-by-section, feature-oriented analytical approach would be worthwhile. (c) Existing studies lack an underlying knowledge structure that can serve as a basis for reasoning and analytics (e.g., considering features and contexts). Thus, a well-constructed (in terms of coverage, accuracy, and domain specificity) ontological structure is needed that can be leveraged for analytical tasks on IPO prospectus content. These research gaps have been translated as key requirements in the design of our proposed analytical framework discussed in Section 3.2.

Research Question

The key research question for our study is: "*How does the management's confidence about the firm's outlook (expressed through the sentiments in the Form 424 filings MD&A section) affect IPO valuations?*" This research question relates to the management's prospects toward the firm's issuance and future performances as well as their impacts on the pre-IPO price adjustment(s) as well as post-IPO initial returns. As discussed in the finance literature (Ferris et al. 2013; Hanley and Hoberg 2010; Loughran and McDonald 2013), from the disclosure of the initial range of the offering price in the initial prospectus (Form S-1) to the offering price is determined in the final prospectus (Form 424), there might be none, one, or more price adjustments. The attitude of the issuer's management is one of the key determinants of such adjustment(s). The forward-looking statements within the textual contents of the prospectus are valid proxies of such attitude, which mostly fall in the MD&A sections of the prospectus (Hanley and Hoberg 2010; Li 2010). The forward-looking statements may involve certain topical contents, such as valuation, operation strategies, marketing, products/services, and so forth. The goal of the research question is to analyze these topical contents in the forward-looking statements within the MD&A sections of the prospectus, and then assess their impacts on the IPO price valuation. The underlying premise is that more positive contents in the forward-looking statements indicate higher level of management's confidence.

Study Data

The research study data mainly includes IPO prospectuses retrieved from the US SEC's EDGAR system, selected for the companies that went public in the most recent decade (January 1, 2003 – December 31, 2013). In accordance with similar prior studies, we have excluded: Financial firms (i.e. banks and loan firms), American Depository Receipts (*ADRs*), real estate investment trusts (*REITs*), close-end funds, and firms with offering price less than five dollars (Hanley and Hoberg 2010; Loughran and McDonald 2013). We have further restricted our sample by the stock type of common stock or ordinary stock. The Center for Research in Security Prices (CRSP) serves as the source for stock offering prices, initial returns, and value-weighted returns of each IPO.

Form 424 is the final IPO prospectus, which is filed on or within several days after the IPO day. Among all 424B variants, we have selected Form 424B4 in this study. All 424B4 entering the sample must have a valid SEC Central Index key (CIK), CRSP *permno*, and key sections including the MD&A section. Using a standard web crawling algorithm, we retrieved 424B4 filings from SEC EDGAR for 713 US IPOs, which were then filtered through a filtering algorithm to meet aforementioned sampling boundary restrictions, resulting in 513 filings.

Variable Name	Description
Up Revision	Set to $\|\Delta P\|$ if $\Delta P > 0$, otherwise 0
Days between S-1 and 424B4	The log of calendar days between the initial S-1 filing and the filing of Form 424B4 from EDGAR.
Top-tier Dummy	Dummy variable, set to 1 if the leading underwriter of IPO has a rating of 8 or higher, otherwise 0
Positive EPS Dummy	Dummy variable, set to 1 if trailing earnings per share is positive at the time of IPO, otherwise 0
Share Overhang	The ratio of retained shares divided by the number of shares in the IPO
Sales	Trailing annual sales/revenues in thousands of dollars at the time of IPO
Prior NASDAQ 15-day returns	The buy-and-hold returns of the CRSP NASDAQ value-weighted index of the 15 trading day period prior to the IPO date, ending on day t-1.

Table 1. IPO Characteristics – Predictors from Prior Studies

The target variables in this study include the pre-IPO price adjustment (ΔP) and the post-IPO initial returns (IR) defined as follows:

$$\Delta P = \frac{P_{ipo} - P_{mid}}{P_{mid}}, IR = \frac{P_{1day} - P_{ipo}}{P_{ipo}} \quad (1)$$

P_{mid}, P_{ipo}, and P_{1day} are the mid-point of the initial offering price range, the final offering price in the 424B4 filings, and the first-day trading price, respectively. The initial offering price range is obtained from the S-1 filings. Table 1 shows the control variables used in this study, based on extant IPO pricing literature.

We next describe the *FOCAS-IE* analytical framework developed and implemented to conduct text analytics, involving Natural Language Processing (NLP) and Information Extraction (IE) activities, followed by predictive modeling on the sample data.

FOCAS-IE Framework

Drawing on the needs of this study and the research gaps identified in Section 2, we articulate following design requirements for the *FOCAS-IE* analytical framework: (1) *Parsing the filings* - The first step of parsing is to read the HTML tags in the documents, the second step is to conduct a deep NLP parsing to identify the linguistic elements (i.e. *tokens/words*, *sentences*) in the textual contents of the filings. (2) *Segmentation of the documents* – Segmentation involves partition key relevant sections within each prospectus, including the MD&A section. (3) *Named entity recognition* - Three types of named entities need to be identified in this study, namely, *features*, *forward-looking indicators*, and *sentiment-indicative words*. Features refer to the major factors discussed in the MD&A sections that affect the IPO pricing strategy. Forward-looking indicators ensure that the sentences are discussing about the future outlook. Sentiment-indicative words denote the attitudes of the issuer/underwriters toward certain feature(s). (4) *Relationship recognition*: We need to further discover the relationships between the three types of entities mentioned. We are interested in two levels of relationships for this study: the first level relationships are co-occurrence relationships, which imply that at least one instance of each of three types

of terms appears at a certain contextual level (e.g., *sentence-level*). The second level relationship is modification, implying that the forward-looking indicator(s) and the sentiment-oriented word(s) are indeed describing a certain feature and they co-occur in a context.

The *FOCAS-IE* framework is depicted in Figure 1. The framework is based on GATE (General Architecture for Text Engineering) as the linguistic platform (Cunningham et al. 2002). *Deep Parsing* is conducted using GATE's native IE system called ANNIE, while *Segmentation* is conducted using GATE's pattern-matching rule engine, namely JAPE (Java Annotations Patterns Engine).

We have developed an underlying knowledge structure, namely the '*Prospectus Ontology*', for recognizing relevant name entities from the annotated sections. The prospectus ontology has three major classes, namely features, forward-looking indicators, and sentiment-oriented words. Financial domain experts were asked to help us building the seed concept list for the prospectus ontology, then ontology enrichment techniques are applied for ensuring accuracy and coverage. We have five features in this study: *sales/revenues*, *costs/losses*, *investments*, *net incomes*, and *cash flows*. Related terms of each feature are listed in the prospectus ontology as properties; a total of 50 terms are included for all the 5 features together after the ontology enrichment (e.g., *market shares/earnings belongs in sales/revenues*). 25 terms indicating the prediction/anticipation of future situations are categorized under the class of forward-looking indicators (e.g., *foresee, estimate, expect*). As for the sentiment-indicative words, we use three word lists from Loughran & McDonald (2011) (*L&M Wordlist*), which are *positive*, *negative*, and *uncertain*, respectively. The L&M Wordlist is specialized and widely accepted in the Finance domain. The Prospectus Ontology is enriched via an ontology enrichment approach (Tao et al. 2015). Building on a seed concept list constructed by domain experts, the ontology is enriched (by discovering new ontological classes, i.e. new features) and populated (by associating mentions of ontological classes in the corpus with the class itself, i.e., earnings → sales/revenues). The enriched ontology are clarified via a word sense disambiguation algorithm, and validated against human judgments. After applying disambiguation and enrichment techniques, the *positive* word list contains 144 words (e.g., *pleasant*, *ideal*, and *honorable*), the *negative* word list contains 746 words (e.g., *dangerous*, *defective*, and *tragic*), and the *uncertain* word list contains 119 words (i.e. *inexact*, *undecided*, and *intangible*). Given its wide acceptance in studies published in reputed avenues, the L&M wordlist is deemed to be accurate in identifying sentiment indicators (Ferris et al. 2013; Loughran and Mcdonald 2011; Loughran and McDonald 2013). However, as acknowledged in these studies, the accuracy of capturing the overall sentiment seems to be under expectation.

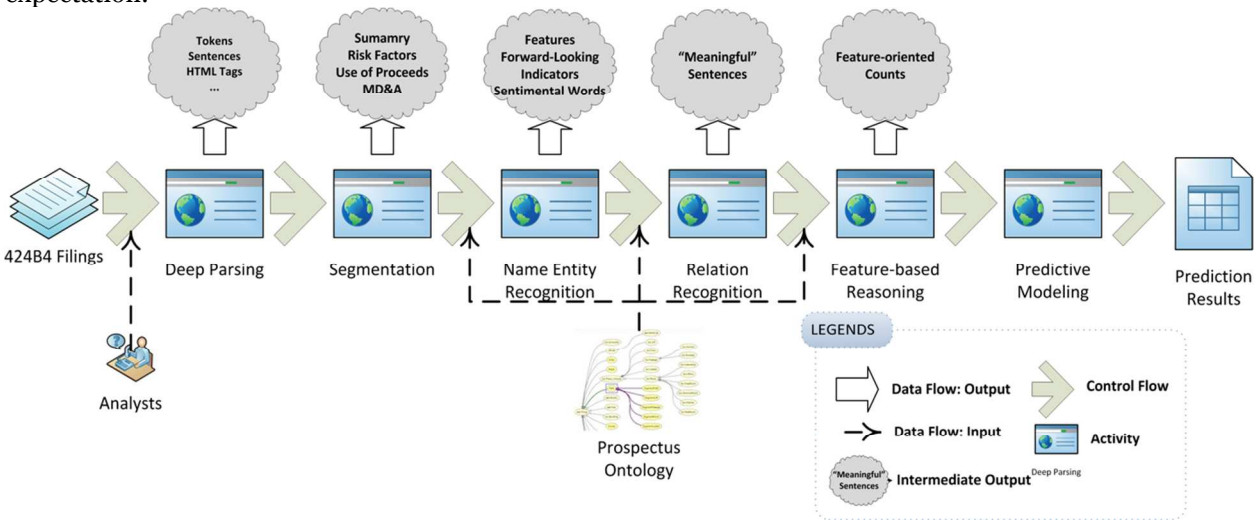


Figure 1. FOCAS-IE Analytical Framework

Based on the prospectus ontology, relevant named entities are identified using GATE language resources as follows. The Snowball stemmer is used to pick up the stem of each word; while APOLDA and onto-gazetteers are used to provide the semantic annotations based on the terms contained in the prospectus ontology. Next, the context-aware relation recognition using JAPE involves identifying the relations among these terms in the context of sentences. A sentence is annotated as "*informative*" if and only if at least one instance of each major class (*features*, *forward-looking indicators*, and *sentimental words*)

appears in it. The modification relations are determined through feature-based reasoning using JAPE, i.e. if only the sentimental words appear in the *n bag-of-words* (*n* adjacent words of the target term, in this study, n is set to 5) of the feature, we consider the sentimental words modifies the feature. We ensure the quality of the *FOCAS-IS* framework through a hybrid (automatic and manual) checking approach, incorporating exception handling where needed.

The *feature-based reasoning* module returns the counts of "*informative*" sentences by features and sentiments. For each 424B4 filing for an IPO, a total of 15 counts are returned (5 features X 3 sentiments). It may be noted that a sentence might overlap over different features/sentiments – for instance, a sentence might discuss both *sales* and *net income*, or contains a transition structure where the first clause is denoting a positive sentiment while the second one is negative. The counts are then normalized to account for variation in lengths of different MD&A sections. Following this, the *predictive modeling* module analyzes the directions of the changes (remain unchanged, positive price revision, negative price revision) in both pre-IPO price adjustments and post-IPO initial returns considering the 15 feature-sentiment counts as predictors generated previously. This predictive modeling is detailed in the next section.

IPO Valuation Predictive Modeling

We first calculate the percentage of "*informative*" sentences with respect to the overall total sentences in the MD&A section (*MDAWeight*). On average, 9.85% (standard deviation: 2.81%) of the total sentences in the MD&A sections are categorized as "*informative*" (feature-oriented, forward-looking, and sentiment-indicative), with a similar median of 9.56%. We further break the "*informative*" sentences down by the selected features (*sales/revenues, costs/losses, investments, net income*, and *cash flows*) and three sentiments (*positive*, *negative*, and *uncertain*). 15 variables representing all the pairs of (*feature*, *sentiment*) are computed (e.g., the percentage of forward-looking sentences discussing the feature *sales/revenues* with the *positive* sentiment is denoted as *Sale*Positive*. The descriptive statistics indicate that the percentages of the feature *costs/losses* are relatively higher than other features, which suggests that managers tend to discuss more about the costs and losses of the company's future operations. Also, four out of five features have more *positive* sentiments than *negative* sentiments (except for *net income*), which is consistent with the findings in Hanley and Hoberg (2010) that the MD&A sections are generally positive in tone. The exception of '*net income*' feature suggests that managers are tending to be more conservative discussing about future incomes.

In order to illustrate the practical value of the extracted information from the MD&A sections, we leverage predictive modeling techniques to predict the trends of pre-IPO price adjustments and post-IPO initial returns, given the feature-oriented, sentimental forward-looking information extracted from the MD&A sections of the Form 424B4 filings.

The target variables are binary variables, which are created based on the values of ΔP and IR in our sample, which are defined as follows:

- Pre-IPO Price Revision (*PRCREV*): set to 0 if the final offering price is lower than or equal to the midpoint of the initial offering price range ($P_{ipo} \leq P_{mid}$), set to 1 if the final offering price is greater than the midpoint of the initial offering price range ($P_{ipo} > P_{mid}$);
- Post-IPO First Day Return Change (*X1stDay*): set to 0 if the final offering price is lower than or equal to the midpoint of the initial offering price range ($P_{1Day} \leq P_{ipo}$), set to 1 if the final offering price is greater than the midpoint of the initial offering price range ($P_{1Day} > P_{ipo}$).

We use *PRCREV* and *X1stDay* as the target variables in our predictive models, respectively. The 15 feature-sentiment variables are used in each model as predictors, where are the IPO characteristic variables are included as control variables. The sample data of 513 observations is randomly divided into training (355 observations) and testing set (158 observations). Four widely accepted predictive modeling techniques are adopted in this study, namely *Decision Tree* (DT), *Artificial Neural Networks* (ANN), *k-Nearest-Neighbors* (k-NN), and *Ensemble* (EN). These models also include the 7 IPO characteristic variables as control variables (see Table 1).

Several steps are conducted in order to handle the *overfitting* and *unbalanced class* issues in the data. To avoid overfitting of the models to the training dataset, we employed the partitioning approach in which the data is split into training, validation, and testing partitions (Zhong, Li, & Wu 2012). The training partition is used for model building, while the validation dataset is used to compare and assess models during the model selection process. Finally, the testing dataset is used for assessment of selected models. We used a commonly accepted rule of thumb approach is partitioning our data into training, validation, and testing data sets with the ratio of 65% (336 records), 20% (104 records), and 15% (77 records). Furthermore, since the distributions of the classes in the target variable are noted to highly unbalanced, we perform oversampling on the minority class in the training data set to balance the data.

In order to illustrate the practical value of the extracted information from the *MD&A* sections, we leverage predictive modeling techniques to predict the trends of pre-IPO price adjustments and post-IPO initial returns, given the feature-oriented, sentimental forward-looking information extracted from the *MD&A* sections of the Form 424B4 filings.

Hypotheses H1 and H2 are proposed as follows.
H1: The 15 feature-sentiment pairs are effective predictors of pre-IPO price revisions (*PRCREV*); and
H2: The 15 feature-sentiment pairs are effective predictors of post-IPO price changes (*X1stDay*).
Above hypotheses are tested through the validations of the predictive models and results. If the models are efficient, and predictions are accurate, the hypotheses are supported, and vice versa.

The hypotheses proposed above are tested through the comparison of the predictive results from models with and without the aforementioned features. We discuss the results of the predictive models as well as the validation of the features in the following section.

Results and Discussion

The contingency tables (1: positive, 0: negative) for each of the predictive models are shown in Table 2. We rely on the comprehensive validation metric, namely the F-score, on the validation data set, for model comparison and selection purposes.

First, pre-IPO price revision (*PRCREV*) is selected as the target variable in the predictive models. As discussed above, oversampling is used for the imbalanced *PRCREV* in the training data set. Thus, the training data set has 436 data points, the validation data set has 92 data points, and the testing data set has 84 data points. We trained two sets of predictive models: one set with the MD&A features and the IPO Characteristics as predictors; while the other just with the IPO characteristics. As discussed above, the comparison between these two sets tests hypothesis H1.

The confusion matrices of the two sets of models are reported in Table 2 (a: Models with MD&A features and IPO Characteristics; b: Models with IPO Characteristics only). Based on the confusion matrices from Table 2, we compute the accuracy metrics for the two sets of predictive models, which include *precision*, *recall*, *accuracy*, and *F-score* (shown in Table 3). These metrics are widely adopted in the context of Machine Learning, including predictive modeling, as accuracy metrics. As discussed above, the F-scores from the validation data set are used for model comparison and selection purposes. From Table 3, it is clear that the F-scores are higher with the predictive models using both MD&A features and IPO characteristics as predictors, comparing to the predictive models with only IPO characteristics, with the validation data set. Thus, H1 is *supported*, which indicates that the extracted MD&A features are more effective in terms of predicting pre-IPO price revisions. Further, the predictive modeling technique *decision tree* (DT) achieved the highest value of F-score; thus, DT is selected as the final model for predicting PRCREV with both MD&A features and IPO characteristics as predictors.

Similarly, predictive models are constructed with post-IPO initial returns (X1stDay) as the target variable. One thing worth noting is that the IPO characteristic *Up Revision* is included in the predictive models as an additional predictor. The rationale behind this design decision is that pre-IPO price revisions have substantial effects on post-IPO initial returns, which is consistent with Loughran and McDonald (2013). The oversampling technique is employed in the training data set, which yields results in a training data set with 643 data points, a validation data set with 92 data points, and a testing data set with 84 data points. Two sets of predictive models are trained based on the data set, in order to test hypothesis H2.

PRCREV	(a) Models with MD&A Features and IPO Characteristics								(b) Models with IPO Characteristics Only							
	Training Data Set				Validation Data Set				Training Data Set				Validation Data Set			
Model	**FP**	**TP**	**FN**	**TN**	**FP**	**TP**	**FN**	**TN**	**FP**	**TP**	**FN**	**TN**	**FP**	**TP**	**FN**	**TN**
DT	106	172	32	126	17	48	13	14	62	126	79	170	11	34	27	19
EN	52	148	56	180	12	40	21	19	59	117	87	173	12	35	26	18
ANN	49	150	54	183	15	40	21	16	52	122	83	180	10	32	29	20
k-NN	36	129	75	196	12	29	32	19	34	137	67	198	10	27	34	20

Table 2. Confusion Matrices for Classification of PRCREV

PRCREV	(a) Models with MD&A Features and IPO Characteristics								(b) Models with IPO Characteristics Only							
	Training Data Set				Validation Data Set				Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.619	**0.843**	0.683	0.714	0.738	**0.787**	**0.674**	**0.762**	0.670	0.615	0.677	0.641	0.756	0.557	0.582	0.642
EN	0.74	0.725	0.752	0.733	**0.769**	0.656	0.641	0.708	0.665	0.574	0.665	0.616	0.745	0.574	0.582	0.648
ANN	0.754	0.735	**0.764**	**0.744**	0.727	0.656	0.609	0.69	0.701	0.595	0.691	0.644	0.762	0.525	0.571	0.621
k-NN	0.782	0.632	0.745	0.699	0.707	0.475	0.522	0.569	0.801	0.672	0.768	0.731	0.762	0.442	0.571	0.559

Table 3. Confusion Matrices and Accuracy Metrics for Classification of PRCREV

X1stDay	(a) Models with MD&A Features and IPO Characteristics								(b) Models with IPO Characteristics Only							
	Training Data Set				Validation Data Set				Training Data Set				Validation Data Set			
Model	**FP**	**TP**	**FN**	**TN**	**FP**	**TP**	**FN**	**TN**	**FP**	**TP**	**FN**	**TN**	**FP**	**TP**	**FN**	**TN**
DT	29	198	80	336	9	44	28	11	29	172	106	336	6	37	35	13
EN	5	185	93	360	6	46	26	14	0	153	125	365	1	34	38	18
ANN	16	208	70	349	6	44	28	14	0	161	117	365	3	30	42	16
k-NN	49	148	130	316	9	38	34	11	55	151	127	310	10	37	35	9

Table 4. Confusion Matrices for Classification of X1stDay

X1stDay	(a) Models with MD&A Features and IPO Characteristics								(b) Models with IPO Characteristics Only							
	Training Data Set				Validation Data Set				Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.872	0.712	0.83	0.784	0.83	**0.611**	0.598	0.704	0.856	0.619	0.790	0.718	0.860	0.514	0.549	0.643
EN	0.974	0.665	0.848	0.791	0.885	0.639	**0.652**	**0.742**	1.000	0.550	0.806	0.710	0.971	0.472	0.571	0.636
ANN	0.929	**0.748**	**0.866**	**0.829**	0.88	**0.611**	0.63	0.721	1.000	0.579	0.818	0.733	0.909	0.417	0.505	0.571
k-NN	0.751	0.532	0.722	0.623	0.809	0.528	0.533	0.639	0.733	0.543	0.717	0.624	0.787	0.514	0.505	0.622

Table 5. Accuracy Metrics for Classification of X1stDay

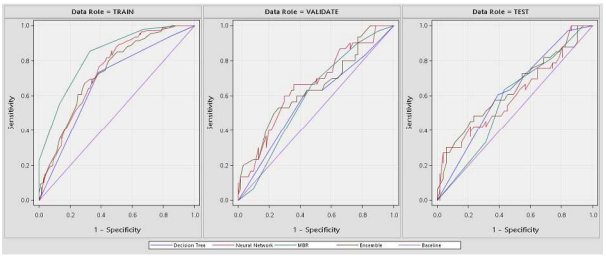
The confusion matrices are reported in Table 4, and the accuracy metrics toward these models are reported in Table 5. It is observed that the F-scores with the validation data set are higher using both MD&A features and IPO characteristics as predictors. Thus, H2 is *supported*, indicating that the feature-sentiment pairs are informative in the sense of predicting post-IPO initial returns. Moreover, ensemble (EN) is selected as the final model of predicting post-IPO initial returns.

For testing the performance of the predictive models, we select *Area Under Curve* (AUC, calculated from *Receiver Operating Characteristics*, ROC) to evaluate the model effectiveness. AUC is also used to interpret the results of the predictions, which is consistent with prior related studies (Khansa and Liginlal 2011; Wasikowski and Chen 2010). The values of the AUC metric with respect to the predictive models regarding the final selected models are reported in Table 6; while the ROC curves upon the partitioned data sets and all for modeling techniques are illustrated in Figure 2. The efficiency metric (AUC) validates the predictive models, as well as supports the acceptances of H1 and H2. As an interpretation, for predicting both pre- and post-IPO pricing trends, using *MD&A features*, with the help of IPO characteristics, the prediction possesses better quality than a random prediction. This indicate that the features selected in the *MD&A* sections are informative in terms of predicting the pre-IPO price revisions and the first day initial returns, which is consistent with prior related studies (Ferris et al. 2013; Hanley & Hoberg 2010).

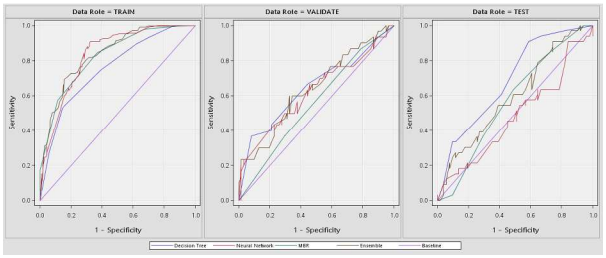
Target Variable	Predictors	Modeling Technique	AUC
PRCREV	MD&A Features + IPO Characteristics	DT	0.689
X1stDay	MD&A Features + IPO Characteristics	EN	0.646

Table 6. Efficiency Metric (AUC) of Final Selected Models

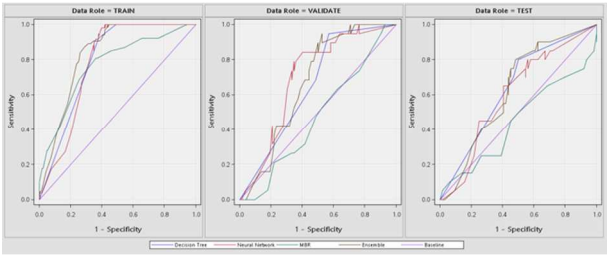
Figure 2 below illustrates the predictive models with MD&A features and IPO characteristics as predictors delivering higher AUC values, comparing to the ones with only IPO characteristics as predictors, which also confirms that the MD&A features add value in terms of predicting pre- and post-IPO pricing trends.



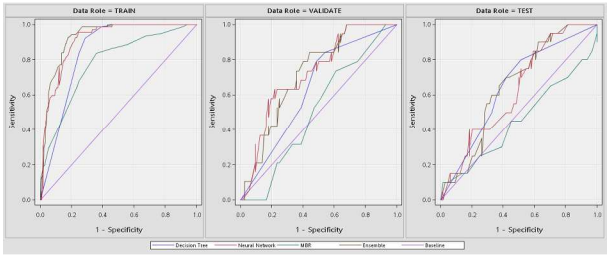
(a) AUC Curves of Predictions toward PRCREV using IPO Characteristics as Predictors Only



(b) AUC Curves of Predictions toward PRCREV using MD&A Features and IPO Characteristics as Predictors



(c) AUC Curves of Predictions toward X1stDay using IPO Characteristics as Predictors Only



(d) AUC Curves of Predictions toward X1stDay using MD&A Features and IPO Characteristics as Predictors

Figure 2. AUC Curves of Predictive Models

The findings from our predictive analysis extend the findings from current literature (i.e. Laughran & McDonald (2013), Hanley & Hoberg (2010)), by providing insights regarding the disclosed information regarding the future organizational performances with finer granularity. As a practical implication, it is clear that under-experienced investors and underwriters should spend more time reading and writing statements regarding the MD&A features, particularly those that are *forward-looking* and *sentiment-subjective*. Also, the findings are valuable in guiding the direction for future studies. It is evident that

further study of predictive ability of individual MD&A features for predicting pre- and post-IPO pricing trends is warranted. Furthermore, by comparing Figure 2(a) to 2(c), and 2(b) to 2(d) respectively, we can observe the improvement in predictive ability of the models. This can be arguably because of the inclusion of Up Revision in the models for post-IPO returns. In other words, the pre-IPO price revisions are taken into account as a predictor while predicting post-IPO returns. This observation points out a direction for future study in terms of employing time series analysis for understanding the IPO pricing trends in a longitudinal manner.

Conclusion

This study focused on studying impact of management's confidence reflected in MD&A sections in the IPO prospectus on IPO valuation. The key contributions of the study include an analytical framework, *FOCAS-IE* that integrates essential NLP and IE components, while including novel components. We propose a sentiment-based metric as a proxy for management's tone and confidence toward the company's future outlook. We further develop a domain-specific ontology that guides our ontology-based IE approach along with guided feature derivation and eventual application for predictive modeling. The predictive models built using these sentiment-indicative features are shown to have good accuracy and performance, particularly when compared to prior study. In sum, our research provides an alternative method for understanding the IPO pricing strategies, through the sentiment-based text analytics of IPO prospectuses. We provide a means to proxy *en ante* IPO uncertainty, which is also useful for similar analyses. Methodology-wise, the novel information extraction and sentiment-based text analytics approach proposed in this article can be used in other domains, such as medicine or bioinformatics.

We acknowledge the limitations of our study as well as discuss future directions. Along the methodology dimension, we recognize the need to improve our reasoning mechanism to better depict the sentiments at sentence level by reducing the overlapping across features/sentiments, so that we can provide an even more accurate proxy of the management's tone. We are considering using artificial intelligence (AI) based document-understanding techniques for this purpose. From an application standpoint, we need to further test and apply our approach on a larger sample to get more comprehensive data. We are working on extending the study to include other key sections of the IPO prospectuses. Another future direction under consideration is to investigate other types of IPO prospectus (Form S-1 or other Form 424 variants).

REFERENCES

Arnold, T., Fishe, R. P. H., and North, D. 2010. "The Effects of Ambiguous Information on Initial and Subsequent IPO Returns," *Financial Management (Blackwell Publishing Limited)* (39:4), pp. 1497–1519.

Bhabra, H. S., and Pettway, R. H. 2003. "IPO Prospectus Information and Subsequent Performance," *The Financial Review* (38:3), pp. 369–397.

Cornelli, F., Goldreich, D., and Ljungqvist, A. 2006. "Investor Sentiment and Pre-IPO Markets," *Journal of Finance* (61:3), pp. 1187–1217.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. 2002. "GATE: an architecture for development of robust HLT applications," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, , pp. 168–175.

Ferris, S. P. S., Hao, Q., and Liao, M.-Y. 2013. "The Effect of Issuer Conservatism on IPO Pricing and Performance*," *Review of Finance* (17:3), pp. 993–1027.

Geva, T., and Zahavi, J. 2014. "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news," *Decision Support Systems* (57)Elsevier B.V., pp. 212–223.

Hagenau, M., Liebmann, M., and Neumann, D. 2013. "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decision Support Systems* (55:3)Elsevier B.V., pp. 685–697.

Hanley, K. W., and Hoberg, G. 2010. "The Information Content of IPO Prospectuses," *Review of Financial Studies* (23:7), pp. 2821–2864.

Hanley, K. W., and Hoberg, G. 2012. "Litigation Risk , Strategic Disclosure and the Underpricing of Initial Public Offerings," *Journal of Financial Economics* (103:2), pp. 235–254.

Khansa, L., and Liginlal, D. 2011. "Predicting stock market returns from malicious attacks: A comparative analysis of vector autoregression and time-delayed neural networks," *Decision Support Systems* (51:4)Elsevier B.V., pp. 745–759.

Li, F. 2010. "The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach," *Journal of Accounting Research* (48:5), pp. 1049–1102.

Lin, M.-C., Lee, A. J. T., Kao, R.-T., and Chen, K.-T. 2011. "Stock price movement prediction using representative prototypes of financial reports," *ACM Transactions on Management Information Systems* (2:3), pp. 1–18.

Loughran, T., and Mcdonald, B. 2011. "When is a Liability not a Liability ? Textual Analysis , Dictionaries , and 10-Ks," *Journal of Finance* (66:1), pp. 35–65.

Loughran, T., and McDonald, B. 2013. "IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language," *Journal of Financial Economics* (109:2), pp. 307–326.

Loughran, T., and Ritter, J. R. 2004. "Why Has IPO Underpricing Increased Over Time?," *Financial management* (33:3), pp. 1–47.

Ritter, J. R., and Welch, I. 2002. "A Review of IPO Activity, Pricing, and Allocations," *Journal of Finance* (LVII:4), pp. 1795–1828.

Schumaker, R. P., and Chen, H. 2009. "Textual analysis of stock market prediction using breaking financial news," *ACM Transactions on Information Systems* (27:2), pp. 1–19.

Shmueli, G., and Koppius, O. 2009. "The challenge of prediction in information systems research," *Robert H. Smith School Research Paper No. RHS 06-152*, .

Tao, J., El-gayar, O. F., Deokar, A. V., and Chang, Y. 2015. "Term Extraction and Disambiguation for Semantic Knowledge Enrichment : A Case Study on Initial Public Offering (IPO) Prospectus Corpus," in *Proceedings of HICSS 48' -- 2015 48th Hawaii International Conference on System Sciences*, , pp. 3719–3728.

Wasikowski, M., and Chen, X. 2010. "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering* (22:10), pp. 1388–1400.

Zhong, N., Li, Y., and Wu, S. 2012. "Effective Pattern Discovery for Text Mining," *IEEE Transactions on Knowledge and Data Engineering* (24:1), pp. 30–44.