Textual Information and IPO Underpricing: A Machine Learning Aapproach

Katsafados, Apostolos G. and Androutsopoulos, Ion and Chalkidis, Ilias and Fergadiotis, Manos and Leledakis, George N. and Pyrgiotakis, Emmanouil G.

Athens University of Economics and Business, Athens University of Economics and Business, Athens University of Economics and Business, Athens University of Economics and Business, Athens University of Economics and Business, University of Essex

27 October 2020

Textual information and IPO underpricing: A machine learning approach

by

Apostolos G. Katsafados
Department of Accounting and Finance
Athens University of Economics and Business
Greece

Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business
Greece

Ilias Chalkidis
Department of Informatics
Athens University of Economics and Business
Greece

Manos Fergadiotis
Department of Informatics
Athens University of Economics and Business
Greece

George N. Leledakis*
Department of Accounting and Finance
Athens University of Economics and Business
Greece

and

Emmanouil G. Pyrgiotakis
Essex Business School
University of Essex
U.K.

This version: October, 2020

*Corresponding author: Department of Accounting and Finance, School of Business, Athens University of Economics and Business, 76 Patission Str., 104 34, Athens, Greece; Tel.: +30 210 8203459. E-mail addresses: katsafados@aueb.gr (A. Katsafados), ion@aueb.gr (I. Androutsopoulos), ihalk@aueb.gr (I. Chalkidis), fergadiotis@aueb.gr (M. Fergadiotis), gleledak@aueb.gr (G. Leledakis), e.pyrgiotakis@essex.ac.uk (E. Pyrgiotakis). We would like to thank Athanasios Episcopos, Prodromos Malakasiotis, Thanos Verousis, and the participants at the 2017 National Conference of the Financial Engineering and Banking Society (FEBS) for their valuable comments and suggestions. Apostolos Katsafados acknowledges financial support co-financed by Greece and the European Union (European Social Fund- ESF) through the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the State Scholarships Foundation (IKY). George Leledakis greatly acknowledges financial support received from the Research Center of the Athens University of Economics and Business (EP-2256-01). All remaining errors and omissions are our own.

Textual information and IPO underpricing: A machine learning approach

Abstract

This study examines the predictive power of textual information from S-1 filings in explaining IPO underpricing. Our empirical approach differs from previous research, as we utilize several machine learning algorithms to predict whether an IPO will be underpriced, or not. We analyze a large sample of 2,481 U.S. IPOs from 1997 to 2016, and we find that textual information can effectively complement traditional financial variables in terms of prediction accuracy. In fact, models that use both textual data and financial variables as inputs have superior performance compared to models using a single type of input. We attribute our findings to the fact that textual information can reduce the *ex-ante* valuation uncertainty of IPO firms, thus leading to more accurate estimates.

JEL classification: C63, G12, G14, G40

Keywords: Initial public offerings; First-day returns; Machine learning; Natural language processing

1. Introduction

One of the most heavily examined issues in the corporate finance literature is the underpricing of initial public offerings (IPOs). Underpricing is estimated as the percentage difference between the offer and the closing price at the end of the first-trading day (Ljungqvist, 2007). When the closing price is higher than the offer price, the IPO is considered to have been underpriced. In their early studies, Logue (1973) and Ibbotson (1975) find that when firms go public, they tend to sell their shares at a substantial discount. This underpricing discount is also confirmed by subsequent empirical studies (Ritter and Welch, 2002; Loughran and Ritter, 2004; Ljungqvist, and Wilhelm, 2005; Banerjee et al, 2011; Loughran and McDonald, 2013; Butler et al., 2014). Despite this extensive research however, there is still much to explore on how the IPO shares are priced (Hanley and Hoberg, 2010).

The pricing of an IPO is a difficult task for two reasons. First, the issuing firms have no observable market price prior to the IPO, and second, they have little or no historical data (Ibbotson et al., 1994). Apparently, this lack of available information leads to higher valuation uncertainty for IPO firms. Hence, the higher the valuation uncertainty, the higher the underpricing, as investors are less able to accurately price the newly issued stock (Beatty and Ritter, 1986). This argument builds on the winner's curse model of Rock (1986). According to this model, there is information asymmetry between different types of investors, and as a result, underwriters underprice IPOs to ensure the participation of uninformed investors in the market. In fact, the vast majority of work in the field explains IPO underpricing under the information asymmetry perspective (Ljungqvist and Wilhelm, 2005).

A recent strand of the literature examines whether the textual information of IPO prospectuses affects the ability of investors to price a new issue. When U.S. firms go public, the Securities and Exchange Commission (SEC) requires them to submit the IPO prospectus,

or S-1 filing, on the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. This requirement ensures that investors are properly informed regarding the issuing firms' valuation, future business strategies, and potential risks (Ferris et al., 2013). In fact, the IPO prospectuses are the most informative sources for investors, as the amount of publicly available information for new issues is rather limited (Ding, 2016). In this respect, Hanley and Hoberg (2010) find that a more informative IPO prospectus leads to more accurate offer prices and less underpricing. Loughran and McDonald (2013) proxy the *ex-ante* valuation uncertainty of IPOs firms using the tone of the S-1 filing. The authors find that higher percentages of uncertain, weak modal, and negative words in the IPO prospectuses are associated with higher levels of underpricing.

At this point, it is worth mentioning that the majority of prior studies examine IPO underpricing under an econometric set-up. However, there are some recent papers that investigate this issue using machine learning algorithms (Basti et al., 2015; Quintana et al., 2017). Among them, there is a handful of studies which examine the IPO underpricing as a binary classification task (Cheng et al., 2007; Chen et al., 2010; Kim et al., 2019). In other words, these studies use machine learning algorithms in order to distinguish underpriced from overpriced IPOs. In several finance classification tasks, machine learning algorithms have advantages over traditional techniques for two reasons: (1) they often can produce accurate predictions (Mai et al., 2019), and (2) they do not depend on statistical distributions (Pasiouras et al., 2010). To the best of our knowledge however, a common element of such studies in IPO underpricing classification prediction is that they ignore the textual information of IPO prospectuses, as they merely focus on financial variables. Considering that S-1 filings contain vital information regarding the IPO firms' valuation, the IPO prediction literature leaves a lot of available information unexploited.

The primary aim of this paper therefore is to examine whether and to what extent the

textual information of the S-1 filings can improve the predictive power of machine learning algorithms in an IPO classification task. We choose to treat the IPO underpricing issue as a classification task, inspired by the fact that the literature on investment strategies primarily focuses on the direction of stock price movements instead of the magnitude (Nardo et al., 2016).

To address our research question, we utilize a large and comprehensive sample of 2,481 U.S. IPOs over the period 1997 to 2016. For the purpose of our analysis, we extract textual features from the 4 major sections of the S-1 filing, their combination, and the entire S-1 filing. In addition, we collect data on financial variables frequently-used in the IPO literature. Then, we use both sources of data separately or together as inputs in our machine learning algorithms, and we evaluate their predictive ability according to the out-of-sample performance of our models. In our classification task, we use the following machine learning models: (1) support vector machine, (2) logistic regression, (3) random forest, and (4) multilayer perceptron.

One issue that emerges when we attempt to combine the plethora of textual features with financial variables is the curse of dimensionality. More precisely, the high dimensional space of our textual features overrules the importance of financial variables, leading to less accurate estimates. To overcome this issue, we employ the singular value decomposition dimensionality reduction technique. Notably, when we do so, we find that textual data can effectively complement financial variables in our IPO classification task. In fact, when both sources of data are inserted as inputs in our machine learning algorithms, we achieve out-of-sample accuracy scores that in some instances exceed 70%. In practice, the magnitude of this accuracy score means that our models are able to correctly distinguish between overpriced and underpriced IPOs in more than 70% of the future cases. In terms of our models performance, random forest produces the highest accuracy scores, followed by logistic

regression and multilayer perceptron.

As an additional step in our empirical analysis, we also investigate whether the tone of the S-1 filing can improve the predictive power of our machine learning algorithms. We measure the S-1 tone using the six sentiment word lists of Loughran and McDonald (2011). Our findings indicate that when sentiment scores are used as mixed inputs with financial variables in our classification models, the prediction accuracy is comparable to our aforementioned findings. Interestingly, each sentiment score performs better in a different section of the S-1 filling, a finding which is consistent with prior relevant studies (Hanley and Hoberg, 2012; Ferris et al., 2013; Brau et al., 2016). Finally, our results remain robust to a series of robustness tests that deal with sample selection and methodological issues.

Our findings are important to all key parties of an IPO transaction. In fact, a machine learning model designed to predict underpricing can profoundly benefit investors, managers, and underwriters, as it would allow them to *ex-ante* identify whether the IPO will be underpriced, or not. On the one part, our study complements the existing literature that examines IPO underpricing under the perspective of information asymmetry. We propose that the textual information of the S-1 filing reduces the valuation uncertainty of IPO firms, and helps the involving parties to value the issuing firm more accurately. On the other part, our research contributes to the literature, as we introduce new methodological insights on how textual disclosure can efficiently be combined with numerical data as mixed inputs in the machine learning algorithms.

The remainder of the paper is organized as follows. Section 2 briefly discusses the relevant literature of IPO underpricing. Sections 3 and 4 describe our sample collection and methodology. Section 5 discusses our empirical results, and Section 6 concludes the paper.

2. Literature Review

Prior empirical studies indicate that underpricing is a persistent phenomenon in the U.S. IPOs market. In fact, during the previous decades, underpricing in the U.S. averages between 7% and 65% (Ritter and Welch, 2002; Loughran and Ritter, 2004; Ljungqvist, and Wilhelm, 2005; Loughran and McDonald, 2013; Butler et al., 2014). However, the level of underpricing varies substantially across time periods (Loughran and McDonald, 2013).

In their early study, Ritter and Welch (2002) examine a sample of 6,249 U.S. IPOs from 1980 to 2001, and report an average underpricing of 18.8%. Loughran and Ritter (2004) use a sample of 6,391 IPOs occurred from 1980 to 2003, and breakdown their examination period into four sub-periods. Their findings point out a substantial degree of variation in the level of underpricing across time periods. More precisely, in the 1980s, the average IPO was underpriced by 7%. During 1990-1998, the average underpricing was 15%, and then jumped to 65% in the dot-com bubble years (1999 and 2000). In the post-bubble years (2001-2003) however, first-day returns dropped to an average of 12%. In a more recent study, Loughran and McDonald (2013) analyze the first-day performance of 1,887 U.S. IPOs for the period 1997-2010, and find a mean first-day return of 34.8%. Consistent with previous empirical evidence, they document that the more underpriced IPOs took place in the dot-com bubble years.

2.1. IPO underpricing and information asymmetry

Information asymmetry is considered to be an important determinant of IPO underpricing (Banerjee et al., 2011). According to the winner's curse model of Rock (1986), some investors are better informed regarding the true value of the shares being offered than other investors. In the context of this model, informed investors only buy shares of attractive IPOs, whereas uninformed investors bid for every new share issued. This information asymmetry problem results in a winner's curse for uninformed investors. More precisely, uninformed

investors get the full supply of unattractive IPOs, while in attractive IPOs, their demand is partly crowded out by the informed investors. In that event, first-day returns for uninformed investors should be zero or even negative (Ritter and Welch, 2002). Thus, uninformed investors may choose not to bid for any IPO. Under this scenario, underwriters underprice IPOs in order to retain uninformed investors in the market.

Ritter (1984) and Beatty and Ritter (1986) extend the winner's curse model by focusing on firms' valuation uncertainty. More precisely, the authors hypothesize that underpricing increases with the level of *ex-ante* uncertainty regarding the value of the IPO firm, as it is more difficult for investors to correctly price the new issue. Interestingly, several empirical studies provide empirical support for this hypothesis, by using proxies to account for firms' valuation uncertainty. Among these proxies are: firm age (Ritter, 1984; Ljungqvist and Wilhelm, 2003; Loughran and Ritter, 2004; Chahine, 2008), growth opportunities as measured by price to earnings ratio (Chen et al., 2004; Hauser et al., 2006), or industry sector (Benveniste et al., 2003). Younger firms, with more growth opportunities, and/or high-tech firms should have higher levels of *ex-ante* valuation uncertainty, which translates to higher IPO underpricing (Engelen and Van Essen, 2010).

Information asymmetry can impact other key parties of an IPO transaction besides investors. In fact, Baron and Holmstrom (1980) and Baron (1982) focus on the information asymmetry between underwriters and the issuing firm. The authors suggest that higher levels of ex-ante uncertainty regarding the value of the IPO firm are associated with more information asymmetry between underwriters and issuers, which results in higher underpricing.

2.2. Ex-ante uncertainty and IPO prospectuses

There are several early studies that use information from IPO prospectuses to proxy for *ex-ante* valuation uncertainty. Such proxies include the number of uses of IPO proceeds (Beatty

and Ritter, 1986) or the number of risk factors declared in the prospectus (Beatty and Welch, 1996). Other studies attempt to link the information content of prospectuses with IPO underpricing. For instance, Bhabra and Pettway (2003) find that financial data included in the prospectus can explain part of the IPO performance. Furthermore, Leone et al. (2007) investigate whether the uses of IPO proceeds included in the prospectuses can relate to IPO underpricing. Their findings suggest that the more specific the use of proceeds disclosure, the less the *ex-ante* uncertainty, which in turn leads to less underpricing.

A more recent strand of the literature examines whether the textual information contained in IPO prospectuses can explain underpricing. Arnold et al. (2010) analyze the Risk Factors section of IPO prospectuses and find that soft information contained in this section is a strong determinant of first-day returns. Hanley and Hoberg (2010) use textual information based on the word content analysis of IPO prospectuses. Further, the authors decompose the prospectuses into their four main components: (1) Summary, (2) Risk Factors, (3) Use of Proceeds, and (4) Management Discussion and Analysis. Their results indicate that greater informative content results in more accurate offer prices and less underpricing. In addition, stronger disclosure of information in IPO prospectuses is associated with lower litigation risk (Hanley and Hoberg, 2012).

Loughran and McDonald (2013) emphasize on the tone of IPO prospectuses, using word lists as a proxy of *ex-ante* uncertainty. The authors document that higher levels of text uncertainty translate to higher first-day returns. Ferris et al. (2013) indicate that more conservative language in the IPO prospectus leads to higher underpricing. Moreover, Brau et al. (2016) find that the strategic tone of the IPO prospectus is related to underpricing. In fact, more frequent usage of positive and/or less frequent usage of negative strategic words is associated with higher first-day returns.

2.3. Machine learning approach of IPO underpricing

As mentioned in the previous section, there is a growing literature that relates textual information with first-day returns. However, all these studies use the textual content of IPO prospectuses in an econometric framework. Recently, there are a handful of papers which attempt to predict underpricing with the usage of machine learning algorithms (Quintana et al., 2017). Supervised machine learning algorithms learn from historical data in order to predict future outcomes. The benefits of these algorithms are that they are able to produce accurate predictions, and they work well with both numerical and categorical data (Bastı et al., 2015).

In Table 1, we provide a list of all papers that use machine learning algorithms in an IPO classification task. In their early study on IPO underpricing, Mitsdorffer et al. (2002) utilize several machine learning models, such as Bayesian classifications, support vector machines, decision trees, and artificial neural networks. More precisely, the authors conduct a classification task by developing models which try to identify IPOs with first-day returns of 50% or higher. In a similar fashion, several other studies use financial variables as inputs in machine learning algorithms to classify whether the IPOs will realize positive or negative first-day returns (Cheng et al., 2007; Chen et al., 2010; Kim et al., 2019).

In the non-IPO literature, there is an ongoing effort to combine textual information with financial variables in machine learning models (Mai et al., 2019). Yet, to the best of our knowledge, evidence along these lines is rather elusive when it comes to IPO classification prediction. In fact, there is one paper by Ly and Nguyen (2020), which uses only the prospectus sentiment (without the addition of financial variables) to predict whether an IPO will be underpriced, or not. The authors utilize several machine learning algorithms, such as random forest, decision trees, and logistic regression. Their findings indicate that most of these models are not able to produce accurate estimates, as only the logistic regression

consistently achieved an accuracy score of higher than 50%. In this paper therefore, we attempt to fill the gap in the literature by examining whether the inclusion of both textual information and financial variables in machine learning models can enhance our knowledge on IPO underpricing classification prediction.

Insert **Table 1** here

3. Data and textual analysis

3.1. Sample selection

We collect our IPO sample from Thomson Financial Securities Data. Our sample includes completed U.S. IPOs during the period 1997 to 2016.[1] In line with Loughran and McDonald (2013), we exclude all IPOs with an offer price of less than $5. Furthermore, all financial firms (savings institutions and banks), real estate investment trusts (REITs), American Depository Receipts (ADRs) and closed-end funds are excluded from our sample. After applying those criteria, our final sample consists of 2,481 IPOs. We obtain stock price data for the close of the first-trading day from the Center for Research in Security Prices (CRSP).

3.2. Matching process and datasets

Our sample includes 576 IPOs with negative first-day returns and 1,905 IPOs with positive first-day returns, respectively. Apparently, the higher proportion of underpriced IPOs suggests that our sample is imbalanced. Hence, if we apply random sampling, our estimates would be less efficient, as this approach will probably generate a sample consisting of more underpriced than overpriced IPOs (Palepu, 1986; Pasiouras et al., 2007; Veganzones and Severin, 2018). To get a better insight on this issue, Figure 1 depicts the imbalance rate of our sample. In each year of our examination period, the imbalance rate is the ratio of underpriced IPOs to overpriced IPOs. Hence, an imbalanced rate of 2 suggests that there are two times more underpriced IPOs than overpriced IPOs in a given year. Notably the highest imbalance

[1] The earliest offer date of our sample is on January 9, 1997 and the latest on December 14, 2016.

rate is observed during the dot-com bubble years.

Insert **Figure 1** here

To mitigate this imbalanced dataset problem, we adopt the undersampling approach as in Veganzones and Severin (2018). More precisely, this method generates a balanced subset from our original dataset by excluding observations from the majority category. This method is widely-used in several classification tasks in finance, such as bankruptcy or acquisition prediction (Barnes, 1998; Laitinen and Kankaanpaa, 1999; Doumpos et al., 2004; Neophytou and Mar Molinero, 2004; Pasiouras et al., 2007, 2010).

According to the undersampling approach, we need to match our 576 overpriced IPOs with an equal number of underpriced IPOs.[2] To do so, we follow the approach of Pasiouras et al. (2010), and we use time (year of the S-1 filing) as the matching criterion. This enables us to do direct comparisons between overpriced and underpriced IPOs, without having to control for any time effects. We do not include any other variables in our matching approach, because, if we use a financial variable for matching purposes, then we have to exclude it from our classification models (Hasbrouck, 1985).

3.3. Textual data and methodology

Textual data are obtained from S-1 filings from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR). The S-1 filing is the initial registration form for IPOs (IPO prospectus), as required by the SEC. We collect the S-1 filings for all IPOs of our sample, using a web-crawling algorithm.

3.3.1. Textual sources and parsing process

Knowledge retrieval from text is a quite sensitive and demanding process. All the retrieved IPO prospectuses are encoded in hypertext markup language (HTML). For each retrieved S-1, we follow the parsing procedure of Loughran and McDonald (2013), and we

[2] Imbens (1992) points out that an equally shaped sample may provide more relevant information than random sampling.

eliminate HTML formatting and any other non-textual information, such as embedded images or spreadsheets that might be present in the text (Bodnaruk et al., 2015). Furthermore, we remove all identified HTML tables, unless their alphabetic character content exceeds 85%.

Following Hanley and Hoberg (2010) and Ferris et al. (2013), we decompose the S-1 filings into their four key sections: Management Discussion and Analysis (MD&A), Risk Factors, Use of Proceeds, and Summary. In addition, we create an artificial S-1 filing by aggregating the four sections. Then, we are interested in examining the kind of information each section really contains. To do so, we apply Latent Dirichlet Allocation (LDA) to the four sections separately, in the spirit of Nguyen et al. (2015).[3] LDA is an algorithm used for topic extraction. In our case, LDA intuitively detects the most common topics in the S-1 corresponding section, and tries to find the words that belong to each topic. By inspecting the most frequent words, we manually assign an overall label to each topic to reflect its meaning. In fact, we investigate which topics are typically covered by each one of the four sections across all filings.

Table 2 presents the results of this analysis. Our findings indicate that each section contains specific information regarding the offering process. In particular, the MD&A section describes how managers intend to increase their share price, and boost future revenues and sales (Ferris et al., 2013). The Risk Factors section outlines the potential risks of the firm such as the uncertainty of future products, or regulatory issues (Ding, 2016). Next, the Use of Proceeds section indicates how the firm intends to use the money raised, such as acquisition events, or dividend payments. Finally, the Summary section briefly refers to the key aspects of the S-1 filing. Particularly, it describes the offering process and provides an overview of the balance sheet and cash flow statements (Ferris et al., 2013).

Insert **Table 2** here

[3] The authors focus on social media messages considering them as a mixture of hidden topics.

3.3.2. Pre-processing and Bag-of-Words

The extent to which pre-processing is vital regarding the performance of any classification algorithm is frequently-highlighted in the textual analysis literature (Nassirtoussi et al., 2014; Kumar and Ravi, 2016). In particular, pre-processing includes a range of sub-processes where the raw text is converted into meaningful inputs for our predictive models.

As a first step, we remove all acronyms, abbreviations, single letter words, numbers, punctuation marks, and stop words (Gandhi et al., 2019; Katsafados et al., 2020). This filtering procedure has the benefit of reducing the informational opaqueness of the textual inputs, which contributes to superior prediction performance. Furthermore, we impose a minimum occurrence threshold in order to remove words with low frequency (Schumaker and Chen, 2009). In the spirit of Mai et al. (2019), we take into account the 20,000 most frequent words of the S-1 filing. In fact, having an excessive number of textual features reduces the effectiveness of any learning algorithm and yields inferior results (Pestov, 2013).

At a second step, we need to convert our textual data into numerical units that a learning algorithm can understand (Mai et al., 2019). We do so because textual data, like natural language, have an unstructured format that cannot be inserted as input in our models. Hence, we follow the bag of words (BOW) approach in order to convert our unstructured textual data into inputs with explicit numerical structure. According to this approach, we tokenize text into words using the Natural Language Toolkit (NLTK). More precisely, we consider each unique word as a different feature, and we create a document-term matrix, where each row and column represent a document and a word, respectively; the value of each cell of the matrix is the value of the corresponding word feature in the particular document (Kumar and Ravi, 2016). We discuss how feature values are computed below.

One limitation of the BOW approach is that it does not effectively account for the presence of polysemous words in the text. Polysemous are words with multiple meanings.

Hence, a perfect model would also use a textual feature for each meaning of polysemous words. To alleviate this problem, we also employ word n-gram features, in effect using a bag of n-grams representation. This model is a set of sequential n tokens (Sun et al., 2017). The BOW representation is a special case of the bag of n-grams representation with n equal to 1 (unigram). For other values of n, we obtain bigrams (n equal to 2), trigrams (n equal to 3), as in Kumar and Ravi (2016). To some extent, word n-grams make word representations aware of their context (surrounding words), especially for larger values of n. However, n values larger than 3 are rarely used, because the size of the corresponding feature set (possible n-grams) increases exponentially. In this paper, we present results for unigrams and bigrams.[4]

Finally, we compute the values of the features, where we represent each textual feature with a numeric value. To do so, we employ the two classical term weighting schemes: (1) the term frequency (TF) normalized by document length, and (2) the term frequency-inverse document frequency (TF-IDF). The former calculates the proportion of each word in each document and assigns equal weight to each of them, while the latter downweights the TF scores based on the document frequency of each word in our sample of IPO prospectuses (Kearney and Liu, 2014; Nassirtoussi et al., 2014). If $TF(t_{ij})$ is the number of times a word i appears in a document j, divided by the total word count of the same document for normalization purposes, then we calculate *TF-IDF* weight of word i in the j^{th} document as follows:

$$\text{TF-IDF}(t_{ij}) = \text{TF}(t_{ij}) \times \left[-\log\left(\frac{n_i}{N}\right)\right]$$

where N represents the number of documents in our entire dataset, and n_i the total number of documents including at least one occurrence of the i^{th} word.

At this point it is worth mentioning that the TF-IDF approach is considered to be more effective than the TF approach, as it assigns lighter weights to very common words, which

[4] We have also used trigrams in our models with no substantial improvement in our results.

are often not useful for document classification purposes (Balakrishnan et al., 2010; Brown and Tucker, 2011; Kumar et al., 2012; Hagenau et al., 2013; Loughran and McDonald, 2016; Mai et al, 2019). Moreover, we also use as separate textual features the sentiment scores based on the LM word lists (Loughran and McDonald, 2011). As in Loughran and McDonald (2013), we compute sentiment scores using the TF approach.[5]

3.3.3. Financial variables

Besides textual data, we also use eight financial variables frequently-used in the relevant literature (Loughran and Ritter, 2004; Loughran and McDonald, 2013).[6] More precisely, we use the following control financial variables: (1) *Sales* is the logarithm of firm annual sales in the 12 months prior to the IPO; (2) *Positive EPS* is a dummy variable which equals to 1 if the IPO has positive earnings per share in the year before going public, and 0 otherwise; (3) *Share overhang* is the amount of shares retained divided by the amount of shares in the IPO;[7] (4) *Venture capital* is a dummy variable which equals to 1 if the IPO is backed by venture capital, and 0 otherwise; (5) *Prior Nasdaq 15-day returns* is the buy and hold returns of the Nasdaq index 15 trading days before IPO date; (6) *Up-revision* is the percentage upward revision from the mid-point of the filing range, if the offer price is higher than mid-point, and 0 otherwise; (7) *Top-tier* is a dummy variable which equals to 1 if at least one underwriter has been classified as top-tier according to the rankings of Carter and Manaster (1990), Carter et al. (1998) and Loughran and Ritter (2004), and 0 otherwise, and (8) *Days between S-1 and 1-trading day* is the logarithm of days between S-1 filing date and first-trading day.

Table 3 reports the summary statistics of our final (imbalanced) sample. In general, the

[5] In untabulated results, we repeat the analysis using the TF-IDF approach. The results were qualitatively similar.

[6] We use only those variables that have a statistically significant impact on first-day returns, as documented in previous empirical studies (Loughran and McDonald, 2013). In unreported analysis, we have also used the dummy variable *TECH,* which equals 1 for IPO firms in the technology sector, and 0 otherwise (Loughran and Ritter, 2004). The inclusion of this variable does not impact our results.

[7] According to Loughran and McDonald (2013), this variable may act as a proxy of scarcity. Fewer shares provided to the market for the initial offering are associated with stronger investors' demand for the stock.

statistics are comparable with Loughran and McDonald (2013) and Butler et al. (2014). *First-day returns* is the underpricing measure, and is calculated as the percentage change from the offer price to the closing price. The average IPO of our sample is underpriced by 31.1% with a corresponding median value of 12.5%.[8] Average annual firm sales are $451.9 million and the mean upward revision is in the order of 10.5%. In line with Gao et al. (2013), only 39% of our sampled firms have positive EPS in the year before the IPO. Moreover, 53% of our IPO firms are backed by venture capitalists and 78% of issuing firms use a top-tier underwriter. In addition, the average time interval between the S-1 filing and the first-trading day is approximately four months (115.9 calendar days). Finally, we also report summary statistics for the percentages of Loughran and McDonald's (2011) sentiment words lists.

Insert **Table 3** here

4. Machine learning models

To perform our classification task, we use the following machine learning algorithms: (1) support vector machine, (2) logistic regression, (3) random forest, and (4) multilayer perceptron. In this section, we will briefly describe the details of these techniques.[9]

4.1. Support vector machine

Support vector machine (SVM) is a machine learning algorithm developed by Vapnik (1998). SVM has been widely-used in various finance tasks, such as IPO underpricing prediction (Bastı et al., 2015; Quintana et al., 2017; Quintana et al., 2018), bankruptcy prediction (Min and Lee, 2005; Shin et al., 2005; Wu et al., 2007; Veganzones and Severin, 2018; Mai et al., 2019), time-series forecasting (Cao, 2003; Huang et al., 2005; Pai and Lin,

[8] The highest value of underpricing (697.5%) in our sample comes from the IPO of VA Linux Systems on December 9, 1999. The offer price was $30 and the closing price at the first-trading day was $239.25.

[9] In all our models, all the financial variables are standardized. Textual features are also standardized when they are combined with financial variables.

2005) and merger prediction (Pasiouras et al., 2008). In its simplest form, SVM is a non-probabilistic supervised linear classifier, which draws a decision boundary that has the form of a hyperplane in the original feature space. The training instances at the boundaries of the margin, or (when allowing 'slack' in the separation) inside the margin, or on the wrong side of the hyperplane are called support vectors. Finding the maximum margin hyperplane is a quadratic programming optimization problem, and the solution depends only on the support vectors. In the case of non-linearly separable data, it is common to use the SVM with non-linear kernel functions such as the radial basis function kernel (RBF). This approach ensures that the training data are projected in a higher dimensional space, where they become linearly separable (Nassirtoussi et al., 2014). For this reason, we repeat our experiments using: (i) a linear SVM, and (ii) an SVM with RBF kernel in our empirical study.[10]

4.2. Logistic regression

The logistic regression model (LOGIT) is probably the most popular predictive model in finance (Hasbrouck, 1985; Palepu, 1986; Ambrose and Megginson, 1992; Barnes, 1998; Powell, 2001; Espahbodi and Espahbodi, 2003; Pasiouras and Tanna, 2010; Veganzones and Severin, 2018; Mai et al., 2019; Ly and Nguyen, 2020). LOGIT estimates a non-linear sigmoid function between the binary output and the control variables. The model's parameters are learned by maximizing the conditional log-likelihood of the training data, typically using stochastic gradient ascent or variants. Regularization terms are also typically added to the log-likelihood to avoid overfitting the training data. We use L2 regularization, which subtracts the squared L2 norm of the weights vector from the log-likelihood. The mathematics behind this model is described as follows:

[10] The hyper-parameters of our SVM models are tuned based on the 5-fold cross-validation performance of the training set.

$$P(Y_{t+1} = 1 \mid X_{i,t}) = \frac{\exp\left(bo + \sum_{i=1}^{n} b_i X_{i,t} \right)}{1 + \exp\left(bo + \sum_{i=1}^{n} b_i X_{i,t} \right)}$$

where Y is the binary output (in our case Y equals 1 if the IPO is underpriced, and 0 otherwise), $X_{i,t}$ is a vector of n control variables at time t, b_i are parameters of the model, and b_0 is a bias term.

4.3. Random forest

Random forest (RF) is a machine learning algorithm suitable for both regression and classification tasks. It was initially developed by Breiman (2001) and is a variant of the Bagging ensemble learning method (Breiman, 1996). In our classification task, RF generates a number of uncorrelated decision trees trained on bootstrap copies of original samples by randomly choosing a subset of features. Each individual tree predicts a class and the category with the most votes becomes the output of the model (Mai et al., 2019). The benefit of this approach is that it reduces variance without increasing bias substantially. Notably, in the IPO literature, Quintana et al. (2017) highlight the high predictive power of RF models.

4.4. Multilayer perceptron

Recent work in the field of Natural Language Processing (NLP) is dominated by neural network models (Goldberg, 2017). One of the simplest kinds of neural networks are Multi-Layer Perceptrons (MLPs), which have also been used in previous work in the financial domain (Kumar and Ravi, 2016). Currently in NLP, more complex neural models are often used, such as Recurrent Neural Networks (Goldberg, 2017), and Transformer-based models. However, in this paper, we experiment with MLPs because they are more directly applicable to BOW text representations and, hence, also more directly comparable to the other models we consider. In an MLP, there is an input layer of neurons, in which our variables are

introduced as inputs into the network. In addition, there are one or more hidden layers.[11] Once the hidden layer receives the content from the input layer, non-linear functions are used before transferring the estimated values to the next hidden or the output layer. Finally, the output layer chooses the predictive class based on the received input from the hidden layers.[12] Unlike logistic regression, which can be seen as a degenerate form of MLP with no hidden layer, MLPs can learn non-linear functions. In the IPO literature, there are several studies that use MLPs to predict underpricing (Jain and Nag, 1995; Reber et al., 2005; Cheng et al.; 2007; Wang et al., 2018).

5. Empirical results and discussion

5.1. Evaluation

It is significant to ensure that our IPO underpricing prediction models are properly evaluated with respect to their out-of-sample performance. In this paper, we split the data into training and testing datasets, as in Mai et al. (2019). In line with previous studies, we partition our data by selecting 80% of our sample as the training set and the remaining 20% as the testing set (Geng et al., 2015; Doumpos et al., 2017; Routledge et al., 2017). In addition, we select our testing set from a future period rather than in random (Pasiouras et al., 2008; Pasiouras and Tanna, 2010). We do so, in order to test our model against a future period. In fact, the usefulness of a classification model depends on its ability to correctly classify objects in the future (Espahbodi and Espahbodi, 2003; Pasiouras et al., 2008).

Next, we describe the two measures we use to evaluate the out-of-sample performance of

[11] The networks with a large number of layers of hidden neurons are known as deep networks, thus leading to the terminology of deep learning (Goldberg, 2017).

[12] We use 5-fold cross-validation for hyper-parameter tuning. As a result, our MLP model has 3 hidden layers, each of which has 200 neurons. Given that MLP is a feedforward model that maps inputs (financial variables and textual features) to a binary outcome (underpricing or not), we apply backpropagation algorithms to train the model. Furthermore, we use cross-entropy as the loss function, Adam as the optimizer algorithm, and rectified linear unit (ReLU) as the activation function in each hidden layer. ReLU is defined as $f(x) = \max(0, x)$. Finally, we use early stopping to mitigate overfitting (Mai et al., 2019). To do so, we set aside 10% of training data as validation or development set.

our models. The first metric we use is the accuracy measure, which is widely-used in many finance tasks (Palepu, 1986; Mitsdorffer et al., 2002; Pasiouras et al., 2007; Pasiouras and Tanna, 2010; Pasiouras et al., 2010; Nguyen et al., 2015; Bastı et al., 2015; Mai et al., 2019; Ly and Nguyen, 2020). *Accuracy* ranges from 0 to 1. The higher the score, the better the out-of-sample performance of the model. Considering that our dataset is fully balanced, our models would be considered effective when they perform better than chance, which corresponds to an accuracy score of higher than 50%. In sum, *Accuracy* can be expressed as follows:

$$Accuracy = \frac{(|TP| + |TN|)}{(|TP| + |FP| + |FN| + |TN|)}$$

where *TP* is the number of observations correctly labeled as positive (underpriced) IPOs by the classifier, *TN* is the number of observations correctly identified as negative (overpriced) IPOs by the model, *FP* the number of observations incorrectly labeled as positive IPOs by the classifier and *FN* is the number of observations incorrectly identified as negative IPOs by the model.

The second evaluation measure we use is the receiver operating characteristic (ROC) curves. ROC has been frequently used in many classification tasks in finance, such as IPO underpricing prediction (Bastı et al., 2015), bank merger prediction (Pasiouras et al., 2008; Pasiouras and Tanna, 2010) bankruptcy prediction (Veganzones and Severin, 2018; Mai et al., 2019), among others. The ROC curve plots the true-positive rate of the classifier on the vertical axis, and the false positive rate on the horizontal axis, by varying the classification threshold. Models closer to the upper and left corner of the diagram have a better out-of-sample performance. For comparison reasons, we plot a 45-degree line which indicates a random assignment of class labels. Based on ROC curves, we calculate the area under the curve (AUC). This measure ranges between 0 and 1. An uninformative classifier yields an AUC value of 0.5, while an AUC value of 1 indicates perfect classification.

5.2. Prediction using textual features and financial variables separately

In this section, we examine whether the language used by managers and investment bankers in the S-1 filing has any predictive power in distinguishing underpriced from overpriced IPOs. Hence, in the first six panels of Table 4, we report out-of sample accuracy scores of our prediction models, using only textual data as inputs. Each panel draws data from a separate section of the S-1 filing: entire S-1 filing (Panel A), the aggregate four major sections (Panel B), Risk Factors (Panel C), Summary (Panel D), Use of Proceeds (Panel E) and Management Discussion and Analysis (Panel F). Further, we use four different types of textual features: (1) term frequency (TF), (2) term frequency inverse document frequency (TF-IDF), (3) term frequency with bigrams (TF+bigrams), and (4) term frequency inverse document frequency with bigrams (TF-IDF+bigrams).[13]

In most cases, our models perform better than chance, which suggests that textual information is important in our classification task. Among the four sections (Panels C to F), MD&A yields the highest accuracy score (0.644), when the MLP model is used with TF-IDF as the textual input. As described before, MD&A provides a detailed explanation of a firm's operation, in a way that is comprehensible to the average investor. Hence, its high informative content allows investors to properly value the issuing firm. Furthermore, both Risk Factors and Summary sections produce adequate accuracy scores. More precisely, the highest score of the Risk Factors section equals 0.628 and is achieved when TF+bigrams are used as inputs in the logistic regression model. In the Summary section, the highest score is 0.625 and it is achieved with the logistic regression with TF+bigrams and the RF with TF features. The Use of Proceeds appears to be the least informative section, as its best score equals 0.592 when we use TF+bigrams as inputs in the random forest model.

One interesting insight from our analysis is the difference in accuracy scores between the

[13] Types 1 and 2 use only unigrams, and types 3 and 4 use a combination of unigrams and bigrams.

entire S-1 filing (Panel A) and the aggregate four sections (Panel B). In general, the four major sections produce slightly better results compared to the entire filing, as the combined section reflects unique information from each separate section and minimizes the noise introduced in the model. Notably, the four sections produce the highest accuracy score (0.649), when we use TF+bigrams in the logistic regression model. In terms of our models performance, RF, LOGIT, and MLP outperform SVMs either with linear kernel or with non-linear kernel (RBF).

Further, we also investigate whether financial variables alone can distinguish between underpriced and overpriced IPOs. In Panel G of Table 4, we use only financial variables as inputs in our prediction models. By looking at the accuracy scores, we observe that the MLP outperforms the other benchmarking models (0.675), while the RF has the second-best accuracy score (0.671). Overall, our results indicate that financial variables can effectively contribute to our classification task.

Insert **Table 4** here

5.3. Prediction with both textual and financial data

We now examine the classification ability of our models when we use both textual data and financial variables as inputs. The critical question that arises here is whether textual data include further incremental information beyond the financial variables, and if so, which models can achieve better prediction accuracy.

Table 5 presents the results of this analysis. At a first glance, it seems that the combination of textual data with financial variables worsens the predictive ability of our models, as the accuracy scores are lower compared to Table 4. We attribute this underperformance to the curse of dimensionality. In fact, one concern with the massive quantity of textual features is that they may overrule the role of financial variables and decrease the performance of our models. We address this issue in the next session.

Insert **Table 5** here

5.4. Prediction with singular value decomposition

To deal with the curse of dimensionality, we apply the singular value decomposition (SVD) dimensionality reduction technique. This is a very popular method in machine learning and NLP tasks, since it can project high-dimensional document vectors into a low dimensional space (Howland et al., 2003; Kim et al., 2005). The benefit of this approach is twofold: (1) it approximates the document very well by preserving the meaningful information, and (2) it copes with the curse of dimensionality (Mai et al., 2019). In our empirical specification, we use SVD to project the original feature vectors to 100 dimensions (SVD-100).[14] By using such a low level of textual representation, we expect that both textual data and financial variables can effectively increase the predictive power of our classification models.

Table 6 presents the accuracy scores of our models, when we use a combination of SVD-100 textual features and financial variables.[15] In line with our expectations, we find that the out-of-sample performance is substantially improved. More specifically, we observe that in many cases, the prediction accuracy exceeds 70%, which suggests that our models can now better capture the most important information from the original features. Again, the most informative section appears to be the combination of the 4 major sections (Panel B of Table 6). In fact, when we use the TF-IDF$_{SVD100}$ unigrams along with financial variables as inputs in the random forest model, we achieve an accuracy score of 0.736. This means that our model is able to correctly classify underpriced from overpriced IPOs in 73.60% of the future cases. In addition, consistent with our previous results, RF, LOGIT, and MLP outperform both types of SVMs.

[14] It is worth-mentioning that we take into account merely the 100 first SVD components since they were found to explain almost 80% of the joint variance of the 20,000 most frequent textual features in the S-1 filings.
[15] We also report results using only the SVD100 features as inputs in our classification models (see Table A1 in the Appendices).

Insert **Table 6** here

Figure 2 presents the ROC curves of our three best machine learning algorithms (RF, LOGIT, and MLP). We observe that AUC values are consistently above 0.7, with the TF-IDF features producing the highest scores (AUC values range from 0.75 to 0.79). When we compare the three models, we find that LOGIT and RF compete, since each one prevails across a specific spectrum of cut-off probabilities. Overall, our findings indicate that textual information in a decreased-dimension form can effectively supplement the financial variables and increase the prediction accuracy of our models.

Insert **Figure 2** here

5.5. Prediction with a combination of lexicon features and financial data

To get further insight of how textual information impacts our classification task, we also employ the document tone in our analysis, using the word lists of Loughran and McDonald (2011). In detail, we use these word lists to classify words into the six following categories: (1) negative, (2) positive, (3) uncertain, (4) weak modal, (5) strong modal, and (6) legal. Then, we compute sentiment scores based on how frequently a word appears in each one of these lexicons. In line with Loughran and McDonald (2013), we use each sentiment score separately in our analysis, due to the correlations and word overlap among word lists. Moreover, in each one of our models, we also include our 8 financial variables.

Table 7 presents the results of this analysis. Interestingly, we report results comparable with Table 6. In fact, RF and MLP models yield accuracy scores that in many cases exceed 70%. What is intriguing in our findings is that each lexicon produces higher results in a separate section of IPO prospectus. In the entire S-1 filling for instance, the best lexicons are the weak modal and the legal, which both produce an accuracy score of 0.690 with the random forest model. In the combined 4 sections, the positive lexicon yields the highest accuracy score (0.706) again with the use of RF. This finding is consistent with Brau et al.

(2016), who document that the S-1 section may reflect some kind of strategic optimism. In both Risk Factors and Use of Proceeds sections, the highest scores are achieved with the legal word list, a finding which could be related to the potential litigation risks associated with these sections. Turning to the MD&A section, strong modal results in the highest accuracy score (0.706) with the use of MLP. This finding is expected to some extent, as this section of the IPO prospectus reflects managers' confidence regarding the future prospects of their firm (Ferris et al., 2013). In the Summary section, the weak modal word list produces the best outcome. Considering that Summary is written by the underwriter in most cases, a weaker tone may be chosen to deal with potential litigation risks (Hanley and Hoberg, 2012). Finally, Table 8 reviews all the aforementioned findings.

Insert **Table 7** here

Insert **Table 8** here

5.6. Robustness tests

In this section, we conduct three main robustness tests to ensure the stability of our results. First, we exclude all crisis years from our sample. More precisely, we remove all IPOs from years 2000-2001 as the years of dot-com bubble, and from years 2008-2009 as the years of the financial crisis (Cohen et al., 2020).[16] Our results remain qualitative similar to the ones reported in Table 6 (see Table A2 in the Appendices).

As a second robustness test, we adopt two alternative sample splits for our training and testing datasets. First, we select 70% of our data as the training set, and the remaining 30% as the testing set (Veganzones and Severin, 2018; Kim et al., 2019). Second, we select 75% of our data as the training set, and the remaining 25% as the testing set (Gogas et al., 2018). Notably, our results are not influenced by these changes in the sample proportions (see Tables A3 and A4 in the Appendices).

[16] Now, our balanced sample consists of 503 overpriced and 503 underpriced IPOs.

As a final robustness check, we change the threshold of the most frequent words from 20,000 to 10,000. Our findings suggest that the main inferences of this article do not change when we use a different maximum number of textual features (see Table A5 in the Appendices).

6. Conclusions

Our study adds to the extensive literature of IPO underpricing. We utilize various machine learning classification models to distinguish between overpriced and underpriced IPOs. This allows us to explore if textual information from S-1 filings provides useful information in IPO underpricing prediction. The intuition behind this approach is based on the findings of Loughran and McDonald (2013), who document that textual information may proxy for the *ex-ante* uncertainty of the issuing firms' valuation. Our empirical approach goes beyond the scope of previous studies for two reasons. First, instead of using an econometric set up, we consider the task as a binary classification by using machine learning algorithms. Second, we examine the predictive power of unigram and bigram textual features, which do not require sentiment lexica and can, thus, be used more easily in less widely spoken languages where linguistic resources are typically more difficult to obtain.

For the purposes of our study, we collect a sample of 2,481 IPOs during the period 1997 to 2016. For each issuing firm, we retrieve the S-1 filing and we decompose it into its four major sections. Then, we construct a plethora of textual features from each one of these sections, the combination of all four sections, and the entire S-1 filing. We use the textual features along with several frequently-used financial variables as inputs in our machine learning models. We do so, in order to investigate whether our classification models can extract meaningful information from IPOs prospectuses, and if so, whether this information can effectively be combined with financial variables.

In our empirical analysis, we first examine the predictive power of textual information and financial variables separately. Our findings indicate that both types of inputs achieve an out-of-sample accuracy that in most cases exceeds 50%. However, when we combine the two sources of data, the accuracy scores are reduced, due to the high dimensionality of textual features. For this reason, we lower the dimensional space of our textual features by using the SVD dimension reduction technique. Notably, when we do so, we obtain superior results. In fact, when we combine SVD textual features with financial variables, we are able to achieve accuracy scores that in some cases exceed 70%. When it comes to our classification models, LOGIT, RF, and MLP yield the highest scores. Furthermore, AUC values complement the outperformance of these three models, as AUC values range between 0.72 and 0.79.

As a concluding remark, our research points out new methodological insights on how the plethora of textual features should efficiently be integrated with numerical data as mixed inputs into the machine learning algorithms. Apart from identifying the best classification model, our findings highlight the importance of textual information in IPOs underpricing prediction. We show that textual information from S-1 filings could reduce the *ex-ante* uncertainty of IPOs valuation, and thereby increase the predictive power of classification models. On this end, we hope that our study will provide fertile ground for future research, as there is still much to explore on this issue. For instance, future research may utilize other machine learning algorithms, including more complex neural models from recent NLP research, to predict IPOs underpricing.

References

Ambrose, B. W., & Megginson, W. L. (1992). The role of asset structure, ownership structure, and takeover defenses in determining acquisition likelihood. *Journal of Financial and Quantitative Analysis*, 27, 575-589.

Arnold, T., Fishe, R. P. H., & North, D. (2010). The effects of ambiguous information on initial and subsequent IPO returns. *Financial Management*, 39, 1497-1519.

Balakrishnan, R., Qiu X. Y., & Srinivasan P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202, 789-801.

Banerjee, S., Dai, L., & Shrestha, K. (2011). Cross-country IPOs: What explains differences in underpricing?. *Journal of Corporate Finance*, 17, 1289-1305.

Barnes, P. (1998). Can takeover targets be identified by statistical techniques? Some UK evidence. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 573-591.

Baron, D. P. (1982). A model of the demand for investment banking advising and distribution services for new issues. *Journal of Finance*, 37, 955-976.

Baron, D. P., & Holmstrom, B. (1980). The investment banking contract for new issues under asymmetric information: Delegation and the incentive problem. *Journal of Finance*, 35, 1115-1138.

Bastı, E., Kuzey, C., & Delen, D. (2015). Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems*, 73, 15-27.

Beatty, R. P., & Ritter, J. R. (1986). Investment banking, reputation, and the underpricing of initial public offerings. *Journal of Financial Economics*, 15, 213-232.

Beatty, R. P., & Welch, I. (1996). Issuer expenses and legal liability in initial public offerings. *Journal of Law and Economics*, 39, 545-602.

Benveniste, L. M., Ljungqvist, A., Wilhelm Jr., W. J., & Yu, X. (2003). Evidence of information spillovers in the production of investment banking services. *Journal of Finance*, 58, 577-608.

Bhabra, H. S., & Pettway, R. H. (2003). IPO prospectus information and subsequent performance. *Financial Review*, 38, 369-397.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50, 623-646.

Brau, J. C., Cicon, J., & McQueen, G. (2016). Soft strategic information and IPO underpricing. *Journal of Behavioral Finance*, 17, 1-17.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Brown, S. V, & Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, 49, 309-346.

Butler, A. W., Keefe, M. O. C., & Kieschnick, R. (2014). Robust determinants of IPO underpricing and their implications for IPO research. *Journal of Corporate Finance*, 27, 367-383.

Cao, L. (2003). Support vector machines experts for time series forecasting. *Neurocomputing*,

51, 321-339.

Carter, R., & Manaster, S. (1990). Initial public offerings and underwriter reputation. *Journal of Finance*, 45, 1045-1067.

Carter, R. B., Dark, F. H., & Singh, A. K. (1998). Underwriter reputation, initial returns, and the long-run performance of IPO stocks. *Journal of Finance*, 53, 285-311.

Chahine, S. (2008). Underpricing versus gross spread: New evidence on the effect of sold shares at the time of IPOs. *Journal of Multinational Financial Management*, 18, 180-196.

Chen, G., Firth, M., & Kim, J. B. (2004). IPO underpricing in China's new stock markets. *Journal of Multinational Financial Management*, 14, 283-302.

Chen, Y. S., Chen, J. S., & Cheng, C. H. (2010). An alternate method of examining IPO returns. *Intelligent Automation and Soft Computing*, 16, 151-161.

Cheng, C.H., Chen, Y.S., & Chen, J.S. (2007). Classifying initial returns of electronic firm's IPOs using entropy based rough sets in Taiwan trading systems. Second International Conference on Innovative Computing, Information and Control (ICICIC).

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75, 1371-1415.

Ding, R. (2016). Disclosure of downside risk and investors' use of qualitative information: Evidence from the IPO prospectus's risk factor section. *International Review of Finance,* 16, 73-126.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research*, 262, 347-360.

Doumpos, M., Kosmidou, K., & Pasiouras, F. (2004). Prediction of acquisition targets in the UK: A multicriteria approach. *Operational Research: An International Journal*, 4, 191-211.

Engelen, P. J., & Van Essen, M. (2010). Underpricing of IPOs: Firm-, issue-and country-specific characteristics. *Journal of Banking and Finance*, 34, 1958-1969.

Espahbodi, H., & Espahbodi, P. (2003). Binary choice models for corporate takeover. *Journal of Banking and Finance*, 27, 549-574.

Ferris, S. P., Hao, Q., & Liao, M. Y. (2013). The effect of issuer conservatism on IPO pricing and performance, *Review of Finance*, 17, 933-1027.

Gandhi, P., Loughran, T., & McDonald, B. (2019). Using annual report sentiment as a proxy for financial distress in U.S. banks. *Journal of Behavioral Finance*, 20, 424-436.

Gao, X., Ritter, J. R., & Zhu, Z. (2013). Where have all the IPOs gone? *Journal of Financial and Quantitative Analysis*, 48, 1663-1692.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, 241, 236-247.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting bank failures and stress testing: A machine learning approach. *International Journal of Forecasting*, 34, 440-455.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55, 685-697.

Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *Review of Financial Studies*, 23, 2821-2864.

Hanley, K. W., & Hoberg, G. (2012). Litigation risk, strategic disclosure and the underpricing of initial public offerings. *Journal of Financial Economics*, 103, 235-254.

Hasbrouck, J. (1985). The characteristics of takeover targets q and other measures. *Journal of Banking and Finance*, 9, 351-362.

Hauser, S., Yaari, U., Tanchuma, Y., & Baker, H. (2006). Initial public offering discount and competition. *Journal of Law and Economics*, 49, 331-351.

Howland, P., Jeon, M., & Park, H. (2003). Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25, 165-179.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32, 2513-2522.

Ibbotson, R. G. (1975). Price performance of common stock new issues. *Journal of Financial Economics*, 2, 235-272.

Ibbotson, R. G., Sindelar, J. L., & Ritter, J. R. (1994). The market's problems with the pricing of initial public offerings. *Journal of Applied Corporate Finance*, 7, 66-74.

Imbens, G. W. (1992). An efficient method of moments estimator for discrete choice models with choice-base sampling. *Econometrica*, 60, 1187-1214.

Jain, B. A., & Nag, B. N. (1995). Artificial neural network models for pricing initial public offerings. *Decision Sciences*, 26, 283-302.

Katsafados, A. G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G. N., & Pyrgiotakis, E. G. (2020). Using textual analysis to identify merger participants: Evidence from U.S. banking industry. *Working Paper*. Available at SSRN: https://ssrn.com/abstract=3474583.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185.

Kim, H., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6, 37-53.

Kim, J., Shin, S., Lee, H. S., & Oh, K. J. (2019). A machine learning portfolio allocation system for IPOs in Korean markets using GA-rough set theory. *Sustainability*, 11, 6803.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114, 128-147.

Kumar, R. B., Kumar, B. S., & Prasad, C. S. S. (2012). Financial news classification using SVM. *International Journal of Scientific and Research Publications*, 2, 1-6.

Laitinen, T., & Kankaanpaa, M. (1999). Comparative analysis of failure prediction methods: The Finnish case. *European Accounting Review*, 8, 67-92.

Leone, A. J., Rock, S., & Willenborg, M. (2007). Disclosure of intended use of proceeds and underpricing in initial public offerings. *Journal of Accounting Research*, 45, 111-153.

Ljungqvist, A. (2007). IPO Underpricing. In B.E. Eckbo, (Eds.), *Handbook of Corporate Finance: Empirical Corporate Finance* (pp. 375-422). Amsterdam: Elsevier-North Holland.

Ljungqvist, A., & Wilhelm Jr, W. J. (2003). IPO pricing in the dot-com bubble. *Journal of Finance*, 58, 723-752.

Ljungqvist, A., & Wilhelm Jr, W. J. (2005). Does prospect theory explain IPO market behavior? *Journal of Finance*, 60, 1759-1790.

Logue, D. E. (1973). On the pricing of unseasoned equity issues: 1965-1969. *Journal of Financial and Quantitative Analysis*, 25, 133-141.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66, 35-65.

Loughran, T., & McDonald, B. (2013). IPO First-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, 109, 307-326.

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

Loughran, T., & Ritter, J. (2004). Why has IPO underpricing changed over time? *Financial Management*, 33, 5-37.

Ly, T. H., & Nguyen, K. (2020). Do words matter: Predicting IPO performance from prospectus sentiment. 2020 IEEE 14th International Conference on Semantic Computing.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274, 743-758.

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

Mitsdorffer, R., Diederich, J., & Tan, C. (2002). Rule extraction from the technology IPOs in the US stock market. Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02) (IEEE Press, Piscataway), 2328-2334.

Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys,* 30, 356-369.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ling Ngo, D. C. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41, 7653-7670.

Neophytou, E., & Mar Molinero, C. (2004). Predicting corporate failure in the UK: A multidimensional scaling approach. *Journal of Business Finance and Accounting*, 31, 677-710.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42, 9603-9611.

Pai, P. F., & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, 33, 497-505.

Palepu, K. G. (1986). Predicting takeover targets: A methodological and empirical analysis. *Journal of Accounting and Economics*, 8, 3-35.

Pasiouras, F., & Tanna, S. (2010). The prediction of bank acquisition targets with discriminant and logit analyses: Methodological issues and empirical evidence. *Research in International Business and Finance*, 24, 39-61.

Pasiouras, F., Gaganis, C., & Zopounidis, C. (2010). Multicriteria classification models for the identification of targets and acquirers in the Asian banking sector. *European Journal of Operational Research*, 204, 328-335.

Pasiouras, F., Gaganis, C., Tanna, S., & Zopounidis, C. (2008). An application of support vector machines in the prediction of acquisition targets: Evidence from the EU banking sector. In C. Zopounidis, M. Doumpos, & P. Pardalos (Eds.), Handbook of financial engineering (pp. 431-456). Boston: Springer.

Pasiouras, F., Tanna, S., & Zopounidis, C. (2007). The identification of acquisition targets in the EU banking industry: An application of multicriteria approaches. *International Review of Financial Analysis*, 16, 262-281.

Pestov, V. (2013). Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65, 1427-1437.

Powell, R. G. (2001). Takeover prediction and portfolio performance: A note. *Journal of Business Finance and Accounting*, 28, 993-1011.

Quintana, D., Chávez, F., Luque Baena, R. M., & Luna, F. (2018). Fuzzy techniques for IPO underpricing prediction. *Journal of Intelligent & Fuzzy Systems*, 35, 367-381.

Quintana, D., Sáez, Y., & Isasi, P. (2017). Random forest prediction of IPO underpricing. *Applied Sciences*, 7, 636.

Reber, B., Berry, B., & Toms, S. (2005). Predicting mispricing of initial public offerings. *Intelligent Systems in Accounting, Finance and Management*, 13, 41-59.

Ritter, J. R. (1984). The "hot issue" market of 1980. *Journal of Business*, 57, 215-240.

Ritter, J. R., & Welch, I. (2002). A review of IPO activity, pricing, and allocations. *Journal of Finance*, 57, 1795-1828.

Rock, K. (1986). Why new issues are underpriced. *Journal of Financial Economics*, 15, 187-212.

Routledge, B. R., Sacchetto, S., & Smith, N. A. (2017). Predicting merger targets and acquirers from text. *Working Paper*, Carnegie Mellon University.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions of Information Systems*, 27, 1-19.

Shin, K. S., Lee, T. S., & Kim, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28, 127-135.

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36, 10-25.

Vapnik, V. N. (1998). Statistical learning theory. (1st ed.). New York: Wiley.

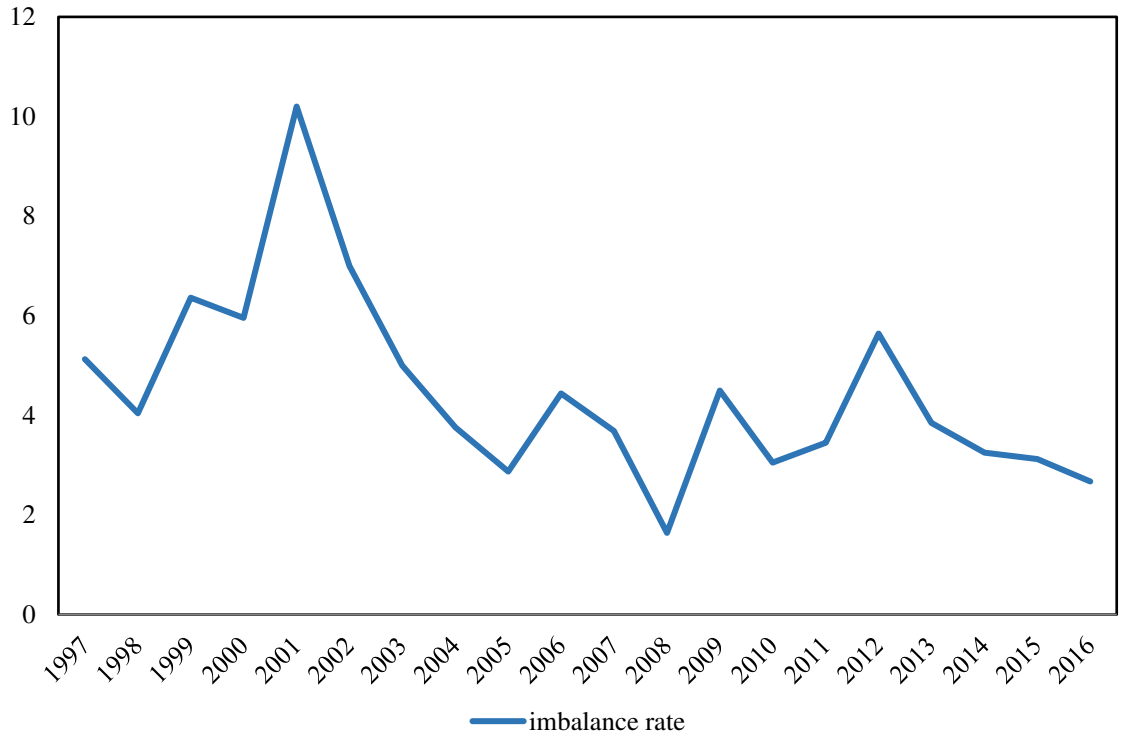
Veganzones, D., & Severin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, 112, 111-124.

Wang, D., Qian, X., Quek, C., Tan, A. H., Miao, C., Zhang, X., Ng, G. S., & Zhou, Y. (2018). An interpretable neural fuzzy inference system for predictions of underpricing in initial public offerings. *Neurocomputing*, 319, 102-117.

Wu, C. H., Tzeng, G. H., Goo, Y. J., & Fang, W. C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications*, 32, 397-408.

Figure 1

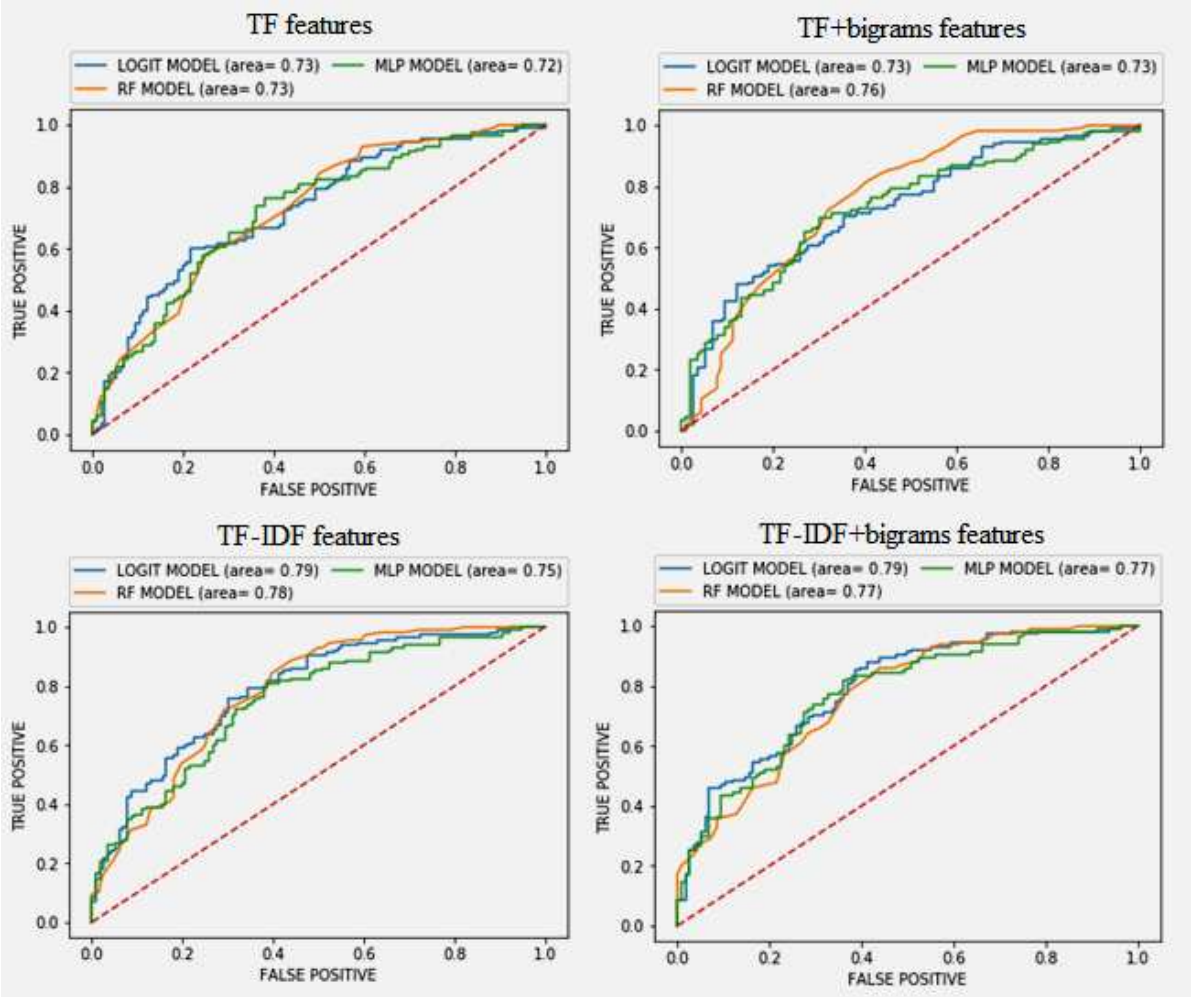
Imbalance rate of the final IPO sample



This figure represents the imbalanced rate of our final (imbalanced) sample on an annual basis. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. In each year of our examination period, the imbalance rate is the ratio of underpriced IPOs to overpriced IPOs.

Figure 2

ROC curve with both SVD-100 textual data and financial variables



This figure depicts the receiver operating characteristic (ROC) curves for three machine learning algorithms: (1) logistic regression (LOGIT MODEL), (2) random forest (RF MODEL), and (3) multilayer preceptor (MLP MODEL). The red dotted line represents a 45-degree line which indicates a random assignment of class labels. Area stands for the area under curve (AUC) measure. TF and TF-IDF represent the two term weighting schemes. TF stands for the term frequency scheme, and TF-IDF for the term frequency-inverse document frequency scheme. Bigrams are word pairs represented as a single textual feature.

Table 1
Studies on IPO underpricing classification

Authors	Country	IPO deals	Period	Models
Mitsdorffer et al. (2002)	USA	182	1996-2000	ANN, Bayesian classifier, Decision Tree, Rule learners, SVM
Cheng et al. (2007)	Taiwan	220	1985-2003	ANN, Naive Bayes, Rough Sets
Chen et al. (2010)	Taiwan	220	1985-2003	ANN, Bayes Net, Decision Tree, Rough Sets
Kim et al. (2019)	South Korea	718	2007-2018	Genetic algorithms-rough sets
Ly and Nguyen (2020)	USA	N/A	1993-2019	Decision Tree, Logit, Naive Bayes, Random Forest

This table summarizes the relevant literature that uses machine learning algorithms in an IPO classification task. ANN stands for artificial neural networks and SVM for support vector machine. The number of IPO deals is not available (N/A) in the study of Ly and Nguyen (2020).

Table 2

Topic extraction through LDA method for each separate section

LDA	Topic 1	Topic 2	Topic 3	Topic 4
Risk Factors	Future products	Regulatory approval	Financial operations	
Summary	Common shares offering	Financial data information		
Use of Proceeds	Acquisition	Development	Dividends	Credit
MD&A	Stock value development	Revenue increase value	Net sales	

This table summarizes the results of applying Latent Dirichlet Allocation (LDA) to the sections of the S-1 filing. LDA is conducted in each separate section of the S-1 filing (Risk Factors, Summary, Use of Proceeds, and MD&A). LDA identifies the most relevant topics (up to 4 in our case) of each separate section, and represents each topic by its most frequent words. Here we show our own labels of the topics, which reflect our own understanding of the meaning of each topic, based on the frequent words of each topic.

Table 3

Summary statistics

Variables	Mean	Median	Std. Dev.	Min	Max
First-day returns	31.1%	12.5%	59.7%	-33.1%	697.5%
Sales	$451.9	$39.6	$3,149	$0	$104,589
Positive EPS	0.39	0	0.49	0	1
Share overhang	3.63	3.03	2.73	-0.16	68.18
Venture capital	0.53	0.50	1	0	1
Prior Nasdaq 15-day returns	1.12%	1.24%	5.38%	-0.22%	0.25%
Up-revision	10.5%	0%	20.1%	0%	220%
Top-tier	0.78	1	0.42	0	1
Days between S-1 and 1-trading day	115.9	86	111.4	21	1,659
% Negative	1.36	1.38	0.30	0.50	2.27
% Positive	0.75	0.75	0.15	0.37	2.08
% Uncertain	1.35	1.36	0.20	0.76	2.04
% Weak modal	0.70	0.71	0.18	0.26	1.36
% Strong modal	0.51	0.50	0.11	0.21	1.23
% Legal	0.75	0.72	0.20	0.34	2.29

This table reports the summary statistics of our final (imbalanced) sample. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. *First-day returns* is the underpricing measure, and is calculated as the percentage change from the offer price to the closing price. *Sales* is the logarithm of firm annual sales in the 12 months prior to the IPO. *Positive EPS* is a dummy variable which equals to 1 if the IPO has positive earnings per share in the year before going public, and 0 otherwise. *Share overhang* is the amount of shares retained divided by the amount of shares in the IPO. *Venture capital* is a dummy variable which equals to 1 if the IPO is backed by venture capital, and 0 otherwise. *Prior Nasdaq 15-day returns* is the buy and hold returns of the Nasdaq index 15 trading days before IPO date. *Up-revision* is the percentage upward revision from the mid-point of the filing range, if the offer price is higher than mid-point, and 0 otherwise. *Top-tier* is a dummy variable which equals to 1 if at least one underwriter has been classified as top-tier according to the rankings of Carter and Manaster (1990), Carter et al. (1998) and Loughran and Ritter (2004), and 0 otherwise. *Days between S-1 and 1-trading day* is the logarithm of days between S-1 filing date and first-trading day. Finally, *% negative*, *% positive*, *% uncertain*, *% weak modal*, *% strong modal*, and *% legal* are the sentiment scores of the S-1 filing, calculated using Loughran and McDonald's (2011) word lists.

Table 4

Out-of-sample performance using textual features and financial variables separately

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF	0.505	0.500	0.565	0.580	0.600
TF-IDF	0.505	0.465	0.600	0.600	0.585
TF + bigrams	0.460	0.500	0.565	0.600	0.585
TF-IDF + bigrams	0.460	0.460	0.620	0.600	0.595
Panel B: 4 Sections					
TF	0.511	0.498	0.641	0.580	0.615
TF-IDF	0.532	0.524	0.606	0.580	0.610
TF + bigrams	0.567	0.494	0.649	0.602	0.641
TF-IDF + bigrams	0.498	0.541	0.619	0.605	0.623
Panel C: Risk Factors					
TF	0.502	0.502	0.615	0.576	0.602
TF-IDF	0.500	0.532	0.589	0.597	0.597
TF + bigrams	0.500	0.502	0.628	0.567	0.610
TF-IDF + bigrams	0.472	0.502	0.589	0.602	0.597
Panel D: Summary					
TF	0.517	0.522	0.616	0.625	0.612
TF-IDF	0.547	0.522	0.612	0.616	0.612
TF + bigrams	0.535	0.517	0.625	0.586	0.603
TF-IDF + bigrams	0.547	0.517	0.608	0.621	0.616
Panel E: Use of Proceeds					
TF	0.519	0.494	0.571	0.575	0.562
TF-IDF	0.490	0.524	0.571	0.575	0.549
TF+ bigrams	0.506	0.502	0.549	0.592	0.579
TF-IDF + bigrams	0.472	0.455	0.571	0.571	0.588
Panel F: MD&A					
TF	0.588	0.545	0.584	0.609	0.627
TF-IDF	0.566	0.528	0.635	0.588	0.644
TF + bigrams	0.575	0.532	0.588	0.631	0.627
TF-IDF + bigrams	0.515	0.541	0.618	0.601	0.635
Panel G: No textual data					
Economic variables	0.545	0.608	0.641	0.671	0.675

This table reports the accuracy scores for our machine learning models, using textual information and financial variables separately as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F use only textual information as inputs. Panel A reports results when we use textual information from the entire S-1 filing, Panel B from the combination of the 4 major sections, and Panels C to F from each separate section (Risk Factors, Summary, Use of Proceeds, and MD&A). Panel G reports results when we use only financial variables. The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table 5

Out-of-sample performance using both textual features and financial variables as inputs

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF	0.515	0.500	0.560	0.595	0.590
TF-IDF	0.460	0.510	0.560	0.590	0.600
TF + bigrams	0.505	0.505	0.560	0.600	0.590
TF-IDF + bigrams	0.560	0.560	0.595	0.595	0.590
Panel B: 4 Sections					
TF	0.545	0.502	0.623	0.606	0.593
TF-IDF	0.580	0.502	0.593	0.610	0.589
TF + bigrams	0.506	0.502	0.623	0.615	0.641
TF-IDF + bigrams	0.602	0.511	0.602	0.597	0.590
Panel C: Risk Factors					
TF	0.506	0.511	0.567	0.632	0.563
TF-IDF	0.498	0.472	0.593	0.597	0.615
TF + bigrams	0.506	0.502	0.575	0.545	0.584
TF-IDF + bigrams	0.528	0.494	0.567	0.597	0.571
Panel D: Summary					
TF	0.487	0.500	0.552	0.629	0.595
TF-IDF	0.522	0.573	0.595	0.642	0.625
TF + bigrams	0.530	0.504	0.491	0.612	0.582
TF-IDF + bigrams	0.530	0.513	0.595	0.638	0.569
Panel E: Use of Proceeds					
TF	0.541	0.489	0.562	0.627	0.558
TF-IDF	0.519	0.532	0.524	0.627	0.545
TF + bigrams	0.506	0.528	0.562	0.627	0.562
TF-IDF + bigrams	0.506	0.481	0.554	0.609	0.541
Panel F: MD&A					
TF	0.554	0.528	0.605	0.597	0.597
TF-IDF	0.545	0.519	0.614	0.639	0.592
TF + bigrams	0.515	0.502	0.597	0.627	0.609
TF-IDF + bigrams	0.515	0.489	0.575	0.631	0.627

This table reports the accuracy scores for our machine learning models, using both textual information and financial variables as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

.

Table 6

Out-of-sample performance using both SVD-100 textual features and financial variables as inputs

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF_{SVD100}	0.525	0.530	0.630	0.700	0.655
$TF\text{-}IDF_{SVD100}$	0.580	0.575	0.665	0.725	0.670
$(TF + bigrams)_{SVD100}$	0.540	0.610	0.635	0.680	0.645
$(TF\text{-}IDF + bigrams)_{SVD100}$	0.630	0.610	0.670	0.720	0.700
Panel B: 4 Sections					
TF_{SVD100}	0.528	0.558	0.628	0.700	0.684
$TF\text{-}IDF_{SVD100}$	0.550	0.584	0.688	0.736	0.688
$(TF + bigrams)_{SVD100}$	0.571	0.545	0.658	0.701	0.667
$(TF\text{-}IDF + bigrams)_{SVD100}$	0.519	0.589	0.684	0.727	0.714
Panel C: Risk Factors					
TF_{SVD100}	0.506	0.506	0.658	0.680	0.654
$TF\text{-}IDF_{SVD100}$	0.580	0.589	0.632	0.658	0.653
$(TF + bigrams)_{SVD100}$	0.519	0.515	0.645	0.641	0.623
$(TF\text{-}IDF + bigrams)_{SVD100}$	0.580	0.558	0.645	0.684	0.628
Panel D: Summary					
TF_{SVD100}	0.664	0.565	0.711	0.694	0.694
$TF\text{-}IDF_{SVD100}$	0.655	0.582	0.655	0.698	0.707
$(TF + bigrams)_{SVD100}$	0.552	0.530	0.711	0.703	0.703
$(TF\text{-}IDF + bigrams)_{SVD100}$	0.599	0.586	0.681	0.703	0.668
Panel E: Use of Proceeds					
TF_{SVD100}	0.481	0.528	0.661	0.661	0.631
$TF\text{-}IDF_{SVD100}$	0.554	0.558	0.687	0.682	0.635
$(TF + bigrams)_{SVD100}$	0.562	0.549	0.644	0.691	0.652
$(TF\text{-}IDF + bigrams)_{SVD100}$	0.519	0.545	0.652	0.661	0.648
Panel F: MD&A					
TF_{SVD100}	0.519	0.575	0.644	0.652	0.661
$TF\text{-}IDF_{SVD100}$	0.562	0.554	0.682	0.700	0.674
$(TF + bigrams)_{SVD100}$	0.558	0.605	0.627	0.652	0.678
$(TF\text{-}IDF + bigrams)_{SVD100}$	0.545	0.519	0.674	0.717	0.704

This table reports the accuracy scores for our machine learning models, using both textual information and financial variables as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. The dimensions of textual features are further reduced to 100 using the singular value decomposition (SVD) dimensionality reduction technique. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table 7

Out-of-sample performance using both text sentiment proportions and financial variables as inputs

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
Negative	0.595	0.585	0.605	0.665	0.655
Positive	0.535	0.560	0.625	0.680	0.665
Uncertain	0.645	0.640	0.615	0.665	0.665
Weak Modal	0.485	0.590	0.620	0.690	0.645
Strong Modal	0.520	0.620	0.625	0.645	0.680
Legal	0.585	0.570	0.620	0.690	0.675
Panel B: 4 Sections					
Negative	0.574	0.638	0.660	0.698	0.689
Positive	0.638	0.583	0.660	0.706	0.685
Uncertain	0.626	0.630	0.655	0.702	0.681
Weak Modal	0.660	0.613	0.660	0.698	0.685
Strong Modal	0.583	0.664	0.655	0.685	0.689
Legal	0.596	0.638	0.651	0.689	0.689
Panel C: Risk Factors					
Negative	0.638	0.587	0.643	0.672	0.672
Positive	0.609	0.609	0.660	0.677	0.677
Uncertain	0.570	0.630	0.660	0.685	0.668
Weak Modal	0.570	0.626	0.647	0.681	0.681
Strong Modal	0.519	0.630	0.655	0.689	0.702
Legal	0.532	0.549	0.643	0.685	0.706
Panel D: Summary					
Negative	0.570	0.515	0.651	0.677	0.681
Positive	0.523	0.532	0.660	0.677	0.681
Uncertain	0.596	0.587	0.651	0.681	0.677
Weak Modal	0.655	0.562	0.664	0.698	0.723
Strong Modal	0.549	0.604	0.660	0.685	0.672
Legal	0.502	0.570	0.664	0.672	0.668
Panel E: Use of Proceeds					
Negative	0.519	0.655	0.660	0.689	0.664
Positive	0.515	0.626	0.660	0.694	0.698
Uncertain	0.549	0.553	0.647	0.685	0.677
Weak Modal	0.609	0.587	0.655	0.664	0.702
Strong Modal	0.651	0.626	0.664	0.694	0.689
Legal	0.583	0.651	0.660	0.716	0.689
Panel F: MD&A					
Negative	0.506	0.647	0.655	0.694	0.677
Positive	0.451	0.562	0.660	0.685	0.689
Uncertain	0.536	0.600	0.660	0.689	0.698
Weak Modal	0.532	0.583	0.647	0.694	0.694
Strong Modal	0.468	0.600	0.664	0.681	0.706
Legal	0.562	0.591	0.655	0.685	0.681

This table reports the accuracy scores for our machine learning models, using both lexicon-based sentiment features and financial variables as inputs. We use Loughran and McDonald's (2011) word lists to classify words into negative, positive, uncertain, weak modal, strong modal, and legal categories. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP).

Table 8

Overall conclusions of Table 7

	S-1	4 Sections	Risk Factors	Summary	Use of Proceeds	MD&A
Best lexicon	weak modal and legal	positive	legal	weak modal	legal	strong modal
Best model	RF	RF	MLP	MLP	RF	MLP
Best score	0.690	0.706	0.706	0.723	0.716	0.706

This table summarizes the conclusions drawn from Table 7. Best lexicon represents the word lists of Loughran and McDonald (2011) that produces the highest accuracy score in each section of the IPO prospectus (entire S-1 filing, the 4 major sections combined, Risk Factors, Summary, Use of Proceeds, and MD&A). RF stands for the random forest model and MLP for the multilayer perceptron model. Best score represents the accuracy score achieved with the combination of best lexicon with the best model.

Appendices

Table A1

Out-of-sample performance using only SVD-100 textual features

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF_{SVD100}	0.530	0.500	0.565	0.540	0.560
$TF\text{-}IDF_{SVD100}$	0.510	0.500	0.605	0.605	0.605
(TF + bigrams)$_{SVD100}$	0.505	0.500	0.570	0.570	0.580
(TF-IDF + bigrams)$_{SVD100}$	0.515	0.545	0.590	0.595	0.600
Panel B: 4 Sections					
TF_{SVD100}	0.494	0.515	0.580	0.628	0.580
$TF\text{-}IDF_{SVD100}$	0.511	0.515	0.602	0.619	0.610
(TF + bigrams)$_{SVD100}$	0.554	0.494	0.589	0.580	0.593
(TF-IDF + bigrams)$_{SVD100}$	0.563	0.619	0.602	0.645	0.602
Panel C: Risk Factors					
TF_{SVD100}	0.524	0.515	0.589	0.545	0.545
$TF\text{-}IDF_{SVD100}$	0.571	0.520	0.597	0.589	0.571
(TF + bigrams)$_{SVD100}$	0.515	0.532	0.602	0.541	0.602
(TF-IDF + bigrams)$_{SVD100}$	0.494	0.524	0.589	0.567	0.597
Panel D: Summary					
TF_{SVD100}	0.578	0.500	0.649	0.625	0.625
$TF\text{-}IDF_{SVD100}$	0.539	0.535	0.599	0.552	0.603
(TF + bigrams)$_{SVD100}$	0.547	0.500	0.616	0.586	0.642
(TF-IDF + bigrams)$_{SVD100}$	0.522	0.539	0.578	0.556	0.599
Panel E: Use of Proceeds					
TF_{SVD100}	0.500	0.506	0.562	0.562	0.554
$TF\text{-}IDF_{SVD100}$	0.485	0.485	0.558	0.541	0.558
(TF + bigrams)$_{SVD100}$	0.500	0.528	0.515	0.519	0.523
(TF-IDF + bigrams)$_{SVD100}$	0.554	0.567	0.545	0.532	0.562
Panel F: MD&A					
TF_{SVD100}	0.450	0.541	0.618	0.588	0.601
$TF\text{-}IDF_{SVD100}$	0.588	0.588	0.614	0.588	0.656
(TF + bigrams)$_{SVD100}$	0.524	0.537	0.618	0.592	0.622
(TF-IDF + bigrams)$_{SVD100}$	0.519	0.537	0.614	0.592	0.631

This table reports the accuracy scores for our machine learning models, using only textual information as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. The dimensions of textual features are further reduced to 100 using the singular value decomposition (SVD) dimensionality reduction technique. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table A2

Out-of-sample performance using both SVD-100 textual features and financial variables as inputs after excluding crises years

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF$_{SVD100}$	0.585	0.550	0.655	0.685	0.620
TF-IDF$_{SVD100}$	0.591	0.614	0.673	0.696	0.673
(TF + bigrams)$_{SVD100}$	0.573	0.585	0.614	0.684	0.632
(TF-IDF + bigrams)$_{SVD100}$	0.596	0.649	0.673	0.712	0.649
Panel B: 4 Sections					
TF$_{SVD100}$	0.607	0.532	0.672	0.685	0.677
TF-IDF$_{SVD100}$	0.652	0.602	0.716	0.705	0.706
(TF + bigrams)$_{SVD100}$	0.617	0.572	0.687	0.700	0.677
(TF-IDF + bigrams)$_{SVD100}$	0.637	0.632	0.687	0.712	0.682
Panel C: Risk Factors					
TF$_{SVD100}$	0.537	0.498	0.667	0.702	0.667
TF-IDF$_{SVD100}$	0.567	0.602	0.647	0.687	0.617
(TF + bigrams)$_{SVD100}$	0.493	0.577	0.662	0.617	0.662
(TF-IDF + bigrams)$_{SVD100}$	0.587	0.602	0.657	0.702	0.627
Panel D: Summary					
TF$_{SVD100}$	0.576	0.562	0.645	0.670	0.695
TF-IDF$_{SVD100}$	0.606	0.591	0.680	0.680	0.690
(TF + bigrams)$_{SVD100}$	0.552	0.562	0.655	0.714	0.700
(TF-IDF + bigrams)$_{SVD100}$	0.665	0.611	0.700	0.700	0.665
Panel E: Use of Proceeds					
TF$_{SVD100}$	0.512	0.527	0.685	0.690	0.631
TF-IDF$_{SVD100}$	0.507	0.562	0.685	0.704	0.621
(TF + bigrams)$_{SVD100}$	0.547	0.502	0.616	0.680	0.616
(TF-IDF + bigrams)$_{SVD100}$	0.581	0.557	0.645	0.670	0.611
Panel F: MD&A					
TF$_{SVD100}$	0.517	0.611	0.616	0.665	0.631
TF-IDF$_{SVD100}$	0.621	0.522	0.631	0.704	0.641
(TF + bigrams)$_{SVD100}$	0.567	0.552	0.601	0.700	0.675
(TF-IDF + bigrams)$_{SVD100}$	0.527	0.586	0.655	0.709	0.655

This table reports the accuracy scores for our machine learning models, using both textual information and financial variables as inputs. The final (imbalanced) sample consists of 2,085 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 503 underpriced and 503 overpriced IPOs, after we remove all IPOs from years 2000-2001 as the years of dot-com bubble, and from years 2008-2009 as the years of the financial crisis. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. The dimensions of textual features are further reduced to 100 using the singular value decomposition (SVD) dimensionality reduction technique. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table A3

Out-of-sample performance using both SVD-100 textual features and financial variables as inputs with as 70/30 train-test split

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF$_{SVD100}$	0.542	0.510	0.631	0.683	0.651
TF-IDF$_{SVD100}$	0.546	0.590	0.647	0.711	0.659
(TF + bigrams)$_{SVD100}$	0.550	0.530	0.639	0.667	0.655
(TF-IDF + bigrams)$_{SVD100}$	0.635	0.558	0.655	0.711	0.659
Panel B: 4 Sections					
TF$_{SVD100}$	0.635	0.549	0.674	0.691	0.674
TF-IDF$_{SVD100}$	0.594	0.601	0.670	0.729	0.670
(TF + bigrams)$_{SVD100}$	0.563	0.556	0.670	0.684	0.660
(TF-IDF + bigrams)$_{SVD100}$	0.629	0.563	0.663	0.712	0.674
Panel C: Risk Factors					
TF$_{SVD100}$	0.516	0.502	0.630	0.682	0.641
TF-IDF$_{SVD100}$	0.633	0.557	0.651	0.706	0.667
(TF + bigrams)$_{SVD100}$	0.533	0.498	0.637	0.671	0.614
(TF-IDF + bigrams)$_{SVD100}$	0.561	0.574	0.661	0.678	0.647
Panel D: Summary					
TF$_{SVD100}$	0.524	0.531	0.703	0.703	0.676
TF-IDF$_{SVD100}$	0.610	0.579	0.655	0.707	0.662
(TF + bigrams)$_{SVD100}$	0.579	0.535	0.707	0.693	0.669
(TF-IDF + bigrams)$_{SVD100}$	0.593	0.559	0.662	0.745	0.645
Panel E: Use of Proceeds					
TF$_{SVD100}$	0.512	0.557	0.663	0.694	0.660
TF-IDF$_{SVD100}$	0.577	0.605	0.653	0.691	0.615
(TF + bigrams)$_{SVD100}$	0.526	0.553	0.650	0.715	0.640
(TF-IDF + bigrams)$_{SVD100}$	0.605	0.605	0.646	0.667	0.656
Panel F: MD&A					
TF$_{SVD100}$	0.643	0.570	0.653	0.650	0.691
TF-IDF$_{SVD100}$	0.581	0.646	0.670	0.700	0.680
(TF + bigrams)$_{SVD100}$	0.612	0.577	0.632	0.663	0.656
(TF-IDF + bigrams)$_{SVD100}$	0.567	0.605	0.684	0.691	0.663

This table reports the accuracy scores for our machine learning models, using both textual information and financial variables as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. The dimensions of textual features are further reduced to 100 using the singular value decomposition (SVD) dimensionality reduction technique. We use 70% of our sample as the training set and the remaining 30% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table A4

Out-of-sample performance using both SVD-100 textual features and financial variables as inputs with a 75/25 train-test split

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF$_{SVD100}$	0.528	0.542	0.609	0.679	0.639
TF-IDF$_{SVD100}$	0.582	0.599	0.662	0.712	0.662
(TF + bigrams)$_{SVD100}$	0.582	0.532	0.625	0.679	0.642
(TF-IDF + bigrams)$_{SVD100}$	0.552	0.649	0.666	0.709	0.649
Panel B: 4 Sections					
TF$_{SVD100}$	0.572	0.549	0.671	0.705	0.679
TF-IDF$_{SVD100}$	0.610	0.627	0.679	0.720	0.676
(TF + bigrams)$_{SVD100}$	0.543	0.546	0.668	0.691	0.662
(TF-IDF + bigrams)$_{SVD100}$	0.584	0.569	0.691	0.714	0.694
Panel C: Risk Factors					
TF$_{SVD100}$	0.568	0.513	0.657	0.666	0.643
TF-IDF$_{SVD100}$	0.550	0.571	0.637	0.692	0.652
(TF + bigrams)$_{SVD100}$	0.527	0.516	0.660	0.689	0.650
(TF-IDF + bigrams)$_{SVD100}$	0.579	0.582	0.634	0.709	0.657
Panel D: Summary					
TF$_{SVD100}$	0.546	0.552	0.690	0.687	0.678
TF-IDF$_{SVD100}$	0.592	0.560	0.649	0.681	0.649
(TF + bigrams)$_{SVD100}$	0.546	0.517	0.681	0.664	0.670
(TF-IDF + bigrams)$_{SVD100}$	0.609	0.555	0.670	0.690	0.684
Panel E: Use of Proceeds					
TF$_{SVD100}$	0.573	0.576	0.630	0.688	0.636
TF-IDF$_{SVD100}$	0.605	0.607	0.642	0.702	0.641
(TF + bigrams)$_{SVD100}$	0.530	0.579	0.639	0.682	0.639
(TF-IDF + bigrams)$_{SVD100}$	0.564	0.616	0.645	0.711	0.638
Panel F: MD&A					
TF$_{SVD100}$	0.547	0.570	0.662	0.702	0.667
TF-IDF$_{SVD100}$	0.570	0.639	0.671	0.745	0.662
(TF + bigrams)$_{SVD100}$	0.533	0.607	0.639	0.696	0.662
(TF-IDF + bigrams)$_{SVD100}$	0.613	0.610	0.665	0.708	0.676

This table reports the accuracy scores for our machine learning models, using both textual information and financial variables as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 20,000 most frequent words of the S-1 filing. The dimensions of textual features are further reduced to 100 using the singular value decomposition (SVD) dimensionality reduction technique. We use 75% of our sample as the training set and the remaining 25% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table A5

Out-of-sample performance of the 10,000 most frequent textual features using both SVD-100 textual features and financial variables as inputs

	SVM-linear	SVM-RBF	LOGIT	RF	MLP
Panel A: S-1					
TF$_{SVD100}$	0.485	0.545	0.630	0.660	0.655
TF-IDF$_{SVD100}$	0.585	0.665	0.660	0.725	0.680
(TF + bigrams)$_{SVD100}$	0.510	0.600	0.630	0.690	0.665
(TF-IDF + bigrams)$_{SVD100}$	0.560	0.600	0.660	0.715	0.680
Panel B: 4 Sections					
TF$_{SVD100}$	0.550	0.545	0.628	0.705	0.649
TF-IDF$_{SVD100}$	0.641	0.615	0.693	0.718	0.684
(TF + bigrams)$_{SVD100}$	0.554	0.554	0.641	0.701	0.680
(TF-IDF + bigrams)$_{SVD100}$	0.606	0.532	0.688	0.718	0.675
Panel C: Risk Factors					
TF$_{SVD100}$	0.537	0.502	0.658	0.684	0.680
TF-IDF$_{SVD100}$	0.584	0.606	0.628	0.710	0.649
(TF + bigrams)$_{SVD100}$	0.563	0.515	0.662	0.680	0.641
(TF-IDF + bigrams)$_{SVD100}$	0.563	0.567	0.645	0.701	0.649
Panel D: Summary					
TF$_{SVD100}$	0.565	0.560	0.711	0.711	0.690
TF-IDF$_{SVD100}$	0.603	0.591	0.647	0.711	0.707
(TF + bigrams)$_{SVD100}$	0.530	0.513	0.716	0.716	0.728
(TF-IDF + bigrams)$_{SVD100}$	0.582	0.634	0.694	0.703	0.694
Panel E: Use of Proceeds					
TF$_{SVD100}$	0.481	0.528	0.665	0.695	0.635
TF-IDF$_{SVD100}$	0.554	0.558	0.687	0.682	0.631
(TF + bigrams)$_{SVD100}$	0.537	0.519	0.657	0.682	0.614
(TF-IDF + bigrams)$_{SVD100}$	0.562	0.554	0.644	0.678	0.635
Panel F: MD&A					
TF$_{SVD100}$	0.567	0.545	0.644	0.682	0.644
TF-IDF$_{SVD100}$	0.536	0.592	0.674	0.716	0.704
(TF + bigrams)$_{SVD100}$	0.560	0.575	0.631	0.682	0.670
(TF-IDF + bigrams)$_{SVD100}$	0.631	0.562	0.652	0.708	0.712

This table reports the accuracy scores for our machine learning models, using both textual information and financial variables as inputs. The final (imbalanced) sample consists of 2,481 IPOs from 1997 to 2016. The analysis is based on a balanced sample of 576 underpriced and 576 overpriced IPOs. To construct the textual features, we use the 10,000 most frequent words of the S-1 filing. The dimensions of textual features are further reduced to 100 using the singular value decomposition (SVD) dimensionality reduction technique. We use 80% of our sample as the training set and the remaining 20% as the testing set. Panels A to F report results for the entire S-1 filing, the combination of the 4 major sections, and from each separate section, respectively (Risk Factors, Summary, Use of Proceeds, and MD&A). The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams. Bigrams are pairs of consecutive words represented as a single textual feature. We use the following machine learning models: linear support vector machines (SVM-linear), support vector machines with radial basis function kernel (SVM-RBF), logistic regression (LOGIT), random forest (RF), and multilayer perceptron (MLP). TF and TF-IDF are the two term weighting schemes. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.