LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Prospectus Content as Predictor of IPO Outcome:

A topic model approach

by

Christian Emidi

Sebastian Galán

May 2022

Master’s Programme in Data Analytics and Business Economics

Supervisor: Jakob Bergman

Abstract

It is beneficial for both investors and companies to avoid the detrimental consequences of overpricing during an initial public offering (IPO). Prospectuses are an important source of information for potential investors. Through Latent Dirichlet Allocation (LDA) we extract topics from the summary section of prospectuses S-1 for companies holding an IPO in the U.S. in 2019-2020. We represent the uniqueness of the companies through the topic proportions each document is composed of and use them, together with the initial offering price, to predict the outcome of the IPO. For the best performing model, we obtain an AUC of 0.80. In line with signalling theory, we argue that prospectuses may indeed send signals able to influence potential investors.

Keywords: initial public offering, overpricing, signalling theory, topic model

Acknowledgements

Our deepest appreciation goes to our supervisor Dr Jakob Bergman whom we would like to thank for his help and guidance throughout the writing process.

We would also like to give special thanks to our families and friends for their unconditional support.

Table of Contents

List of Tables

List of Figures

1 Introduction

Initial public offerings (IPO) are a common approach for companies to raise capital. On these occasions, companies set an initial price and sell a certain number of shares to the public (Ashford & Schmidt, 2022). While raising new capital for growth is the most common reason, there are other factors influencing a firm's decision to go public. Brau and Fawcett (2006) identify in the literature at least four other reasons: lower the cost of capital, facilitate the process of a take-over, offer an opportunity to insiders for personal financial gain, and finally obtain a first-mover advantage. Since 1999 there have been more than 5000 IPOs only in the U.S., peaking in 2021 with 951 IPOs (Statista Research Department, 2022). At the end of the first day of trade, the price may exceed, stay the same, or be lower than the initial offering price. The former scenario is called underpricing, while the latter is called overpricing (Kagan, 2020).

1.1 Problematisation and Purpose

Underpricing is desirable for investors since it yields positive returns on their investment. For companies instead, the benefits of this outcome are more debated. In fact, correctly pricing the stock would allow them to raise more capital than what is raised when stocks are underpriced. At the same time, underpricing shows a strong demand for the companies' stocks, which is considerable as a success. It is noticeable that some companies set the offering price lower on purpose to fuel demand (Kagan, 2020). On the other hand, overpricing is detrimental to both investors and companies. The former incurs a loss on their investment, while the latter have not managed to convince the public of the validity of their operations (Kagan, 2020).

Therefore, being able to predict the outcome of an IPO is of importance to both investors and companies themselves. Predicting IPO outcomes has not received the same attention as other measures of financial performance. Nevertheless, there has been research attempting to predict IPO outcomes; text data is sometimes used to derive measures used as predictors. Among the text documents used, prospectuses are a common source of text data: the Securities and

Exchange Commission (SEC) requires companies to file such documents. An example is prospectus S-1 which provides a picture of the company and its operations to potential investors (SEC, n.d.).

Much of the research, which we discuss more thoroughly in the literature review, focuses on the sentiment and connotations of the words in the prospectuses. For example, the presence of ambiguous words (Arnold, Fishe & North, 2010), uncertain ones (Loughran and McDonald, 2013), as well as a mix of different sentiments associated with words (Ly and Nguyen, 2020) are used in order to determine whether there is a relationship with IPO outcome. At the same time, the frequencies of words associated with family (Chandler, Payne, Moore, & Brigham, 2019) and commitments to practices such as CSR (Pencle & Mălăescu, 2016) are also investigated. Topics are also created from forward looking statements, and different aspects of such topics such as their sentiment is analysed (Tao, Deokar & Deshmukh, 2018). It is arguable that, in line with signalling theory, the presence and the frequency of such words send signals to potential investors, influencing their behaviour, and consequently, the outcome of the IPO. In fact, it is suggested that signals are sent through prospectuses (Daily, Certo, Dalton & Roengpitya, 2003).

The purpose of our paper is to focus on the signals representing the uniqueness of a company. In doing so, we do not focus on the sentiment and connotations of selected words or statements, but rather on the text as a whole. We define common topics extracted from all the documents analysed, which are the prospectus summaries of prospectus S-1. Each prospectus consists then of a different proportion of the defined topics and is therefore unique in such composition. These proportions are then of interest to us, since we deem them to be representative of the idiosyncratic nature of a company. Thus, our paper contributes to the literature by investigating whether the uniqueness of the prospectuses, represented by their topic proportions, is able to predict the outcome of an IPO.

Hence, the research question we investigate is:

To what extent can the unique content of a prospectus predict IPO outcome?

1.2 Topics and Predictions

We address the research question by firstly extracting the topics each document is represented by. For reasons of time and computing power, the documents from which the topics are extracted represent only a portion of the prospectuses S-1, namely the prospectus summary. Nevertheless, this section is argued to be one of the most informative (Hanley & Hoberg, 2010). Ten topics are extracted through the Latent Dirichlet Allocation (LDA) model, and the topic proportions for each document are used as features in the predictive phase of our paper. We also add the offering price decided by companies as an additional variable.

Finally, we use these features to predict the outcome of an IPO. Such outcome is represented by two scenarios: overpricing (=1) or no overpricing (=0). We split our data into train and test data, and we use four different algorithms to carry out the predictions. The algorithms are classification tree (CT), logistic regression (LR), support vector machines (SVM) and random forests (RF). Given that our data is imbalanced, meaning that underpricing is more recurrent than overpricing, we do not simply rely on accuracy. Rather, we use the AUC (Area Under the Curve), displaying the trade-off between the TPR (True Positive Rate) and the FPR (False Positive Rate). We deem TPR to be of special interest to us, since correctly predicting overpricing would on the one hand prevent investors from losing money, while on the other hand avoid companies a detrimental outcome of the IPO. According to these performance metrics, RF performs best with an AUC of 0.80, followed by LR with an AUC of 0.76, and finally by CT and SVM with AUC scores of 0.69 and 0.66 respectively

1.3 Outline of the Thesis

In chapter 2 we discuss extant literature regarding the use of text data in predicting IPO outcome and the theoretical framework guiding our paper; moreover, we discuss how our paper contributes to both the literature and the theory. After that, in chapter 3, we describe the data and the methods used both in extracting the topics and in the predictive phase. Subsequently, in chapter 4, we discuss the results obtained from the analysis relating them to the literature and to the theory. Finally, in chapter 5, we conclude the paper by discussing its limitations and suggestions for further research.

2 Literature Review

Predicting financial performance has been the focus of much research during past years, taking different points of departure. One indicator of financial performance that has been especially interesting to researchers is stock market fluctuations. Different approaches have been implemented to tackle the challenge of predicting the development of stocks in the market, such as fundamental and technical analysis. Text analysis is yet another approach (Schumaker & Chen, 2009; Nguyen, Shirai & Velcin, 2015), one which will be at the centre of our paper.

Different sources of texts have been used to carry out predictions of stock price movements: examples are financial news (Schumaker & Chen, 2009; Schumaker, Zhang, Huang, & Chen, 2012), social media comments (Nguyen, Shirai & Velcin, 2015), financial disclosures (Kraus & Feuerriegel, 2017), and 10-K forms (Loughran & McDonald, 2011).

Despite much research being devoted to predicting stock price movements through text data, the task of attempting to predict IPO outcome, i.e., how the price of the stock develops the day of the IPO, through text data has not received the same attention. Hence, in this chapter, we discuss existing research in the field and how our paper is located and contributes to the literature. Moreover, we discuss the theory that guides our paper.

2.1 Text Data and IPO Prices

Common sources of text used in predicting the outcome of an IPO are prospectuses. These are documents companies list with the Security and Exchange Commission (SEC) (SEC, n.d.). A study conducted by Arnold, Fishe and North (2010) investigates the role risk-related ambiguity has on first-day returns. They use the section of the prospectuses devoted to risks and measure the relative occurrence of ambiguous words. The results of their research indicate that investors are compensated for the ambiguity in the prospectuses: in fact, there is a significant correlation between the ambiguity of the prospectus and first-day returns. A similar result is obtained in a study conducted by Loughran and McDonald (2013), where they count the frequency of words expressing sentiment of uncertainty in prospectus S-1.

Different angles to the explanation of the phenomenon under investigation can be given. For instance, Pencle and Mălăescu (2016) focus on corporate social responsibility (CSR), and through a content analysis identify its presence in prospectuses which are used to explain its importance in predicting certain metrics. One of their results suggest that there is a significant relationship between some aspects of CSR and underpricing. The reduction of uncertainty obtained by providing information about the company's CSR commitment reduces underpricing. Furthermore, Chandler, Payne, Moore, and Brigham (2019) show the presence of a significant relationship between family-oriented businesses and underpricing. In other words, a higher frequency of words with family connotations lead to more uncertainty for investors, thus leading to more underpricing.

Analysing prospectuses is useful not only to explain the relationship between words' connotations and IPO outcome, but also for predictive purposes: this is of interest to our paper. Ly and Nguyen (2020) adopt a sentiment analysis in order to disentangle the effect words used in prospectuses have on the development of the stock price up to thirty days after the IPO has occurred. One of the days they are trying to predict the return for is the first day, i.e., how the price closes at the end of the IPO date. Specifically, the researchers aim at predicting the stock price based on the meaning of the words used in the prospectus and whether they are connotated with negative or positive feelings, as well as uncertain and litigious words. Subsequently, they compare the predictive power of different algorithms compared both to each other and to random guesses. The result is that through logistic regression, the features extrapolated by the prospectus, were performing better than random guess by more than 9%, with differences witnessed in the different time periods following the IPO.

Moreover, of special interest to us is a study conducted by Tao, Deokar and Deshmukh (2018) in which they attempt to predict IPO outcome through the use of forward-looking statements retrieved from the prospectus 424B4. In doing this, they use a certain portion of the prospectus, namely the Management's Discussion and Analysis, which is argued to be one of the most revealing sections of the prospectus. Topics are then derived from the resulting statements and different features are created such as the sentiment of the topics, the readability, etc. The result shows that while such features are useful in predicting first day return, other predictors commonly used to predict IPO first day return perform better on their own.

Our study contributes in different ways to the extant literature. Firstly, we focus on the prospectus S-1 rather than the 424B4, contrary to some of the papers previously discussed. In line with Loughran and McDonald (2013) we recognise that while the prospectus 424B4 is filed on the day of the IPO or some days after, prospectus S-1 is filed before the IPO date. Thus, considering also the theory we adopt, we prefer to use a text source that potential investors have had the possibility to read prior to the IPO. Furthermore, the topic modelling we will implement will not focus only on forward looking statements as done by Tao, Deokar and Deshmukh (2018). Additionally, we will consider another of the previously mentioned revealing portions of the prospectus, namely the prospectus summary. Finally, the topics we create are not going to be analysed through the sentiment of the words used as done by most of the literature but should rather represent the idiosyncratic nature of a company.

2.2 Theoretical Framework: Signalling Theory

Prospectuses are important sources of signals (Daily et al., 2003). The importance of signals has been theorised by signalling theory. Although originally applied to the job market, the author himself recognises its potential relevance in other markets (Spence, 1973). In fact, one could draw a parallel between the decision to invest in an employee and the decision to invest in a company: both are uncertain ones. In fact, signalling theory is claimed to be one of the most relevant theories in understanding IPOs (Daily et al., 2003).

Companies, through their prospectuses, send signals to investors who elaborate these signals in order to gain an understanding of the company and the profitability of an investment in such a company (Daily et al., 2003). In turn, this “impact[s] the price at which they are willing to purchase IPO shares” (Daily et al., 2003, p. 276). Multiple studies have shown the importance of signals in IPO (Connelly, Certo, Ireland & Reutzel, 2011).

All the research discussed previously shows the importance of words in influencing investors’ perception of a company. As seen, in the case of IPOs, these words are presented in prospectuses, with S-1 giving initial insights into a company. Hence, it is arguable that companies, through said prospectuses, intentionally or unintentionally provide an image of themselves, by sending signals.

Hence, we consider the way a company describes itself to potential investors in the prospectus summary to be important. Given this, in our study we see each company as a composition of topics, and we believe that such composition may be signalling its distinctiveness to potential investors. Consequently, we expect the uniqueness of each company to have an influence on investors' assessment, affecting their willingness to invest and to receive compensation for such an investment. Thus, it is arguable that the signals companies send out by describing their activities may affect the outcome of an IPO. Therefore, our approach takes a different angle, using the topic proportions in the prospectus summary as predictors.

3 Methodology

In this chapter, we present the methods used to conduct the analysis. We start by presenting our data and how we process it. Subsequently, we present the unsupervised model we use to extract topics from the prospectus summaries, namely the Latent Dirichlet Allocation (LDA). Finally, we discuss the algorithms we use to predict IPO outcome, and different metrics used to assess their performance.

3.1 Data Source

The data used in this paper stem from different sources. The core of our study lies in the S-1 forms that companies submit to the SEC when pursuing to go public. We manually download these S-1 forms from the SEC's filing system called EDGAR. The S-1 is a document containing information about the company and its goals, strategies, strengths, and risks (SEC, n.d.).

However, we only focus on the section named prospectus summary. Different reasons justify our choice. The S-1 forms are of varied length, but usually several hundred pages long and require significant computational power to process. Large parts of the text contain standardised and mandatory text required by the SEC; these parts of the text often consist of legal texts and are therefore notably similar across companies. Additionally, some studies suggest that the prospectus summary, compared to other sections, contains more informative content for potential investors (Hanley & Hoberg, 2010). In the end, limiting our focus to the prospectus summary not only makes the text more feasible to process, but allows us also to access more differentiated information about the companies making it easier to extract sensible topics from them.

This study only makes use of S-1 forms from IPOs between 2019 and 2020 and which are available on EDGAR. We exclude some companies from the analysis since we are unable to find their S-1 forms. Other files seem broken when downloaded and are also excluded. In the end, we consider 371 companies.

Our analysis also requires financial data which we retrieve from two different sources in order to calculate the first-day returns. Data about companies’ IPO offering price is downloaded from StockAnalysis.com, while data on the closing price of the stock on its first day on the market is downloaded through the tidyquant library in R from Yahoo Finance.

3.2 Data Processing

As with any work related to text analytics, the text data needs to be processed and cleaned for it to be ready to analyse; we use R for this task. The S-1 prospectuses are separated into individual words in order to filter out numbers, punctuations, and stop words, i.e., words without meaning for the purpose of our analysis such as “the” or “and”. The words are also lemmatised, meaning that words’ inflected forms are substituted by their lemma. For instance, the words “jumping”, “jumped”, and “jumps” simply become “jump”. This method does not come without drawbacks as it requires additional computational power but at the same time it provides us with usable words for our topic model.

While processing our text we use the concept of tf-idf deletion. Tf-idf is the result of the multiplication of two measures: tf (term frequency), i.e., how often a word is repeated in a document, and idf (inverse document frequency), which prioritises words that seldom occur in a corpus (Silge & Robinson, 2022). With tf-idf we would give more importance to words that represent the distinctiveness of a prospectus, and consequently of a company. This is accomplished by choosing words that are frequent and unique to a company at the same time. Once we have the list of words with their tf-idf score, we filter out words that have a tf-idf score lower than 0.001. This approach thus uses a “tf-idf deletion baseline” (Fan, Doshi-Velez and Miratrix, 2019, p. 213). In this way, we reduce the number of observations in our dataset by removing words we deem to be less informative.

After carefully analysing the remaining words, we still notice numerous words that would not make much sense to our topics, such as company-specific terms, names, and abbreviations. Despite already filtering out stop words, we decide to remove these words as well. In order to do so, we use a list from MIT containing 10.000 English words (MIT, n.d.). After extracting our topics and analysing the results, there are still some words that add little or no value, negatively impacting the topics derived. Therefore, we remove them. A list of these words can be seen in Table A1.

Regarding the financial data, we calculate the percentage first-day return $R_{first-day}$ as follows:

$$R_{first-day} = \frac{P_{close} - P_{IPO}}{P_{IPO}} \times 100$$

where P_{close} is the closing price of the stock on its first day of trading and P_{IPO} is the initial offering price of the stock. The distribution of the first-day returns can be seen in Figure 1. We can see a clear concentration close to zero and a skewness towards the right; some outliers also become evident. We inspect the outliers but are not able to find any particularity to explain their occurrence. Nevertheless, since we are pursuing a classification problem, their presence causes little disturbance. The distribution's skewness to the right indicates that our data ultimately is somewhat imbalanced, meaning that we have more observations where the first-day return is positive compared to negative. This is something to bear in mind when evaluating the classifiers' performance, since the prevalence of one class over the other may influence certain metrics.

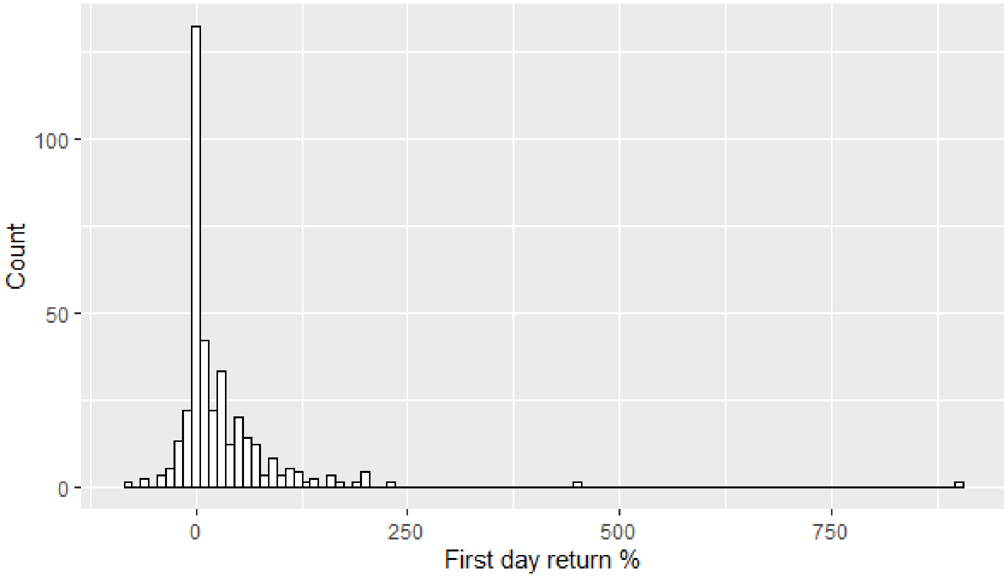


Figure 1. Histogram showing the distribution of the outcome variable first-day return

We use the topic composition of the prospectus summaries to predict the outcome of the IPO, i.e., overpricing or no overpricing. Since we have a classification problem, we assign a 1 to negative returns and 0 to positive or unchanged returns. Finally, as seen in Table A2, we use a 75/25 percent split, dividing the data into training and test data, respectively.

3.3 Latent Dirichlet Allocation (LDA)

As previously mentioned, we seek the composition of topics each document consists of. First, we need to extract topics from the prospectus summaries, and we do that by using probabilistic topic modelling through the LDA model, also known as Latent Dirichlet Allocation, which is an unsupervised technique (Blei, 2012). This model needs to be fed with the unigrams previously obtained in the phase of text processing. The representation of the processed text is called "Bag of Words", where each word is a variable and is linked to each document by the occurrence of that word in the specific document. In our case, we have the number of times each word is present in the prospectus summaries, identifying each document by its ticker (Tao, Deokar & Deshmukh, 2018). One potential drawback of this approach is that it does not consider the order of the words (Blei, 2012).

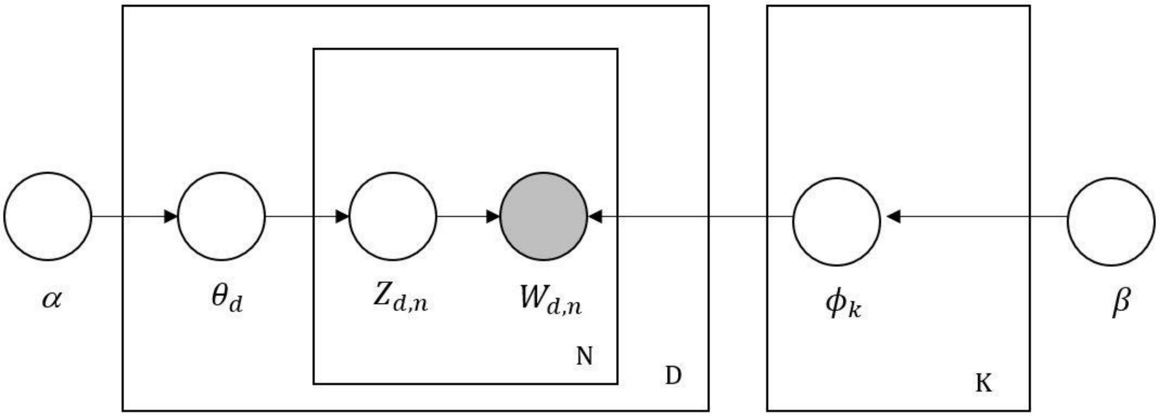


Figure 2. Directed Acrylic Graph of the LDA model (Adapted from Blei, 2012)

The LDA model derives topics from our documents and then it is able to show how much each topic composes a document. The Directed Acyclic Graph (DAG) in Figure 2 visualises the LDA, where each circle represents a random variable, and the shadowing indicates whether such variable is observed (Blei, 2012). The rectangles suggest that there are N words, D documents, and K topics. The values of the random variables positioned in rectangle D changes with each document *d*; the values of the random variables positioned in rectangle N changes with each word *n*; the values of the random variable in rectangle K changes with each topic *k*.

The two random variables which are not positioned in the rectangles, namely α and β, are parameters of two Dirichlet distributions. From these distributions, probabilities are sampled, and these parameters control whether the probabilities sampled are more or less uniform. The probabilities sampled from the Dirichlet distribution with parameter α are collected in θ_d, while the probabilities resulting from the Dirichlet distribution with parameter β are collected in ϕ_k. They are sampled for each document *d* and topic *k* respectively. The former, θ_d , represents a categorical distribution over topics for each document *d*; the latter instead, ϕ_k , represents a categorical distribution over words for each topic *k*. From the former distribution, for the *n*-th word in the *d*-th document, we sample a topic $Z_{d,n}$. Finally, having sampled a topic $Z_{d,n}$, the *n*-th word in the *d*-th document is sampled from the categorical distribution represented by ϕ_k , and is represented by $W_{d,n}$ (Magnusson, 2021).

Therefore, a word in the document is generated following this process. In our case, we do not need to generate words as this is the only element we observe, contrary to the rest that is unobserved. Thus, we revert the procedure, and starting from the words we discover the topics that have generated them and subsequently the topic distribution for each document. Therefore, since the topics are not observed, but are extracted from the words in the documents, they are latent (Blei, 2012).

As a method for our LDA model we use the Gibbs sampling since it is not feasible to compute the distribution of our topic structure given the words, known as posterior (Blei, 2012). This method allows us to sample from the posterior, eventually approximating the underlying distribution. In Gibbs sampling each variable is sampled by being conditioned on the other variables. For the *n*-th word, we compute the probability that this word belongs to a topic *k*. This result is based on both the probability that the specific word belongs to the specific topic and the probability that the specific topic belongs to the document *d*. The algorithm initialises the topic assignment of each word, and then computes the conditional probability for each word. In doing this, it does not consider the current topic assignment of the word for which the conditional probability is being computed. The process is iterated, and, in the end, the algorithm allows us to estimate the parameters θ and ϕ (Griffith & Steyvers, 2004). We are only interested in the former since these probabilities will serve as our independent variables.

We run the model in R, and as mentioned previously we choose as method the Gibbs sampling with the default number of iterations, i.e., 2000. The only parameter we experiment with is the number of topics *k*. In the choice of topics, we are guided both statistically and by practical reasons. We use four optimisation metrics, namely CaoJuan, Arun, Deveaud, and Griffith to guide our choice of topics. The former two need to be minimised while the latter two need to be maximised (Nikita, 2020). The result can be seen in Figure A1. We set an upper limit of candidate topics equal to 20 since we believe that interpretation would suffer with more topics than that. We settle for the *k* that we deem to be a good trade-off between the metrics as well as the usefulness for the analysis. For instance, an adequate choice according to the metrics is *k* = 3, but we discard such a number since we deem it to be too small, potentially jeopardising the subsequent analysis. We try with several different *k*, but a final manual check of the top terms in each topic, which we discuss in the results chapter, convinces us to settle for *k* = 10.

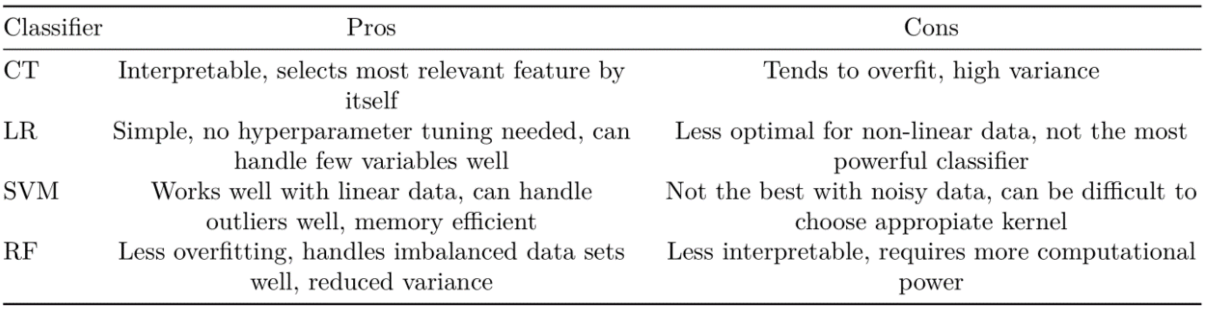
3.4 Algorithms

When predicting IPO outcome, we use different machine learning algorithms for classification problems. The ones considered are classification tree, logistic regression, support vector machines, and random forest.

SVM and RF are by several scholars considered to be the default classification algorithms due to their consistency and accuracy in a wide range of applications (Zhang, Liu, Zhang, & Almpanidis, 2017). RF is an ensemble learning method based on trees. It builds numerous trees, each with a random subset of predictors and in the case of classification, takes a majority vote of the trees to determine the output class. Because these trees are built independently and on randomised subsets of predictors, both variance and bias are reduced (Belgiu & Dragut, 2016). SVM seeks a linear function that separates samples from both categories as widely as possible; when it fails to do so, it relies on kernels to transform the data into a higher dimensional space and then tries to fit a separator in that transformed space (Zhang et al., 2017). LR makes use of the logistic function to model the probability of the outcome variable to belong to a certain class, where the regression coefficients of the logistic function are then usually fit using the maximum likelihood method (James, Witten, Hastie, & Tibshirani, 2021). Given that we define no overpricing as class 0 and overpricing as class 1, this method seeks coefficient estimates such that the predicted probability of overpricing for a company's IPO is as close as possible to the IPO's actual class status (James et al., 2021). Finally, a CT learns several rules of where to split the data into different branches resulting in internal nodes if it has further splits, or terminal nodes if it does not. The output in the terminal nodes is then based on the most commonly occurring class in that terminal node (James et al., 2021).

Even though these are all well-known machine learning algorithms, each one of them brings specific strengths and limitations. The most notable of them are presented in Table 1.

Table 1. Benefits and disadvantages of the different machine learning algorithms used (adapted from Dineva and Atanasova, 2020; Gupta, 2020; Sarker, 2021). Note: classification tree (CT), logistic regression (LR), support vector machines (SVM), random forest (RF).



Classifier	Pros	Cons
CT	Interpretable, selects most relevant feature by itself	Tends to overfit, high variance
LR	Simple, no hyperparameter tuning needed, can handle few variables well	Less optimal for non-linear data, not the most powerful classifier
SVM	Works well with linear data, can handle outliers well, memory efficient	Not the best with noisy data, can be difficult to choose appropiate kernel
RF	Less overfitting, handles imbalanced data sets well, reduced variance	Less interpretable, requires more computational power

3.4.1 Performance metrics

We use several methods to measure and compare the performance of the different classifiers. One of the measures considered is accuracy, which is the ratio of the correct predictions over the total number of predictions. Even though this measure is widely popular, it does not account for the distribution of the dataset and imbalances in the data (Fawcett & Provost, 1997). Hence, other measures are also used to develop deeper insight into the classifiers' performance.

Another measure we use is the area under curve (AUC) and receiver operating characteristics (ROC). The ROC is a probability curve based on the relationship between the true positive rate (TPR, also known as recall or sensitivity) on the y-axis and the false positive rate (FPR) on the x-axis. While the TPR is the number of positives correctly predicted in relation to the total number of actual positives, the FPR computes the number of negatives incorrectly predicted in relation to the total number of actual negatives (Fawcett, 2006). The ROC is helpful because having a high TPR does not necessarily mean that the model is performing well. In other words, predicting every single observation to belong to the first class, for instance, yields a TPR of 100 percent. However, it also means that we have a FPR of 100 percent since we incorrectly classify all observations belonging to the second class. Because the ROC relies on these measures, it does not change if the class distribution changes (Fawcett, 2006). According to He and Ma (2013), analysis with the help of the ROC curve is useful in this sense because it does not have any bias towards either the majority or minority class. Since the ROC graph is a unit square with area 1, it becomes relatively simple to compute the AUC which can help us compare the different classifiers. The AUC gives a measure of how well a classifier can distinguish the data: an AUC of 1 is the best-case scenario while an AUC of 0.5 is only as good as a random guess. Furthermore, we follow the rule of thumb used by Hosmer, Lemeshow, and Sturdivant (2013) and consider:

$AUC = 0.5$	No better than coin toss
$0.5 < AUC < 0.7$	Marginally better than coin toss
$0.7 \leq AUC < 0.8$	Acceptable
$0.8 \leq AUC < 0.9$	Excellent
$AUC \geq 0.9$	Outstanding.

4 Results

In this chapter we present the results of our analysis and discuss them in relation to the literature and the theoretical framework presented previously. We start by presenting the results of the LDA; subsequently, we extract the features which are needed for the following phase; finally, we present the results from the prediction process comparing the performance of the different algorithms presented in the methodology chapter. We conclude by discussing how the results we obtain fit into the existing literature and the theoretical framework.

4.1 Extracting Topics

We run the LDA model with *k* = 10 and manually check the coherence of the topics. We argue that through this setting we are able to identify clear boundaries between the topics, although some are clearer than others. In Figure 3, it is possible to see the top terms for each topic.

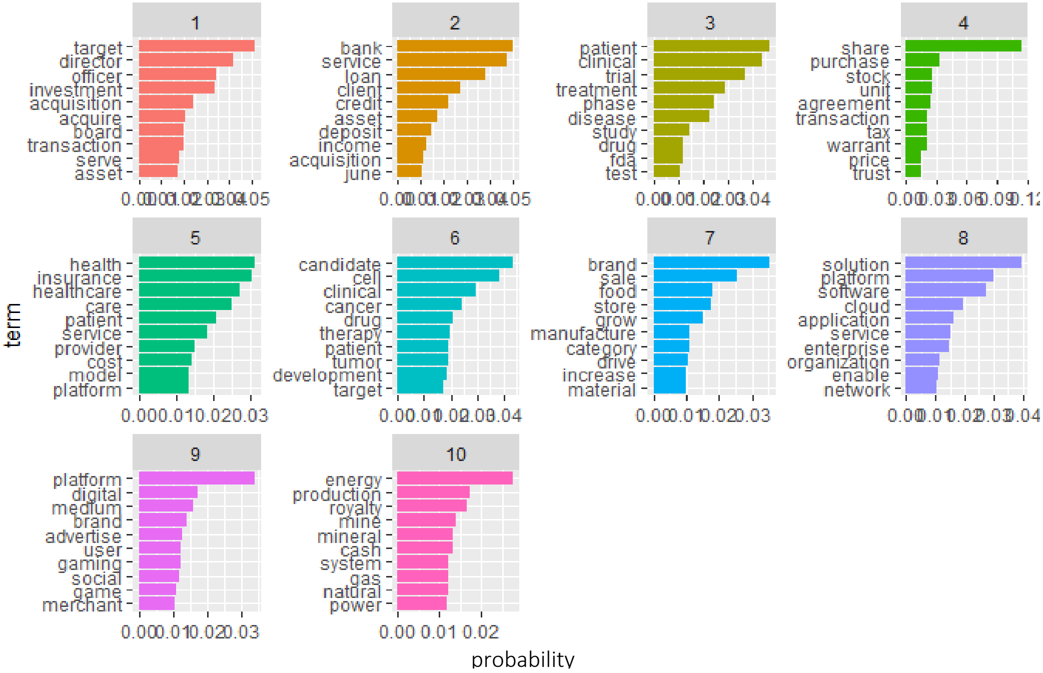


Figure 3. Top terms for each topic, k = 10.

The figure represents the distribution over words for each topic. On the y-axis we see the top 10 terms, while on the x-axis we see the probability for each word to be sampled from the topics.

Having worked with the prospectus summaries we have an idea of the nature of the companies analysed; this knowledge helps us to contextualise the words in the topics and to provide labels for each of them. The labels are presented below:

- Topic 1: SPAC (Special Purpose Acquisition Company)
- Topic 2: Banking
- Topic 3: Research - Medicine
- Topic 4: Finance - Shares
- Topic 5: Insurance - Healthcare
- Topic 6: Research - Oncology
- Topic 7: Commerce - Retail
- Topic 8: Digital Solutions
- Topic 9: Tech - Entertainment
- Topic 10: Resources - Energy

For example, words such as energy, mine, mineral, gas, natural, etc. in topic 10 suggest to us that this topic is related to energetic resources, and thus we label it as resources - energy. Another interesting example is the difference between topic 3 and topic 6. Both have medical terms in them indicating commitment to research in this field. But topic 6 contains words such as cancer and tumour, suggesting to us that this specific topic may be focusing more on oncology. Hence, we label them research - medicine and research - oncology respectively.

Some topics are clearer than others. For instance, topic 7 is more difficult to interpret. Words such as brand, sale, food, store indicate that it may be related to retail. Despite this, other words in this topic are more difficult to position in this category. On the other hand, it is easier to construe topic 8 as belonging to digitalisation since all the words are related to this phenomenon; hence, we name it digital solutions.

For reasons of space, we do not scrutinise each word and argue for why we deem it to be representative of the label provided. We acknowledge that these topics are open to interpretation; nevertheless, we deem them to be in line with our understanding of the prospectuses read.

4.2 Topic distribution for documents

The previous step of extracting the topics is helpful to check their coherence and to label them in order to facilitate the analysis. What we are really interested in is how much each document is composed of a certain topic. That is, we want to know the topic distribution for each document. As mentioned in the methodology part, this is represented by θ, which we estimate by running the Gibbs sampling in the LDA model.

Theta is composed of probabilities: for each document *d* we obtain the probabilities for each topic to compose said document. For example, we expect a company operating in the medical research sector to show a higher probability of belonging to topic 3 (research - medicine) or topic 6 (research - oncology). A company focusing on oncology would probably lean more towards topic 6, while a company focusing on medicine more generally would be composed predominantly by topic 3.

We also expect that certain companies may be composed of a mixture of topics to a larger extent compared to others. For instance, we expect topic 3 (research - medicine) and topic 8 (digital solutions) to contribute extensively to a company operating in medical research with a strong focus on digitalisation. We believe that this flexibility allows us to capture the uniqueness of a company. As in the example above, we are able to differentiate two companies operating in the same sector by the degree to which they focus on digital solutions in the prospectus summary.

We show a sample of companies and their topic distribution in Figure 4. On the y-axis we can see the tickers, which are the unique identifiers of companies on the stock market. On the x-axis instead, we have the probabilities. Each ticker represents a document, which is the prospectus summary for a specific company. The probabilities tell us the topics each document is likely to be composed of. The topics are represented in the legend, and each colour is associated with a topic.

For example, if we consider WISH, we see that it is composed mainly by topic 9 (tech - entertainment) and to a lower extent by topic 7 (commerce - retail) and topic 8 (digital solutions). WISH is the ticker associated with the company Wish, which describes in the prospectus summary its mission to be an “entertaining mobile shopping experience to billions of consumers around the world” (Wish, 2020, p.1). Pinterest, a platform showing “visual recommendations … based on … personal taste and interests" (Pinterest, 2019, p.1), whose ticker is PINS, is arguably well positioned in being composed mostly of topic 9 (tech - entertainment). Some documents are not as clear-cut, and no individual topic takes the lead. For instance, HAAC is fairly distributed between topic 5 (insurance - healthcare) and topic 1 (SPAC), with a lower but still noticeable contribution from topic 8 (digital solutions). According to their prospectus they are a blank check company, thus justifying the presence of topic 1 (HAAC, 2020). Moreover, they state that "the intersection of technology and healthcare is one of the most significant value creation opportunities of this decade" (HAAC, 2020, p.2) and that healthcare ought to be "consumer-centric, data-driven, [and] cloud-based" (HAAC, 2020, p.2). It is arguable that the topics capture the nature of the company in a fairly accurate way. The same could be said for the other tickers which we do not scrutinise further for reasons of space.

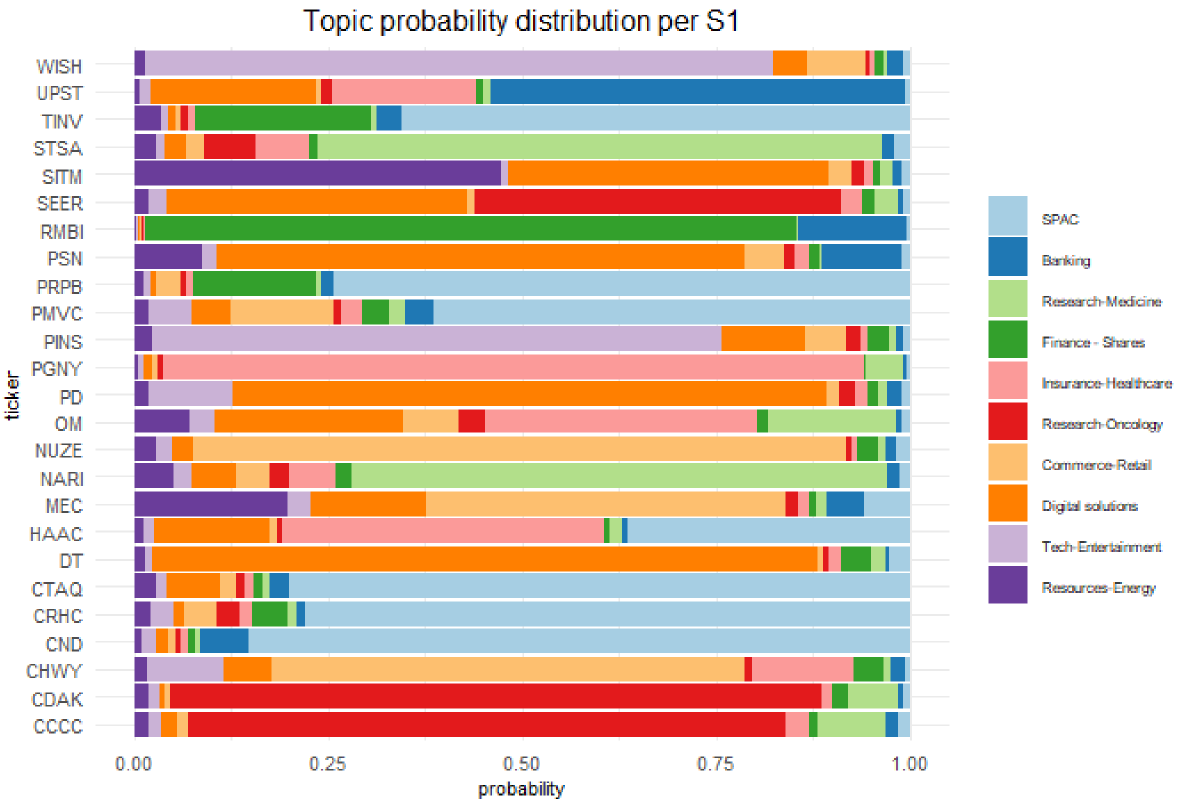


Figure 4. Topic distribution for a sample of documents. Format inspired by Nagelkerke & van Gils (2020).

The result of the LDA is important on its own. Since each company prospectus summary has a unique distribution over topics, we are able to disentangle the unique nature of each company. In the literature review and theoretical framework presented previously, we have seen that companies send all sorts of signals, and that these could be both intentional and unintentional. We argue that the distinctiveness captured in our paper, which is represented by the probabilities of each topic to compose a document, may be sending signals to potential investors affecting the attractiveness of a certain company to their eyes. It is not our purpose to investigate the intentionality of such signals; rather, we are interested in them as such. We therefore focus on whether these signals could be responsible for influencing the outcome of an IPO.

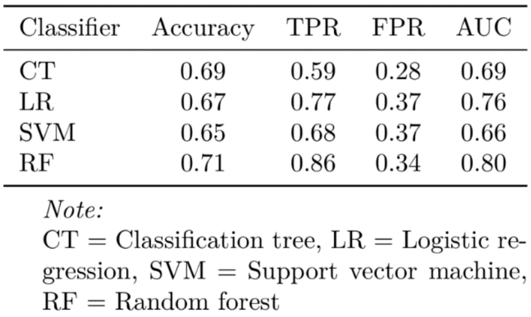
4.3 Independent Variables

The purpose of the first part of our analysis, i.e., the LDA model, is to obtain the topic probabilities for each document. These probabilities, which we have discussed extensively previously, are the features we are going to use in our prediction process. Additionally, we also include the initial offering price. The reason for this is twofold: first, it is information provided in the prospectus summary that is lost when we process the data; secondly, we argue that it is important to consider the offering price since, even though the topics composing a prospectus summary may suggest positive returns, if the price is too high, that may not be the case.

4.4 Predictions and Performance

Using the topic probabilities for each document and the company's IPO price as features, we use four classification methods to fit our different models, namely classification tree, logistic regression, support vector machine and random forest. To compare their performance, we use four different measures which can be seen in Table 2. We begin by discussing each of these four machine learning classifiers and their performance, starting from the classification tree.

Table 2. Four performance metrics for the classifiers used; random forest seems to be the best performing model.



Classifier	Accuracy	TPR	FPR	AUC
CT	0.69	0.59	0.28	0.69
LR	0.67	0.77	0.37	0.76
SVM	0.65	0.68	0.37	0.66
RF	0.71	0.86	0.34	0.80

Note:
CT = Classification tree, LR = Logistic regression, SVM = Support vector machine, RF = Random forest

4.4.1 Classification Tree

The accuracy of the classification tree is 0.69. Despite this, as previously seen, accuracy may not be the most suitable metric given that our data is somewhat imbalanced. In fact, given our test data, simply guessing that every IPO will yield a negative return would result in an accuracy of 0.28, and guessing every IPO to have a non-negative return would result in an accuracy of 0.72. We therefore need to be cautious when assessing the models' accuracy.

Given the purpose of our study, we argue that it becomes relatively more important to have a high TPR than a low FPR. This is because we argue that it is worse for an investor to lose money on an IPO, i.e., investing in a company which is supposed to be underpriced but which instead is overpriced (false negative and type II error), compared to losing out on potential money-making opportunities (false positive and type I error). Our model reaches a TPR of 0.59 which might not be optimal for potential investors. Regarding the AUC, an area of 0.69 is just under the acceptable level in reference to our established thresholds.

After inspecting the model's cross validated error from different complexity parameters, it becomes apparent that a classification tree might just not be the right classifier for our data. As can be seen from Table A3, the cross validated error (xerror) increases after each split and is a potential indicator that the model will generalise poorly. That is, as the complexity parameter (CP) decreases, the tree is penalised less for additional terminal nodes and thus grows as a result. But since the lowest cross validated error is reached with the highest complexity parameter of 0.07, in essence, the table is suggesting that the best possible tree is just the root node with zero splits and will overfit when adding any independent variable. Despite tuning the model with different numbers of folds for cross validation, different number of observations required to split a node, and pruning the tree, results do not improve.

4.4.2 Logistic Regression

We use logistic regression with a threshold of 0.3. Among the cut-offs tried, this one yields the most satisfactory result. In other words, the probability of belonging to class 1 (overpricing) just needs to be above 0.3 to be assigned to that class. The logistic regression model has lower accuracy than the classification tree and a higher FPR. However, it reaches a significantly higher TPR of 0.77 and AUC of 0.76. As argued before, having a higher TPR can arguably be of more value to investors avoiding negative returns on their investments, and the logistic regression model reached the second highest TPR out of the four classification algorithms.

While inspecting the coefficient estimates in Table A4, we can observe that while the IPO price is statistically different from 0 at the 0.01 level, the only statistically significant topic is the one relating to *Tech - Entertainment* at the 0.05 level. The rest of the predictors have such high standard error and small coefficients that the model cannot reject the null hypothesis that the estimates are statistically different from zero.

4.4.3 Support Vector Machine

The support vector machine model has a relatively high TPR, but it also has the poorest performance out of all models according to the accuracy and AUC. This is mainly due to the model focusing on accurately predicting no overpricing, i.e., that there would be non-negative first-day returns. A FPR that high means that out of all the IPOs which had positive first-day returns, the model would predict negative first-day returns around 37 percent of the time. In essence, investors would often be suggested that underpriced IPOs would have negative returns.

Despite trying numerous variations of the kernels, the FPR remained somewhat high across models considered and the AUC does not improve much. Increasing the cost parameter considerably is in the end one way of increasing TPR but comes with overfitting issues. One must be careful in not increasing the complexity of the model too much by having a very high-degree polynomial kernel or adapting the cost parameter too much to the data. This can lead the model to overfit and generalise poorly. Another indication of the relatively poor performance of the support vector machine with this data was that out of the 278 observations used for building and training the model, it resulted in 229 support vectors. One could generally expect the support vectors to be relatively few compared to the total number of observations. As suggested by Burges (1998), this could hint that the predictors used do not have much predictive power for the dependent variable or that the model is overfitting. However, as discussed, increasing the cost parameter does not improve the model to a considerable degree.

4.4.4 Random Forest

From the metrics in Table 2, the random forest is the best performing model since it essentially outperforms the other models with a TPR of 0.86 and an AUC of 0.80. The model is built using 1000 trees with 4 randomly selected variables considered in each split, and a minimal node size of 10.

Upon further analysis, however, we observe that two of the eleven variables have a negative variable importance measured by the mean decrease of accuracy. These are topic 5 (insurance - healthcare) and topic 8 (digital solutions). This implies that including those variables in the model decreases its accuracy, as seen in Figure A2. Meanwhile, IPO price, topic 1 (SPAC), topic 7 (commerce - retail), and topic 9 (tech - entertainment) seem to be the most important variables for this model.

Because of the high TPR and AUC of the random forest model, we deem it to have the most value for a potential investor. While it might have a higher probability to misclassify IPOs with positive returns as IPOs with negative returns, it does well in distinguishing the IPOs that have negative returns.

4.5 Discussion

This paper contributes to the literature concerning text data as a predictor of IPO outcome by using the summary section of the prospectus S-1. Our study shows that it is possible to derive meaningful topics from the prospectus summaries. Additionally, we suggest that the topic proportions representing the uniqueness of companies, together with the initial offering price, can have some predictive power with respect to the IPO outcome. While some studies suggest that first-day returns can be related to ambiguity and sentiments in the prospectuses (Arnold, Fishe, & North, 2010; Loughran & McDonald, 2013), our paper also indicates that the unique content and signals retrieved from the prospectus summaries may also play a role in influencing investors and IPO outcome.

As previously mentioned, we agree with the argument that companies send out signals through their prospectuses (e.g. Daily et al., 2003), and the unique content of the prospectuses can be understood as such. We argue that the models obtained from our analysis may have value in capturing these signals and consequently predict whether a company's stock will increase or decrease in price on its first day of trading. This is especially true in predicting overpricing and could therefore help investors avoid investing in IPOs whose value is predicted to decrease.

However, as can be seen from the results above, not all topics may represent a decisive signal in predicting the outcome of an IPO. In the classification tree, not all variables are considered when doing the splits and are therefore left out. The first split considers the IPO price while the second and third split consider topic 9 (tech - entertainment) and topic 7 (commerce - retail) respectively. The logistic model shows that only the coefficients of topic 9 (tech-entertainment) and the initial offering price are different from zero at a statistically significant level. The variable importance from the random forest shows that omitting topic 5 (insurance - healthcare) and topic 8 (digital solutions) would increase the accuracy of the model, while variables such as IPO price, topic 1 (SPAC), topic 7 (commerce - retail), and topic 9 (tech - entertainment) are considerably more important for the model. It becomes apparent that especially topic 9 (tech - entertainment), topic 7 (commerce - retail) and the IPO price are the most relevant variables in our models. Given this, the presence of certain topics may be perceived as a stronger signal by investors compared to others whose presence may not be decisive in affecting investors' perception of a company.

Furthermore, previous research regarding the utility of the content in prospectuses suggest that the prospectus summary has more informative content relative to other sections (Hanley & Hoberg, 2010). Our study further highlights the relevance of the prospectus summaries as a source of text data for predicting IPO outcomes.

5 Concluding Remarks

In this paper, we have seen that the uniqueness of the prospectuses' summaries, represented by their topic composition, could indeed be seen as sending signals to potential investors. It is arguable that the topic proportions, together with the offering price, predict IPO outcome in a satisfactory way. Our research comes with its own limitations in several regards. Many of these can be an opportunity for further studies wishing to take on the endeavour of financial forecasting using text documents such as the S-1 prospectus.

One limitation of this study is in regard to the creation of the topics. Even though we rely on some performance metrics to guide our decision regarding the number of topics to be extracted, we are also guided by our own interpretation. This might open for interpretability concerns as other researchers most probably have other understandings of the number of topics that suit the data best. Additionally, our study does not control for variables that can be important for the concept of overpricing. It may be possible that certain firm characteristics, such as company size or revenue, sends important signals to investors, thus affecting the IPO outcome. Future studies could strengthen the results by controlling for additional variables. Another possible limitation that can be explored in future research is the scope of our study. We lay our focus on the years 2019 and 2020, but it might be possible that other relationships exist for other years. It is then also worth noting that 2020 was an extraordinary year for the stock market due to the pandemic, and we cannot determine how this might have influenced the stock movements on their first day of trading.

References

Arnold, T., Fishe, R.P.H., & North, D. (2010). The Effects of Ambiguous Information on Initial and Subsequent IPO Returns, *Financial Management*, [e-journal] vol. 39, no. 4, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 17 April 2022]

Ashford, K., & Schmidt, J. (2022). What Is An IPO?, Forbes, 24 March, Available online: https://www.forbes.com/advisor/investing/initial-public-offering-ipo/ [Accessed 10 May 2022]

Belgiu, M. & Drăguţ, L. (2016). Random Forest in Remote Sensing: A Review of Applications and Future Directions, *ISPRS Journal of Photogrammetry and Remote Sensing*, [e-journal] vol. 114, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 7 May 2022]

Blei, D.M. (2012). Probabilistic Topic Models, *Communications of the ACM*, [e-journal] vol. 55, no. 4, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 3 May 2022]

Brau, J.C. & Fawcett, S.E. (2006). Initial Public Offerings: An Analysis of Theory and Practice, *The Journal of Finance*, [e-journal] vol. 61, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 10 May 2022]

Burges, C.J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, Available online: https://www.microsoft.com/en-us/research/publication/a-tutorial-on-support-vector-machines-for-pattern-recognition/?from=http%3A%2F%2Fresearch.microsoft.com%2Fpubs%2F67119%2Fsvmtutorial.pdf [Accessed 15 May 2022]

Chandler, J.A., Payne, G.T., Moore, C., & Brigham, K.H. (2019). Family Involvement Signals in Initial Public Offerings, *Journal of Family Business Strategy*, [e-journal] vol. 10, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 18 April 2022]

Connelly, B.L., Certo, S.T., Ireland, R.D., & Reutzel, C.R. (2011). Signalling Theory: A review and assessment, *Journal of Management*, [e-journal] vol. 37, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 17 April 2022]

Daily, C.M., Certo, S.T., Dalton, D.R., & Roengpitya, R. (2003). IPO Underpricing: A meta-analysis and research synthesis, *Entrepreneurship: Theory & Practice*, [e-journal] vol. 27, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 16 April 2022]

Dineva, K. & Atanasova, T. (2020). Systematic Look at Machine Learning Algorithms - Advantages, Disadvantages, and Practical Implications, in Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds), *Distributed Computer and Communication Networks*, Cham: Springer

Fan, A., Doshi-Velez, F., & Miratrix, L. (2019). Assessing Topic Model Relevance: Evaluation and informative priors, *Statistical Analysis & Data Mining*, [e-journal] vol. 12, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 1 May 2022]

Fawcett, T. (2006). An Introduction to ROC Analysis, *Pattern Recognition Letters*, [e-journal] vol. 27, no. 8, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 6 May 2022]

Fawcett, T. & Provost, F. (1997). Adaptive Fraud Detection, *Data Mining and Knowledge Discovery*, [e-journal] vol. 1, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 10 May 2022]

Griffiths, T.L., & Steyvers, M. (2004). Finding Scientific Topics, *Proceedings of the National Academy of Sciences of the United States of America*, [e-journal] vol. 101, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

Gupta, S. (2020). Pros and Cons of Various Machine Learning Algorithms. Available online: https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6 [Accessed 11 May 2022]

HAAC. (2020). Form S-1 Registration Statement under the Securities Act of 1933 [pdf], Available at: https://www.sec.gov/Archives/edgar/data/1824013/000110465920118046/tm2031762-2_s1.htm [Accessed 12 May 2022]

Hanley, K.C. & Hoberg, G. (2010). The Information Content of IPO Prospectuses. *The Review of Financial Studies*, [e-journal] vol. 23, no. 7, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

He, H. & Ma, Y. (2013). Imbalanced Learning: Foundations, Algorithms, and Applications, New Jersey: Wiley-IEEE Press

Hosmer Jr. D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression, New Jersey: Wiley

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R, 2nd edn, New York: Springer

Kagan, J. (2020). Underpricing, Available online: https://www.investopedia.com/terms/u/underpricing.asp [Accessed 10 May 2022]

Kraus, M., & Feuerriegel, S. (2017). Decision Support from Financial Disclosures with Deep Neural Networks and Transfer Learning, *Decision Support Systems*, [e-journal] vol. 104, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

Loughran, T., & McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, [e-journal] vol. 66, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 17 April 2022]

Loughran, T., & McDonald, B. (2013). IPO First-day Returns, Offer Price Revisions, Volatility, and Form S-1 Language, *Journal of Financial Economics*, [e-journal] vol. 109, no. 2, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 17 April 2022]

Ly, T.H., & Nguyen, K. (2020). Do Words Matter: Predicting IPO performance from prospectus sentiment, IEEE 14th International Conference on Semantic Computing (ICSC), Available online: https://ieeexplore.ieee.org/document/9031486 [Accessed 18 April 2022]

Magnusson, M. (2021). Lecture 7: Statistical Analysis of Large Textual Data, DABN14, powerpoint presentation, LUSEM Lund, 8 December 2021

MIT (n.d.). 10000 Word list - MIT. Available online: https://www.mit.edu/~ecprice/wordlist.10000 [Accessed 10 April 2022]

Nagelkerke, J. & van Gils, W. (2020). NLP With R part 3: Using Topic Model Result to Predict Michelin Stars, Available online: https://medium.com/broadhorizon-cmotions/nlp-with-r-part-3-using-topic-model-results-to-predict-michelin-stars-ba8ec1b182c2 [Accessed 5 May 2022]

Nguyen, T.H., Shirai, K., & Velcin, J. (2015). Sentiment Analysis on Social Media for Stock Movement Prediction, *Expert Systems With Applications*, [e-journal] vol. 42, no. 24, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

Nikita, M. (2020). Select Number of Topics for LDA Model, Available online: https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html [Accessed 1 May 2022]

Pencle, N., & Mălăescu, I. (2016). What's in the Words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses, *Journal of Emerging Technologies in Accounting*, [e-journal] vol. 13, no. 2, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 18 April 2022]

Pinterest. (2019). Form S-1 Registration Statement under the Securities Act of 1933 [pdf], Available at: https://d18rn0p25nwr6d.cloudfront.net/CIK-0001506293/34798acf-d187-444c-bc45-79c97304b466.pdf [Accessed 12 May 2022]

Sarker, I.H. (2021) Machine Learning Algorithms, Real-world Applications and Research Directions, *SN Computer Science*, [e-journal] vol. 2, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 10 May 2022]

Schumaker, R.P., & Chen, H. (2009). Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText system, *ACM Transactions on Information Systems*, [e-journal], vol. 7, no. 2, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

Schumaker, R.P., Zhang, Y., Huang, C-N., & Chen, H. (2012). Evaluating Sentiment in Financial News Articles, *Decision Support Systems*, [e-journal] vol. 53, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

SEC (n.d.). Investing in an IPO [pdf], Available at: https://www.sec.gov/files/ipo-investorbulletin.pdf [Accessed 12 April 2022]

Silge, J., & Robinson, D. (2022). Text Mining with R: A tidy approach, [e-book] Available online: https://www.tidytextmining.com/ [Accessed 2 May 2022]

Spence, M. (1973). Job Market Signalling, *The Quarterly Journal of Economics*, [e-journal] vol. 87, no. 3, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 15 April 2022]

Statista Research Department. (2022). Number of IPOs in the United States from 1999 to 2021, Available online: https://www.statista.com/statistics/270290/number-of-ipos-in-the-us-since-1999/ [Accessed 10 May 2022]

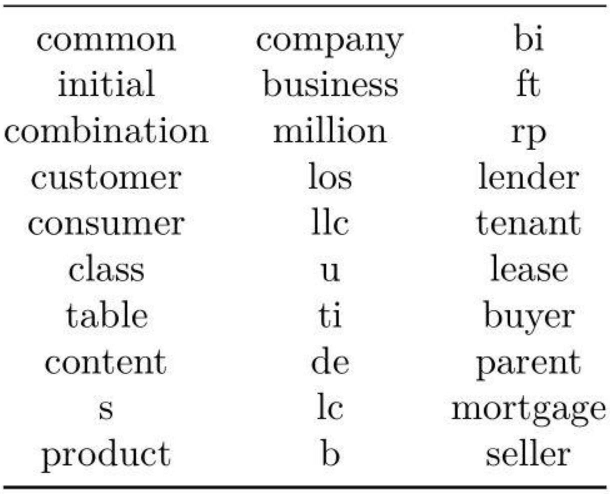
Tao, J., Deokar, A.V., & Deshmukh, A. (2018). Analysing Forward-looking Statements in Initial Public Offering Prospectuses: A text analytics approach, *Journal of Business Analytics*, [e-journal] vol. 1, no. 1, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 19 April 2022]

Wish. (2020). Form S-1 Registration Statement under the Securities Act of 1933 [pdf], Available at: https://ir.wish.com/static-files/d8aa340f-ba72-4b15-be3c-e401d0996b4d [Accessed 12 May 2022]

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An Up-to-date Comparison of State-of-the-art Classification Algorithms, *Expert Systems with Applications*, [e-journal] vol. 82, Available through: LUSEM Library website http://www.lusem.lu.se/library [Accessed 5 May 2022]

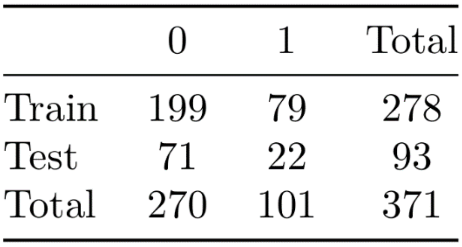
Appendix A

Table A1. Custom list of stop words.



common	company	bi
initial	business	ft
combination	million	rp
customer	los	lender
consumer	llc	tenant
class	u	lease
table	ti	buyer
content	de	parent
s	lc	mortgage
product	b	seller

Table A2. Split of training and test data for class 0 and 1.



	0	1	Total
Train	199	79	278
Test	71	22	93
Total	270	101	371

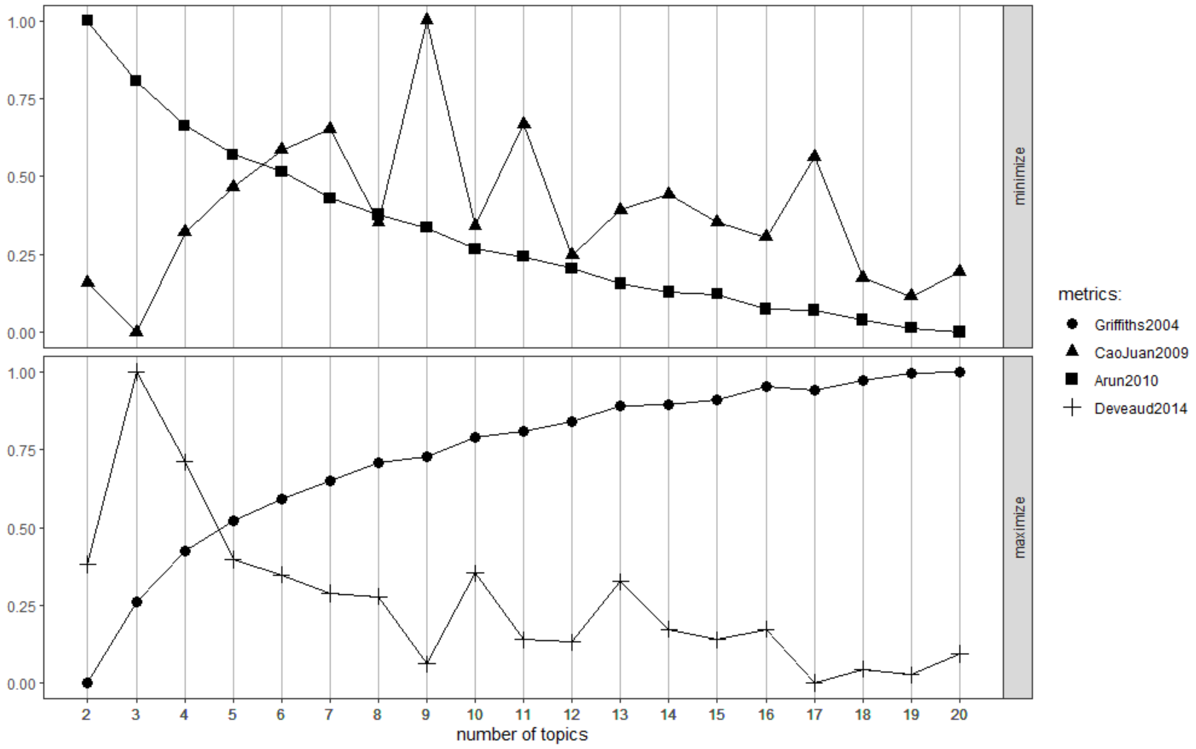
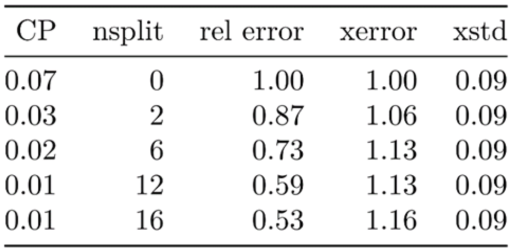


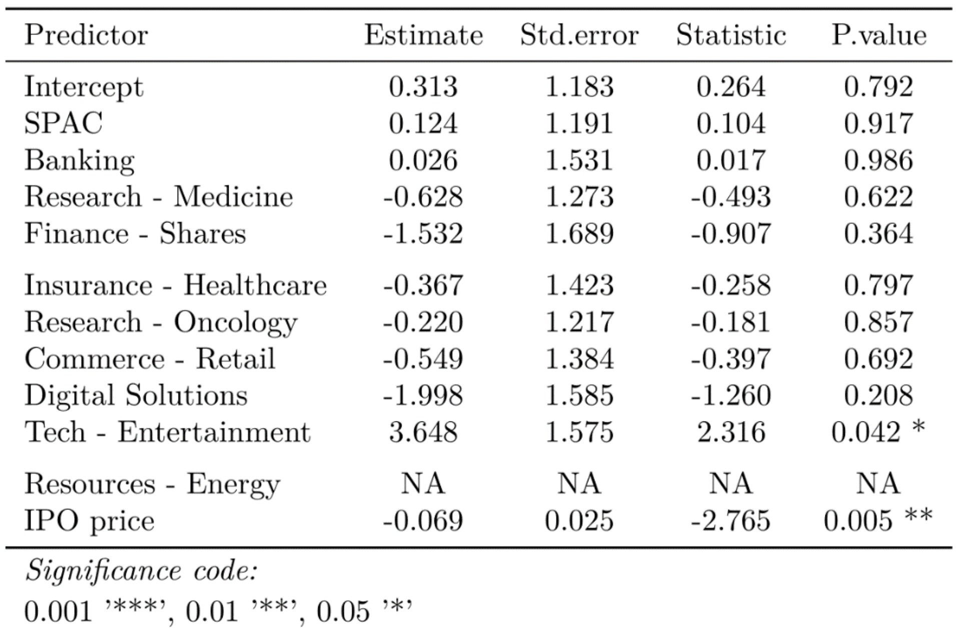
Figure A1. Result of metrics guiding choice of number of topics.

Table A3. Output of the tree model showing the cross validated error (xerror) for different tree sizes, splits, and complexity parameter (CP).



CP	nsplit	rel error	xerror	xstd
0.07	0	1.00	1.00	0.09
0.03	2	0.87	1.06	0.09
0.02	6	0.73	1.13	0.09
0.01	12	0.59	1.13	0.09
0.01	16	0.53	1.16	0.09

Table A4. Coefficient estimates from logistic regression of the predictors. Note: Due to perfect singularities, i.e., all topic proportions summing to 1, the last topic is not defined.



Predictor	Estimate	Std.error	Statistic	P.value
Intercept	0.313	1.183	0.264	0.792
SPAC	0.124	1.191	0.104	0.917
Banking	0.026	1.531	0.017	0.986
Research - Medicine	-0.628	1.273	-0.493	0.622
Finance - Shares	-1.532	1.689	-0.907	0.364
Insurance - Healthcare	-0.367	1.423	-0.258	0.797
Research - Oncology	-0.220	1.217	-0.181	0.857
Commerce - Retail	-0.549	1.384	-0.397	0.692
Digital Solutions	-1.998	1.585	-1.260	0.208
Tech - Entertainment	3.648	1.575	2.316	0.042 *
Resources - Energy	NA	NA	NA	NA
IPO price	-0.069	0.025	-2.765	0.005 **

Significance code:
0.001 ’***’, 0.01 ’**’, 0.05 ’*’

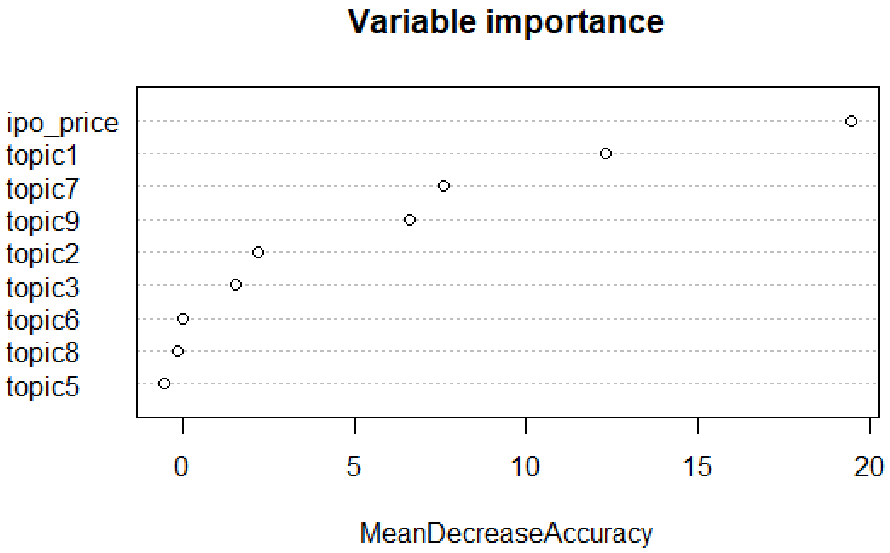


Figure A2. Variable importance of the 10 predictors as measured by the Mean Decrease Accuracy.