UCLouvain

LOUVAIN
School of Management

Louvain School of Management

The information value of the thematic content of IPO prospectus.

Author : BARRIO HERREZUELO Diego
Supervisor : THEWISSEN James
Academic year : 2021-2022
Dissertation for the master of Management (GEST2M)
Corporate Finance
Daytime schedule

.

Abstract

IPOs are known for being systematically underpriced. Prior research tried to link the underpricing to quantitative factors, the tone of words in IPO prospectuses, and how they are disclosed. However, underpricing still seems to be misunderstood. In this paper, we take a new approach by using a topic modeling method, Latent Dirichlet Allocation (LDA), to identify the thematic content of IPO prospectuses and see if the underpricing can be impacted by what is disclosed in these documents. We employ a sample of 1,155 IPO prospectuses that occurred between 1996 and 2021 with their corresponding financials and stock market data. We first show that the most elaborated topics in prospectuses are related to the shares and their characteristics. We show that the identified topics help to explain the underpricing, as it improves our model quality. In a second step, we discuss what topics matter to explain and increase the underpricing. We find that Trading Activities related topic has a positive and significant effect on underpricing. In contrast, Common Shares, Risk, Number of Shares, Strategic Alliances, and Result Statement related topics have negative and significant effects. Overall, we conclude that IPO prospectuses' thematic content helps explain the underpricing but generally decreases it as there is less asymmetric information. This paper shows that IPO prospectuses should be used more carefully by investors and managers as they contain valuable information which can predict underpricing and identify the potential return behind an IPO.

Keywords : IPOs; Prospectuses; Underpricing; Thematic content; Topic Modeling.

First, I would like to thank my supervisor, James Thewissen, professor of Finance at UCLouvain, for his guidance and advice throughout the realisation of this article. He managed to provide me with answers to my questions despite his busy schedule. I would also like to express my gratitude to him for providing me with the data I needed.

I also wish to warmly thank my family, who supported me during my studies and research. Nothing would have been possible without them.

Finally, I am grateful to all my friends and student associations who helped me develop as a person and who made these five years at UCLouvain unforgettable.

Contents

1 Introduction

Initial Public Offerings (called "IPOs") enable managers to sell shares for the first time on a stock market. This phenomenon is an important trend in the business world as it represents one of the critical steps in the development of a company (Daily et al., 2003). IPOs have been occurring for several generations and represent a source of interest amongst managers and external investors. Indeed, existing managers can increase the liquidity of their firm and diversify their portfolios by selling their shares. The risk for a manager of facing a refusal by the existing owners is reduced through an IPO, as the access to funding is easier (Beck, 2017). Most of the time, IPOs appear to be underpriced on their first trading day.

This study looks at whether or not the thematic content of IPO prospectuses based on a machine learning method can serve to identify the IPO underpricing. Prior research looked at theories based on asymmetric information and behavioural finance. These researches were based on quantitative approaches, which tried to detect factors that could have predictive effects. Further research tried to investigate the use of tone and how the information is disclosed to explain the IPO underpricing, using more qualitative data and methods. This type of research was focused on discourse value and the use of words employed in prospectuses (Wales & Mousa, 2016). The goal was to understand how the information is disclosed and linked with the underpricing. Despite the considerable growth and all the research that was made in the field, the underpricing still appears to be misunderstood. Indeed, limited evidence in prior literature investigates the linguistic content of the prospectus. Here, we will answer that gap by looking at the thematic content of IPO prospectuses. An emerging research trend has appeared in the last years, focusing on content analysis and the central importance of qualitative data in documents. This approach differs from the previous literature because it applies a machine learning method to investigate and explain the underpricing instead of looking simply at linguistic features.

In this paper, we identify and classify IPO prospectuses using topic modeling. Our reason for addressing this phenomenon is to fill the gap in the literature regarding the explanation of the underpricing, understand how the topics are built, and provide an additional piece of evidence to the thematic content literature of financial documents. Through this study, we want to understand the topics' impact on the company. In other words, it is to determine whether or not the classification of the content can explain the underpricing by increasing or decreasing it and how it impacts the value of a company and its performance. This article aims to identify the topics that matter the most to increase the underpricing and provide valuable recommendations to investors by bringing them with the best information to value an IPO truly, and from which topics they could benefit.

This research is about textual analysis, searching for information disclosed in publicly available IPO prospectuses, and seeing if there is some information that is not conveyed but embedded (Zou et al., 2020). Here, the prospectus thematic content is characterized as a distribution of underlying topics in a document and is measured by the number of words related to it. We use a sample of American IPO prospectuses, which occurred between 1996 and 2021, as they are regulated by the Securities and Exchange Commission but also because they remain less standardized compared to other kinds of prospectuses, such as the Chinese one, where the whole document is being standardized.

We first provide a literature review focusing on the determinants of underpricing based on asymmetric information and behavioral theory. This review also includes research made about the textual analysis of IPO prospectuses. Once the literature is properly reviewed, we establish research questions. Using the LDA topic-modeling method, we build several topics through our prospectuses by identifying and classifying our different content. Subsequently, we provide answers to our research questions by providing an empirical analysis. In this analysis, we show that the content of IPO prospectuses is tight and presents similarities. Furthermore, we show that the thematic content has information that explains the underpricing. We also find that the thematic content helps, in general, to reduce the underpricing as it decreases asymmetric information. However, we identify the Trading Activities-related topic as having an increasing effect on the underpricing and as a factor of successful IPOs in terms of return and performance.

2 Literature Review

In this literature review, the main topic is to address the causes of underpricing. The basic concept of underpricing is tackled, as well as its origin, its different components, and the impact of the words used on its value. In this order, this review aims at answering the question: *What are the determinants of the underpricing, and how the tone disclosure impacts it?*

2.1 Initial Public Offerings, underwriters and underpricing

IPOs are a business trend that is considered a puzzle, which would explain why most IPOs tend to fail in the market (Kumar & Kumar Singh, 2019). A part of this puzzle identified is that most IPOs face underpricing on the first trading day. The underpricing is a market anomaly represented by the offer price of a newly issued security being lower than the close price of the same security on the first trading day (Injai, 2018). It also can be defined as the money left on the table (Loughran & Ritter,

2002). In other words, by underpricing, the issuing firm leaves a substantial amount on the table, which is expressed as the number of shares times the difference between the first-day market and the offer price. On the opposite, the underpricing can also be considered as a measure of performance. Indeed, the return of an IPO could be measured using the underpricing on the first-trading day since the price increases (Zou et al., 2020).

The issuing firm taking the IPO has to pass through an intermediary called the underwriter, often an investment bank, to succeed in the process. The underwriter plays a crucial role in long and complex financial transactions such as Mergers & Acquisitions, but most of the time deals with IPOs. The motivation between the issuing firm and the underwriter should be mutual, as they can benefit from it. Underwriters use the book-building method to value newly issued stocks (Injai, 2018). This method presents the new issue to potential buyers and then asks them their willingness to pay and their desired quantity. According to a study on the causes of underpricing, the latter can be explained and summarized by two schools of theories: Asymmetric information theory and Behavioral theory (Zou et al., 2019)

2.2 Asymmetric Information Theory

There are three components regarding the Initial Public Offerings (Beck, 2017). There are the underwriter, the issuing firm, and the investors. The reason why there would be underpricing could be due to information asymmetry between these parties. Information asymmetry is when both parties during the IPO issuance have different kinds and levels of information. Information has then a key role in underpricing (Injai, 2018). The underwriter would only serve as a distributor in a perfect world, but there is asymmetry. As compensation, investment banks are rewarded because they know more than the firm. It means that the higher the asymmetry, the lower the first-day offer price and the higher the underpricing.

In this case, the signaling theory plays a significant role in determining underpricing. Here, the underpricing would be volunteered and made to signal the firm's prospects (Beck, 2017). In other words, companies that think they know the best information about their future and positive prospects would lower the offer price of a share on the first trading day to attract investors, as they would make higher returns after. Their study on prospectus information mentions several factors that signal to investment bankers as CEO-retained equity, board prestige, venture capital equity, or firm size and age. However, in their study, they found these factors to be insignificant. They also tested if the underpricing could be more present in the tech industry, but the result was insignificant, as this

industry appears as mature. However, they proposed further research regarding the underpricing for less mature industries.

To continue, the underwriter's reputation, which corroborates a study on Indian SMEs (Arora & Singh, 2019), was also identified as a valuable signal. It highlights the signaling role of the reputation as it states a negative relationship between the reputation and the underpricing, defined as the initial return. Hence, the underwriter's prestige would help to reduce information asymmetry and would signal quality. Good firms only would receive interest and funding from these third-parties. Moreover, unique underwriter content would increase the pricing accuracy given that they have a great experience, reputation, and market share (Hanley & Hoberg, 2010). On the opposite, another study by Kumar Singh and Kumar (2019) shows a positive relation between underwriter's quality and underpricing. The explanation would be that the underpricing is seen as compensation for the high-ranked underwriters. The issuing firm prefers the analysis from a prestigious underwriter over any concern about underpricing (Loughran & Ritter, 2004).

These are signals disclosed by the firm, but another study states that the environment can also determine underpricing (Park & Patel, 2015). Indeed, it defines ambiguity as a critical factor for underpricing. It differs from the uncertainty because ambiguity can not be resolved with time, even if a part of the prospectus is standardized. According to the SEC regulation,*"there is discretion in the prospectus on how the firm is presented."* (Park & Patel, 2015). Ambiguity is soft informational content that can be interpreted differently by different observers. This research found significant that ambiguity is positively correlated with underpricing but that the effect is moderated when senders are strategically deviating from the norm, when the valuation within an industry is heterogeneous, and when firms are medium-sized. Hence, it shows how the signaling environment can affect a recipient and the outcome.

Another theory explains the information asymmetry of underpricing in another way. It is called the winner's curse or the *"Rock Model"*. In this model, asymmetry occurs not between the underwriter and the issuing firm but between the firm and investors. This model separates investors between informed and uninformed investors (Kumar & Kumar Singh, 2019). The informed investors can identify the high-quality security issues, unlike the uninformed ones, which cannot identify the quality within a mix of securities. The underpricing is then made to enable uninformed investors to enter the market as they have better chances of earning a higher rate of return.

2.3 Behavioural Finance Theory

When it applies to IPOs, the behavioural finance theory is when the investor's sentiment and over-optimism can influence the offer price (Zou et al., 2019). Investment bankers can deliberately set the underpricing (Daily et al., 2003). A factor influencing the underpricing can be the type of contract or agreement used in the IPO. To launch an Initial Public Offering, two types of contracts exist : firm commitment agreement and best effort agreement. In the first one, the underwriter agrees to pay a fixed amount for all the securities to the issuing firm before the IPO (Injai, 2018). The best effort agreement is when the underwriter tries to sell IPO shares at best, with a required minimum, to raise capital. The difference between both agreements is the risk repartition. In a firm commitment agreement, the underwriter bears the entire risk, while there is no risk for the underwriter in the best effort. Underpricing could be determined because most IPOs are conducted through a firm commitment agreement and then would be deliberate (Daily et al., 2003). Since the underwriter bears the full risk, and if the offer price is too high, he has an incentive to lower the latter to pursue the sell. Another motivation for investment bankers to deliberately underprice the first-day IPO share is to make a greater return for their clients and, as a consequence, to encourage them to pursue business for the next time.

Moreover, Zou et al. (2019) identified in their study a second type of underpricing, called mis-valuation. It happens when the closing price is higher than the intrinsic value and is driven by biased investors. It operates on the secondary market, while deliberate underpricing occurs on the primary market. The question addressed in their study was to see which behavioral bias would most explain the IPO underpricing. They found that the main cause of underpricing is the mis-valuation. The underpricing would then occur more on the secondary market. This puts in perspective the effects of deliberate underpricing and the type of contracts, as they would not have a sufficient impact to explain why underpricing occurs.

2.4 Textual Analysis in IPO Prospectuses

Suppose we follow the underlying logic behind the existing literature. In that case, the more information between underwriters, issuing firms, and external investors, the less there is ex-ante uncertainty and the less there is underpricing. The thematic content and the textual analysis could impact how the information is disclosed and perceived and how it affects the investors' behaviors and underpricing. The media tone of newspapers and financial documents has been shown to affect how investors can perceive and be impacted by the information. Underpricing could then be influenced

by positive and negative words in financial documents, including prospectuses. It was found that the degree of underpricing would decrease with an increase in the percentage of positive and negative words in the documents, with a more significant impact from the negative media tone (Zou et al., 2020). It supports the view that information in documents helps reduce underpricing. However, it is interesting to see the effect of negative media tone, which has the biggest impact on investors' behaviors. Furthermore, the underpricing has also been linked to how the words are expressed in the IPO prospectuses (Wales & Moussa, 2016; Paulus et al., 2021). Investors' minds could be influenced by how the discourse is used in prospectuses. They identify the affective discourse, which uses emotions and the cognitive discourse, which is the *"use of language that reflects the process of understanding through the application of thought and consideration"* (Wales & Moussa, 2016). The firm should be careful of what discourse it will employ in its disclosure. The affective rhetoric can signal a way for the firm to make investors consider the IPO positively, which might hide other information and cause underpricing. More considerable publicity before the offer leads to more prominent investor sentiment. On the opposite, cognitive discourse is perceived as a sign of value and could reduce the underpricing. It is interesting to notice that research has shown that the quantity of positive words (in %) in the prospectus would reduce the underpricing, contrary to the affective discourse's statement of increasing the IPO initial return.

The IPO prospectuses can be divided into two parts regarding their disclosure of information (Hanley & Hoberg, 2010). There are standardised and informative components. The latter category is considered more important for the pricing decision as investors rely on specific categories in the document, like the risk factor section (Paulus et al., 2021). Indeed, the study has shown that when there is more standardised content, book-building is longer and more expensive for the underwriter, which is linked with the difficulty of pricing the stocks accurately. Furthermore, regarding the similarity of the disclosure in prospectuses, research has shown that if the information is similar to the previous registration statement, there will be more standardised content. As a consequence, there will be less information in the document, thus increasing the asymmetry and the underpricing. Underpricing is then impacted by the repetition and similarity of the document's content. It shows how this content and the qualitative data are important and relevant to analyse the IPO underpricing in prospectuses. Hanley and Hoberg (2010) also examined the most important content related to pricing. They found that the inputs into valuation models (accounting, product market, and corporate strategy) are the most important as they observed lower price changes.

3 Research Questions

We have seen in the literature that the underpricing can be determined in several ways regarding the theories. It depends mainly on the information between the issuing firm, the underwriter, and the external investors. On the other hand, the underpricing is also impacted by the words and the tone used to disclose information. We state that it is not only the quantity of information that prevails, as discussed in the literature review, but also the type of information. Indeed, there is specific information that investors will look for because not all sections on prospectuses are informative, making them standardized. Various researches focused on how the disclosures are communicated, but limited evidence looked at what is being disclosed. Some topic modeling-based research method has started to focus on financial topics. We will use this method to help us identify what content is being disclosed through the IPO prospectuses and to see if a topic can help us explain the underpricing. The method used is called Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003). It is an unstructured and unsupervised method that employs a Bayesian topic-modeling algorithm (Brown et al., 2019). We will be able to identify and classify the content of many IPO prospectuses, which otherwise would be nearly impossible to do. Through our research questions, we propose that the content of IPO prospectuses and the underpricing rely not only on the quantity of information or the way it is disclosed but also on the several types of content disclosed. A study by Brown et al. (2019) identified a link between financial misreporting detection, earnings quality, and the use of topic modeling, which proves the interest in applying a machine learning method. We will extend that type of research to ours to see if thematic content using our IPO prospectuses topic modeling method helps explain the underpricing. We expect that the thematic content will help us explain underpricing, identify successful IPOs and list the topics that are the most interesting for investors. In order to follow our contribution, we will address the following research questions :

- Research question 1: What are the topics elaborated in the IPO prospectuses?

- Research question 2: Does the thematic content help explain the IPO underpricing?

- Research question 3: What topics matter to explain the underpricing?

4 Data Selection and Sample

This section presents the data we used to further perform the LDA method and build our data sample.

4.1 Research question 1

Since we focused on IPOs made by U.S. companies, and since they are regulated by the SEC, we used a bunch of 1,155 IPO prospectuses, the S-1 Forms, that occurred between 1996 and 2021. The S-1 Forms are the registration documents that have to be filled by companies if they want to be listed on a national exchange. Choosing these texts made it easier to identify the companies they describe and find data about their financials, and model our sample. Indeed, with the CIK, the number given to an individual company by the SEC, we could retrieve the financial information corresponding to the companies in our documents.

4.2 Research questions 2 & 3

Since we identified the firms present in our sample of texts, we were provided with data from 1990 to 2021 from the CRSP and Compustat platforms, which provide financial and stock market information for the companies in the prospectuses. By cleaning and merging our data, we modelled a sample gathering our topics and the data from the CRSP and Compustat platforms.

5 Methods and Variables Description

This part describes the methodology we used to get the prospectuses' topics and do the regressions. We also describe the variables chosen to perform the regression tests.

5.1 LDA Topic Modeling

To determine the different topics elaborated in the IPO prospectuses, we used topic modeling in our bunch of 1,155 IPO prospectuses. Here, this analysis aims to clean the documents and then determine the thematic content of the IPO prospectuses. We will get an output of several lists of words that will give us information about a particular topic that we could name afterwards. The Latent Dirichlet

Allocation (LDA) is the method used, which is what we call an unsupervised method. Indeed, we use machine learning to determine the topics elaborated in our documents. However, this machine learning depends on various assumptions and decisions that have to be made to extract categories from our documents.

For our output to be the more precise as possible, we had to process to a document cleaning method. This cleaning is set to make documents more "standardized" and comparable to each other. We first created a corpus that is a collection of documents on which we can apply text mining and language routines like topic modeling and on which we could apply the cleaning. First, we removed the punctuation, the numbers, and the special characters (e.g. "/" or "&"). We also transformed all letters into lower cases. To continue, we also performed language manipulation like lemmatization which consists of replacing all the different forms of a word with a particular term. This avoids redundant and non-useful information. Subsequently, we removed the English stop words, which are words that appear frequently, but that do not give any additional information. Once we did all that, we stripped the white spaces to get clear and coherent documents.

After the cleaning, we created a matrix of our documents on which we specified our different assumptions and decisions to get our output. Here, we worked by induction, setting factors that we increased or decreased after checking the results. We established 3 factors on which our output depends. There is first the frequency, which refers to the first x words of our sample. For example, if we choose for example, a frequency of 5,000, and we define our matrix on this value, the topic modeling will only use the 5,000 most used words. There also are the number of topics which categorize the documents and the number of words related to each topic. From these tests, we chose 9 topics on which we listed 15 words with no frequency. Indeed, reducing or increasing these parameters was either non-informative or redundant. The low number of topics appeared to be repetitive, and a large number of topics appeared to be non-useful regarding the quality of the information.

We also removed words that appeared in our results because of their frequency, but that gave no additional information regarding the topics. This list contains the terms: "table", "ii", "iii", "us", "may", "will", "much", "content", "upon", "can", "two", "include", "holder", "hold", "complete", "will", "article", "per", "require", "act", and "prior". After removing the words present in less than five documents, we obtained a list of 17,864 different words. Once all the assumptions were set and the LDA method performed, we got our set of topics and words.

5.2 Regressions

5.2.1 Models

In this paper, we established two equations to test whether the thematic content could impact the underpricing but also to analyse which topics matter to explain it. These models allow us to test the effect of topics on underpricing. Our first model includes the effect of topics (equation 1), while the second model (equation 2) without topics serves as a test regression and a point of comparison. We used 8 independent variables and 3 control variables in equation 1, while in equation 2 there are just the 3 control variables from equation 1. Our models are expressed below :

$$(1)\ y = \sum_{k=1}^{k} Topics.k + undprcControls + \epsilon$$

$$(2)\ y = undprcControls + \epsilon$$

where $Topics.k$ are the topics generated from the LDA with k the topic number, $undrpControls$ are the control variables, and ϵ is the expression of the standard errors corrected for heteroscedasticity.

5.2.2 Variables

a) Dependent variable

undprc: The underpricing is our dependent variable on which we want to see how it is impacted. As we saw in the literature, the underpricing can be expressed as a performance index as it could measure the return of an IPO on the first trading day (Zou et al., 2020). So we expressed the underpricing using a return formula: (P1-P0)/P0 with P1 and P0, which are close and offer prices, respectively. To compute it, we chose the variable Bid (as an estimator of the price asked at the end of the first trading day) and the Open Price in our data sample. Our formula is now expressed as (BID – Open Price)/Open Price. It means that a negative underpricing is when the open price is higher than the close (bid) price, and, on the inverse, a positive underpricing means that the close price (bid) is higher than the open price.

b) Independent variables

We chose the topics obtained from the LDA topic modeling method as independent variables for our model since they are our variables of interest. As we did some previous tests with our regressions, it appeared that the 9th topic (Real Estate) presented a coefficient expressed in the "NA" form,

which is a sign of multicollinearity. We decided to remove it from our model to have the most precise analysis possible, and we kept as our variable of interest the first 8 topics.

c) Control variables

factor(sic): This variable indicates the industries of the companies of our final data sample. Beck (2017) shows that underpricing could depend on the industry's maturity degree. The effect of the industries may explain why underpricing occurs. This variable is expressed as a factor, as there are several industries, and they cannot take the form of a single value.

factor(fyear): This variable indicates the year when the IPO occurred. IPOs are considered a puzzle. One of the anomalies is that IPOs tend to occur in waves, particularly when companies look to raise capital when there are existing growth opportunities (Injai, 2018). Time may impact the underpricing. Since IPOs often happen together, mis-valuations could also increase in time of growth and have a bigger impact on underpricing. This variable is expressed as a factor, as there are several years, and they can't take the form of a single value.

at: This variable represents the total assets of a company and is chosen as a measure of a firm's size since smaller firms are considered riskier (Beck, 2017). Indeed, there is more uncertainty about smaller firms' prospects, so higher uncertainty and higher underpricing.

6 Empirical Results

This section shows the analysis of the results obtained from LDA and regressions and how they answer the research questions.

6.1 Research Question 1: Topics elaborated in IPO prospectus

Results from LDA are presented in Table 5 in the Appendix. Topics head the columns with their corresponding list of words below. We labelled the topics using the information given by the lists of words. The goal is to get a more coherent meaning behind all the words. We decided to name the topics as follows:

-**Topic 1 - Common Shares**: This topic is related to the unique presence of common shares in the IPO.

-**Topic 2 - Class of Shares**: Since we noticed the terms "share", "combination", "warrant" and "ordinary", we stated that these terms tend to specify this topic as the presence of a hybrid structure of shares in prospectuses. Different types of shares are then offered (ordinary shares, derivatives).

-**Topic 3 - Risk**: Terms like "loss" and "insurance" refer to the potential risk, and its mitigation explained in the prospectuses.

-**Topic 4 - Acquisitions**: This topic refers to the presence of acquisitions made by the companies in the prospectuses.

-**Topic 5 - Number of shares**: This topic refers to the number of shares issued during the IPO.

-**Topic 6 - Trading activities**: This topic gathers words related to trade, commodities, and markets. Prospectuses would refer to buying and/or selling goods, price hedging, and protection activities.

-**Topic 7 - Strategic Alliances**: This topic gathers words related to partnerships and collaborations. It refers to the presence in prospectuses of strategic agreements with independent companies.

-**Topic 8 - Result Statement**: This topic refers to the amount of result information that is disclosed at the end of the exercise.

-**Topic 9 - Real Estate**: This topic refers to the presence in the prospectuses of Real Estate investments and properties of the companies.

If we look at table 5, we can see that some categories appear to be repetitive, but diversity occurs based on some words. We state that it is due to the high frequency of first words and the inherent nature of the prospectus itself, which remains relatively similar in its categories and is tight in the contents disclosed. We could state that since IPOs are regulated and focused on a company's precise point, then the diversity of content is low. Once we identified the topics elaborated in the prospectuses, we looked at their descriptive statistics and respective means to see which topics have the biggest proportions and are the most elaborated in prospectuses. In table 1 below, we can see the results.

Table 1: Descriptive Statistics of Topics

Statistic	N	Mean	St. Dev.	Min	Max
Common Shares	1,155	0.206	0.222	0.00000	0.548
Class of Shares	1,155	0.100	0.177	0.00000	0.492
Risk	1,155	0.064	0.180	0.00000	0.984
Acquisitions	1,155	0.088	0.096	0.00000	0.352
Number of Shares	1,155	0.254	0.203	0.00000	0.503
Trading Activities	1,155	0.062	0.219	0.00000	0.990
Strategic Alliances	1,155	0.038	0.126	0.00000	0.860
Result Statement	1,155	0.124	0.260	0.00000	0.981
Real Estate	1,155	0.064	0.189	0.00000	0.886

Let us take the several means in the table. We notice that 4 topics are substantially present in the prospectuses, which are the Number of Shares (25.3%), Common Shares (20.6%), Result Statement (12,4%), and Class of Shares (10%) related topics. The most elaborated topic in prospectuses refers to the number of shares issued in the IPO. The other topics are less present regarding their proportions. Acquisitions (8,8%), Risk (6.4%), Real Estate (6,4%), Trading Activities (6,2%), Strategic Alliances (3,8%) related topics show lower means that vary between 3% and 8%. The topic which is less elaborated on in the prospectuses is Strategic Alliances.

Interestingly, this analysis shows that the most important topics are pretty interrelated if we refer to their proportions. Indeed, it is straightforward to state that topics like Class of Shares, Number of Shares, or Common Shares refer to common terms. It also shows that IPO Prospectuses are mainly focused on the issuance of shares and their characteristics in general. It comforts us with the idea that IPO prospectuses are relatively tight in their disclosed content.

6.2 Research question 2: The impact of the thematic content

6.2.1 Descriptive statistics and correlations

Before running our regression, we looked at the descriptive statistics of our regression model's dependent and control variables in Table 2. Since we used the descriptive statistics of topics, our independent variables, to answer research question 1, they are represented in Table 1.

These statistics include the number of observations, the mean, the standard deviation, and the minimum and maximum. We observe 3,923 observations for the underpricing, which is lower than the control variable. This is because these values were not available for all companies identified using their CIK. We also see that the average underpricing of our sample is -1.7% which means that, on

Table 2: Descriptive Statistics of dependent and control variables

Statistic	N	Mean	St. Dev.	Min	Max
Dependent Variable					
undprc	3,923	−0.017	0.101	−0.527	0.645
Control Variable					
at	4,298	38,859.220	129,639.200	0.001	1,060,505.000

average, the close price of our sample is lower than the open price on the first trading day. If we also look at the difference between the maximum and the minimum of our sample, the underpricing varies in a range of 112%. Regarding the control variables, we omitted *sic* and *fyear* in this table as it was not relevant since they represent different years and industries. We also looked at the correlations we obtained after doing a Spearman analysis in which we showed how underpricing is correlated with our 9 topics. The results are presented in Tables 6 and 7 in the Appendix. We can see in the last column of Table 6 the correlations of the topics on the underpricing. First, if we look at the p-values in Table 7, we can see that all correlations tend to be significant. We observe in Table 6 that there are positive and negative correlations between topics and underpricing. A positive correlation means that the underpricing would tend to increase with a particular topic. On the contrary, a negative correlation means that the underpricing would tend to decrease with a particular topic. These statistics would be the first indicator that underpricing could be linked to the thematic content. The presence of specific topics could increase or decrease the underpricing according to those tables, and the sign could vary.

6.2.2 Regression Results

In the following Table 3, we observe the results from our two regressions lines, the models with and without topics. This analysis aims to show that by adding the topics, our model improves. If we observe Table 3, we see that by incorporating the different topics into our model, the latter improves. Indeed, if we look at the adjusted R-Square, we see it going from 44,4% in the model without topics to 56% in the model with topics. By comparing the two adjusted R-Squares, we see that incorporating topics increases the adjusted R-Square by 26,12%, which is quite substantial. So adding topics to our regression model would impact the dependent variable, which is the underpricing. The thematic content would then matter to explain if the underpricing occurs or disappears depending on the coefficient sign. On the other hand, we also have to perform a linear restrictions test, which allows us to test if a group of variables is jointly significant. This will inform us if there is a significant difference between the two models and if it is relevant to add our topics to the regression. Our models are significantly different if the p-value of our test is lower than 0,05. These results are in Table 4.

Table 3: Regressions results

	Dependent variable:	
	undprc	
	(without topics)	(with topics)
Common Shares		-0.176^{***}
		(0.033)
Class of shares		0.114
		(0.089)
Risk		-0.131^{***}
		(0.009)
Acquisitions		0.068
		(0.048)
Number of shares		-0.305^{***}
		(0.056)
Trading Activities		0.290^{***}
		(0.058)
Strategic Alliances		-0.354^{***}
		(0.020)
Result Statement		-0.070^{***}
		(0.015)
fixed.sic.effect	YES	YES
fixed.year.effect	YES	YES
at	0.00000^{***}	0.00000^{***}
	(0.000)	(0.000)
Constant	0.118^{***}	0.331^{***}
	(0.040)	(0.043)
Observations	3,870	3,870
R^2	0.453	0.569
Adjusted R^2	0.444	0.560
Residual Std. Error	0.075 (df = 3803)	0.067 (df = 3795)
F Statistic	47.779^{***} (df = 66; 3803)	67.582^{***} (df = 74; 3795)
Note:		*p<0.1; **p<0.05; ***p<0.01

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
without topics	3803	21.65				
with topics	3795	17.09	8	4.56	126.72	0.0000

Table 4: Linear Restrictions Test

We see that the F-Stat of our linear restriction test is 126.72 with a p-value of 0.0000, which is highly significant. This test gives a piece of evidence that our model fit. After assessing the effect of the topics on the Adjusted R-Square and the significant difference between our models, we can conclude that this analysis answers our second research question positively. The thematic content of the IPO prospectus is useful and informative regarding the IPO underpricing. In other words, prospectuses contain information that affects the share price on the first trading day. Hence, these texts are helpful to use and analyse in order to see and predict how a share price can evolve on the IPO date. This research question aimed at showing the general effect of the thematic content. If we want to identify the relevant topics or, in other words, that significantly increase the underpricing, we have to look more precisely at each specific topic.

6.3 Research question 3: The matter of topics to explain the underpricing

We showed that the thematic content could explain the underpricing. We then analysed the results of our regression model deeply to see which topics matter to explain the underpricing. This analysis aims to identify topics that increase significantly underpricing and how investors can benefit from it. To answer this research question, we used the regression with topics from Table 3. With this model, we identified 3 different categories: topics whose coefficients are non-significant and that are worthless to analyse, topics with positive and significant coefficients and topics with negative and significant coefficients.

When we look at Table 3, we see that not many topics matter to explain the underpricing. More precisely, only one subject increases the underpricing significantly, which is the Trading Activities topic with a coefficient of 29% at a 99% level. Hence, except for the Class of Shares and the Acquisitions related topics which are non-significant, all topics decrease significantly the underpricing by showing a negative coefficient. We can highlight the effect of the Strategic Alliances related topic as having the biggest negative impact with a coefficient of -35.4%, which is significant at a 99% level. Underpricing tends to be higher when there is the presence of Trading Activities related terms in the IPO prospectuses and tend to be lower with the presence of Risk, Number of Shares, Strategic Alliances, and Result Statement related terms. This result is in line with the asymmetric information theory (Injai, 2018), as the majority of the topics reduce underpricing. Indeed, information is key

to underpricing. The more content disclosed the less uncertainty and asymmetry, leading to a more efficient allocation and less underpricing.

As we said, the Trading Activities topic increases significantly the underpricing. This means that companies that, in their business model, are active in trading activities and express it in prospectuses would tend to obtain a close price superior to the offer price on the IPO date. Investors could then use this information disclosed in the prospectuses and benefit from it. Underpricing is a complex concept, but it is not necessarily negative even if, as we saw in a study by Loughran and Ritter (2002), it is also called the money left on the table. However, it does not apply to investors but to issuing firms that will lose money because the open price could have been higher but does not reach its efficient potential. As we saw in the study of Zou et al. (2020), underpricing is also expressed as a performance index and can measure the return of an IPO on its first day. Then, as the presence of trading activities' terms increases the underpricing, investors investing in companies that are active in trading on their business model and which disclose this content could benefit from positive returns. Using signaling theory, we can also extend this reasoning to issuing firms. Besides the potential loss, firms that think they best know about their future could voluntarily disclose more Trading Activities content to voluntarily underprice their shares to attract more investors and make the IPO more successful.

7 Limits

This research showed how to create categories of topics in IPO prospectuses. We succeeded in showing the link of the thematic content with the underpricing. We have to remind that, to build our topics, we used an unsupervised machine learning method that depends on several assumptions and decisions which are subjective and different from one to another (frequency, number of topics, etc.). The output can evolve easily depending on who is performing the research. The use of machine learning is evolving at a great pace in finance. We might encounter conclusions that will be more accurate and precise in the future. The fact we failed to show the effect of the Real Estate topic on underpricing is also a limitation in this work.

Since we also showed that the contents disclosed in these prospectuses are tight and similar, we think that we identified fewer topics in the prospectuses compared to the variety of subjects that exist, which could affect the underpricing. We then recommend further research to assess the link between underpricing and thematic content in other financial documents, which disclosed a wider variety of content. This paper can also be the basis of further research on how to efficiently combine

the proportion of topics in prospectuses to avoid potential losses and benefit from returns generated by the underpricing.

8 Conclusion

IPOs are complex and can be characterized by different anomalies. One characteristic of the IPOs is that they tend to be often underpriced on the first trading day. For decades, research has tried to explain why underpricing happens. The textual analysis in financial documents has also started to be used by showing how the tone and the use of words could impact the stock price on the IPO date. In this research, we take a different approach which is more recent in the field of finance, by using the LDA topic modeling and by showing how the content in IPO prospectuses impacts the underpricing. It differs from previous literature as it focuses on what is disclosed and not how it is disclosed.

The first conclusion of this work is that IPO prospectuses are similar in some categories they disclose, as the most elaborated topics are related to the share in itself and its characteristics. We then succeeded in showing that the thematic content has informational value when testing its impact on underpricing. Regarding the specific topics, we found that Trading Activities related topic increases the underpricing while the latter decreases with Risk, Number of Shares, Strategic Alliances, and Result Statement related topics. Subsequently, we established that trading aspects of prospectuses are linked to successful IPOs because they increase the underpricing as well as the initial return. Hence, it is interesting for investors to invest in companies that disclose this type of content because there is a potential for greater returns.

In general, the thematic content decreases the underpricing as more information is disclosed, so there is less uncertainty, less information asymmetry, and less underpricing. It can also be used to identify successful IPOs if we express this success in terms of yield and return. As managerial implications, this paper shows that companies and investors should pay attention to the type of information disclosed in IPO prospectuses. Indeed, it can have a predictive effect on a firm's future potential loss, which is interesting for the managers but also for external investors as they could be able to identify the potential return behind an IPO.

9 References

Arora, N., & Singh, B. (2019). Impact of Auditor and Underwriter Reputation on Underpricing of SME IPOs in India. *Management and Labour Studies, 44(2)*, 193-208.
https://doi.org/10.1177/0258042X19829285

Beck, J. (2017). Determinants of IPO Underpricing : Tech vs Non-Tech Industries. *Major Themes in Economics, 19(5)*, 39-55.
https://scholarworks.uni.edu/mtie/vol19/iss1/5

Blei, D. M., Ng, A. W., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3,*993-1022.
https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Hanley, K. W., & Hoberg, G. (2010). The Information Content of IPO Prospectuses. *The Review of Financial Studies, 23(7)*, 2821-2864.
https://www.jstor.org/stable/40782968

Historic Year Sees Highest Global IPO Activity in a Decade with Surge in Domestic Listings and Continued Growth Predicted for 2021. (2020, 16 décembre). Baker McKenzie. Consulted the 25 november 2021, at the adress
https://www.bakermckenzie.com/en/newsroom/2020/12/ipo-report-2020

Injai, E. B. (2018). *Comparison of IPO underpricing between Financial Institutions and Non-Financial Institutions* (Thesis). Universidade Do Porto.
https://repositorio- aberto.up.pt/bitstream/10216/117417/2/302647.pdf

Kumar Singh, A., & Kumar, M. (2019). IPO's Underpricing : Trends, Determinants and Role of Underwriters. *Asia-Pacific Journal of Management Research and Innovation, 14((3–4))*, 81-93.
https://doi.org/10.1177/2319510X18817450

Loughran, T., & Ritter, J. R. (2002). Why Don't Issuers Get Upset about Leaving Money on the Table in IPOs ? *The Review of Financial Studies.*
https://www.jstor.org/stable/2696783

Loughran, T., & Ritter, J. R. (2004). Why has IPO underpricing changed over time ? *Financial Management, 33(3),* 5-37.
https://doi.org/10.2139/ssrn.331780

Nerissa, C. B., Crowley, R. M., & Elliott, W. B. (2020). What Are You Saying ? Using topic to Detect Financial Misreporting. *Journal of Accounting Research, 58(1)*, 237-291. https://doi.org/10.1111/1475-679X.12294

Park, H. D., & Patel, P. C. (2015). How Does Ambiguity Influence IPO Underpricing ? The Role of the Signalling Environment. *Journal of Management Studies, 52(6)*, 796-818. https://doi.org/10.1111/joms.12132

Paulus, N. M., Koelbl, M., & Schaefers, W. (2021). Can textual analysis solve the underpricing puzzle ? A US REIT study. *Journal of Property Investment Finance.* Published. https://doi.org/10.1108/JPIF-06-2021-0052

Wales, W., & Mousa, F. T. (2016). Examining Affective and Cognitive Discourse at the Time of IPO : *New England Journal of Entrepreneurship, 19(2)*, 12-24. https://digitalcommons.sacredheart.edu/neje/vol19/iss2/2

Zou, G., Cheng, Q., Chen, W., & Meng, J. G. (2020). What causes the IPO underpricing ? New evidence from China's SME market. *Applied Economics, 52(23)*, 2493-2507. https://doi.org/10.1080/00036846.2019.1693017

Zou, G., Li, H., Meng, J. G., & Wu, C. (2020). Asymmetric Effect of Media Tone on IPO Underpricing and Volatility. *Emerging Markets Finance Trade, 56,* 2474-2490. https://doi.org/10.1080/1540496X.2019.1643320

10 Appendix

Table 5: Topics

Comm. Shares	Cl. of Shares	Risk	Acquisitions	# of shares	Trad. Act.	Strat. All.	Result Stat.	Real Est.
stock	share	company	offer	combination	fund	partner	stock	income
stockholder	business	share	public	business	share	unit	share	investment
business	ordinary	stock	trust	share	trust	million	company	share
common	shareholder	financial	class	warrant	trade	general	common	interest
director	company	million	share	initial	investment	interest	statement	asset
public	combination	insurance	security	offer	commodity	cash	financial	property
security	director	bank	account	company	contract	agreement	offer	tax
share	initial	security	director	class	future	common	product	common
trust	public	business	business	purchase	master	financial	market	value
initial	security	loss	target	price	market	tax	security	rate
offer	warrant	december	transaction	sponsor	shareholder	end	director	loan
right	fund	common	issue	account	income	partnership	result	company
time	right	statement	exercise	target	price	distribution	price	financial
account	exercise	loan	agreement	officer	sponsor	increase	year	security
exercise	time	value	interest	exercise	tax	statement	sale	reit

Table 6: Spearman correlations

	Comm. Shares	Cl. of Shares	Risk	Acquisitions	# of Shares	Trad. Act.	Strat. All.	Result Stat.	Real Est.	undprc
Comm. Shares	1	-0.415	-0.226	-0.084	0.100	-0.046	-0.205	0.044	0.031	-0.152
Cl. of Shares	-0.415	1	0.159	0.253	-0.173	0.166	-0.066	-0.027	-0.001	-0.183
Risk	-0.226	0.159	1	-0.186	0.018	-0.108	-0.028	0.017	-0.411	-0.081
Acquisitions	-0.084	0.253	-0.186	1	0.063	0.387	0.127	0.350	-0.242	-0.051
# of Shares	0.100	-0.173	0.018	0.063	1	-0.007	0.191	0.330	-0.377	-0.097
Trad. Act.	-0.046	0.166	-0.108	0.387	-0.007	1	-0.071	0.234	0.008	0.142
Strat. All.	-0.205	-0.066	-0.028	0.127	0.191	-0.071	1	0.517	-0.173	0.123
Result Stat.	0.044	-0.027	0.017	0.350	0.330	0.234	0.517	1	-0.471	-0.139
Real Est.	0.031	-0.001	-0.411	-0.242	-0.377	0.008	-0.173	-0.471	1	0.189
undprc	-0.152	-0.183	-0.081	-0.051	-0.097	0.142	0.123	-0.139	0.189	1

Table 7: Spearman correlations p-values

	Comm. Shares	Cl. of Shares	Risk	Acquisitions	# of Shares	Trad. Act.	Strat. All.	Result Stat.	Real Est.	undprc
Comm. Shares		0	0	0.00000	0	0.003	0	0.004	0.039	0
Cl. of Shares	0		0	0	0	0	0.00001	0.077	0.939	0
Risk	0	0		0	0.223	0	0.061	0.274	0	0.00000
Acquisitions	0.00000	0	0		0.00003	0	0	0	0	0.002
# of Shares	0	0	0.223	0.00003		0.628	0	0	0	0
Trad. Act.	0.003	0	0	0	0.628		0.00000	0	0.589	0
Strat. All.	0	0.00001	0.061	0	0	0.00000		0	0	0
Result Stat.	0.004	0.077	0.274	0	0	0	0		0	0
Real Est.	0.039	0.939	0	0	0	0.589	0	0		0
undprc	0	0	0.00000	0.002	0	0	0	0	0	

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm