Norwegian School of Economics

Bergen, Spring 2018

Sentiment of Prospectus and IPO Underpricing

How textual analysis can explain IPO Underpricing phenomenon

Nurbol Kenessov and Meruyert Kanzhigalina

Supervisor: Kyeong Lee

Master Thesis in MSc in Economics and Business Administration

Finance, International Business

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

IPO Prospectus is one of the most important and informative documents filed by issuers with the Securities and Exchange Commission. Performing textual analysis of such document allows to understand issuers' perspective on the future of their company. In our research we show that while sentiment of the entire document could be not that useful, the analysis of certain parts of it can help understand underpricing. Basing our work on Ferris, Hao and Liao (2012), our approach was to consider the pull of US IPOs over the last 3 years. Having controlled for firm specific characteristics, we find that negative sentiment in the Risk Factors Section is positively related to 1st day underpricing. Moreover, we find a significant relation for a longer period underpricing - 6 months.

Keywords: *Underpricing, prospectus, sentiment, textual analysis, negativity, IPO*

Contents

1 Introduction

The phenomenon of initial public offerings' underpricing has long been debated among financial researchers. Underpricing by itself indicates a global market inefficiency, when firms due to some reason "leave money on the table". Much research has been conducted in this area, trying to identify possible reasons of why this anomaly takes place so often. However, most authors tend to analyze this problem from investor's and market's point of view, ignoring the perspective of the firm itself and its managers. In this paper we decided to step back from traditional approach and shift the emphasis on how issuers' view on the firm performance may affect IPO underpricing. In order to assess issuers' perspective on the future of their firm, we analyzed one of the most important documents filed during the IPO process - the prospectus. Recent developments in textual analysis allowed us to identify the sentiment of prospectus. Our hypothesis is that the negativity of a sentiment in IPO prospectus is positively related to the underpricing of IPO. In this hypothesis we assume that when issuer is uncertain about the future performance of the firm, it would be reflected in to what extent the sentiment is negative in certain parts of the prospectus. In order to compensate for the increased uncertainty, a lower offer price would be set for the investors, which in turn leads to a greater underpricing once trading begins.

This paper will be structured in the following way: in the next section we will give some background information about what IPO prospectus is and why it may be influential for pricing outcome. Further Section 3 will give an extensive literature review, covering all the papers related to our research topic, as there is a relatively limited number of works that need to be considered. This will be followed by the section called Negativity, where we would explain what sentiment of the document is and introduce our main independent variable. Section 5 will describe the process of data collection, which IPOs we chose for analysis and why, what kind of variables we derived and which controls we chose for our model. We continue with the 6th section called Methods, where we describe how automated textual extraction works and how the main independent variable was derived from prospectuses. Further section, Descriptive Statistics, will cover basic

statistics on data used to give some understanding of how our variables look like and how they correlate to one another. Important results of our analysis would be reported in section 8, Regression Results, which would be followed by several Robustness checks in Section 9. Finally, in the Discussion Section we will consider some limitations that we observe in our work, as well as give some ideas for further research in the field. In the end we will conclude and highlight the main results of the analysis.

2 Background Information

To begin with, it is important to understand the role of IPO prospectus in the entire process of firm going public. When the firm is preparing for IPO, there are several important steps and procedures that require filing of necessary documents. One of such documents is prospectus, which is basically the legal document that contains all the information about the firm, its history and performance, its management and future projections, as well as legal matters and important financial information. Prospectus of any firm needs to be filed with the Securities and Exchange Commission (usually referred to as SEC) and contains around 20 sections, such as Summary, Use of proceeds, Dividend policy, Principal Stockholders, etc. From a financial researchers' viewpoint, we would put emphasis on the Risk Factors Section as it contains important information for potential investors about possible risks they incur by purchasing stocks of the firm. As well, we will look at the Management's Discussion and Analysis of Financial Condition and Results of Operations section (to which we by convention refer as MDA), as this section gives investors clear understanding of a company's financial insight and performance of management.

Analyzing different prospectuses of IPOs, we agree with other researchers who claim that Risk Factors and MDA sections are the most representative for firms' performance and most unique for each company, while other sections are usually more formal and similar across prospectuses within an industry (Arnold, Fishe, North, 2010). As a result, we believe that these sections are specifically important for our analysis, as we try to

understand if the well-known issue of IPO underpricing could be driven by how skeptically prospectus is written by issuers. At the same time, in order to obtain a complete picture, we will look at the entire prospectus of each company, which will give us some understanding of general sentiment of the full prospectus across companies.

3 Literature Review

The anomaly of IPO underpricing or how it is sometimes called "money left on table" has been widely discussed and researched both in academia and in business. On average, underpricing, or the initial first day return from IPO purchase, is estimated to be 17%, and the question is what drives this number (Loughran and Ritter, 2004). Mostly existing research models are based on the idea of information asymmetry, when issuers, underwriters and investors have some important pieces of information and lack other pieces, which results in money being left on the table (Lowry, Michaely, Volkova, 2017). Most of the research is concentrated on the role and incentives of underwriters in IPO process (ibid). Moreover, a great portion of IPO underpricing research considers the investors' view and market demand on IPOs (Ferris, Hao, Liao, 2012). However, the existing literature in a way ignores the view of issuers themselves, leaving a gap in understanding the issue of underpricing (ibid). Thus, in our work we will attempt to help filling in this gap by questioning whether the view and sentiment of issuing company's management affects underpricing. As we put textual analysis of IPO prospectus at the center of research, we carefully analyzed the existing literature on the issue. In general, we can see that this area has not yet been considered extensively. Such situation could be present due to relative novelty of textual analysis as a method by itself (Loughran, McDonald, 2016). Even though researchers have been trying to analyze texts since around 14th century, only the recent progress in computer technology and availability of data in digital format allowed scientists to automate certain processes and work in this field on a bigger scale. For example, the very first authors who worked on textual analysis in the financial and accounting areas were Antweiler and Frank in 2004, Tetlock in 2007, Das and Chen in

2007, so this could doubtlessly be called an emerging area (Loughran, McDonald, 2016).

As a result, we can see an even smaller pool of literature where IPO prospectuses were analyzed to explain underpricing phenomenon. In a historical retrospective, the very pioneering researchers were Beatty and Ritter, who were the first to analyze qualitative information in IPO prospectus in 1986. In their method, they counted the number of uses in the Use of Proceeds section to approximate for ex ante uncertainty about IPO value, basing on a SEC requirement for "more speculative issues to provide relatively detailed enumerations of the uses of proceeds, while not requiring more established issuers to be very explicit" (p.218). From such approximation they found positive relation between ex ante uncertainty and underpricing of IPOs.

A more recent research "The Effects of Ambiguous Information on Initial and Subsequent IPO Returns" was done by Arnold, Fishe and North in 2010. Having a sample size of 1400 IPOs, they considered risk factors section as soft and ambiguous information, assuming that risk information by itself is interpretation, thus could be a measure of IPO ambiguity. Authors used different word count ratios of soft (risk factors section) to hard information (the whole prospectus and business-related parts of it) to approximate ambiguity of prospectus. In regression authors used ambiguity ratios, return variables, riskiness variables, control variables (proceeds, firms size, firm age, days in registration, lead underwriter reputation, venture capital backing, market activity), and dummy variables (control for 1990-2000 IPO boom, control for technology). Arnold et al. concluded that investors will require premium for the firms that have more ambiguous prospectuses, or to rephrase it, underpricing will increase with ambiguity in the document.

Further Hanley and Holberg (2010) considered the importance of prospectus in IPO pricing. Authors claim that the information gathered during premarket (by issuer and underwriter) gives more informative and unique content for prospectus, while the information gathered during Book Building (by investor) is more standard, as it is based on the industry data available to all. Having a sample size of 1700 IPOs, Hanley and Holberg (2010) controlled for firm specific characteristics, like firm age, underwriter's market share, venture capital dummy, NASDAQ return, IPO size and technology dummy. Au-

thors measure how unique and informative the content of prospectus is, using diverse textual analysis tools like normalized word vectors, and based on this create measures of degree of similarity between prospectuses. As a result, they show that greater effort in premarket leading to more informative content results in more accurate initial offer pricing, and thus less underpricing.

Same authors have published a more recent work, "Litigation risk, Strategic disclosure Underpricing of IPOs" (2011-2012), where they claim that issuers tradeoff between underpricing and strategic disclosure as hedges against risk. In this paper Hanley and Holberg use word content analysis to understand this trade off, meaning that on one hand greater disclosure may decrease the probability of lawsuits from investors (i.e. for material omissions in prospectus), while on the other hand it could be costlier as the information is usually of a high proprietary value for a company.

Another interesting study that requires attention is "Soft strategic information and IPO underpricing" by Braua, Ciconb and McQueen (2012). In this research authors used textual analysis of IPO prospectus to explain underpricing. Particularly, they looked at strategic tone of the document and found that it is positively related to first day returns from IPO. Importance of this research is that in order to derive strategic tones of documents authors created new dictionary, or how they refer to it "new content-analysis libraries for strategic words" (p.1).

Finally, among the most sophisticated studies that uses textual analysis algorithms is the research of Ferris, Hao and Liao (2012) "The effect of issuer conservatism on IPO Pricing Performance". In this paper authors examined the relation between the prospectus conservatism and IPO pricing, as well as consequent stock return performance. These researchers were among the first who outlined the problem that underpricing phenomenon is actively analyzed from investors' point of view or from market perspective, while the perspective of issuers is mostly ignored. Consequently, authors decided to step in this area and addressed the beliefs of issuers about the future performance of their firms. More particularly, Ferris, Hao and Liao (2012) examined if the use of cautionary language, or, as they call it, "conservatism" in IPO Prospectus is related to the pricing and performance

of the initial public offering.

Looking at data from 1999 till 2005 with a sample size of 1,100 US IPOs, authors begin with the analysis of the whole prospectus followed by a separate analysis for each of the four sections: summary, risk factors, use of proceeds and MDA (later dropping the use of proceeds section due to insignificance of results). In order to define the negativity tone of a document, authors derive a main independent variable, Conservatism defined as a proportion of negative words to total words. In their research, authors tested 3 different dictionaries (Loughran McDonald, Harvard Psycho-Social and Diction) to derive this variable, concluding that Loughran McDonald one allows to construct best measure of negativity tone. Controlling for firm specific characteristics, authors looked at firm size, age, VC backing, lead underwriter score, auditor market share and technology dummy: such method of controlling is consistent with what most researchers in IPO area do. Results show that there is a positive relation between conservatism of IPO prospectus and underpricing, especially in technology firms.

Admittedly, Ferris, Hao and Liao in their paper chose such period where they encompass the so-called dot-com bubble around 2000-s when fast growth of Internet usage took off, changing industries worldwide. In their analysis authors often highlight the difference between the tech and non-tech firms, deriving significantly different results for these groups. For example, researchers emphasize that valuation of tech firms is more difficult than non-tech ones, and thus Conservatism score for them is much higher. This could partially be driven by the fact that the chosen time period was the beginning of digitalization era where technology firms were a newly emerging trend with many uncertainties around them. In general, all the listed sources cite each other and base on one another, which again highlights the limited amount of research in this area.

In our research we have decided to replicate the study of Ferris, Hao and Liao (2012), with a more recent data period, as with time market trends change and new results may differ from what was true more than 10 years ago. As well, we are contributing with more up to date text extraction and analysis algorithms that allow for automated data collection process. We are using the same principle for conservatism estimation, however

recent developments in textual analysis allowed us to automate this part of the estimation.

4 Negativity

Our research revolves around the sentiment of the prospectus text, the sentiment that issuer puts into the text when describing the company and its post-IPO future. In order to determine the sentiment of the text under consideration we used the SentimentAnalysis package in R, which includes the AnalyzeSentiment function. This function outputs a proportion of words with negative meaning to the total number of words. It is worth mentioning that the function removes any stop words that might affect the calculation. The list of stop words includes words such as articles (a, the, an) or auxiliary verbs (be, have), their inclusion might introduce a downward bias to negative words ratio by increasing the total word count. Let us consider a trivial example of how this function works: the text that says "the stock price has declined" would be analyzed in the following way. First, stop words "the" and "has" would not be included in the word count, and the text left to analyze is "stock price declined". The function determines whether the words in the supplied text are within the negative, positive or neutral sets of words based on the dictionaries embedded into the function. In our small example, the only word with the negative tone is "declined", the negativity score attributed to this would be 1. Thus given that it's the only negative word, and our set has three words in total: negativity score of the text will be 1/3 or 0.33. On the other hand, if the function would count the stopwords in the total word count, the negativity score of the text would have been equal to or 0.2, being less that what it should actually be. Overall, the function allows us to obtain the negative tone measurements for the supplied text, and we used the "AnalyzeSentiment" to obtain negativity scores for our dataset.

When it comes to estimation of the negativity scores, we did that for the entire prospectus as well as for two of its separate sections. First of the separate sections that we considered is the Risk Factors: it outlines the uncertainties arising from micro and macroeconomic factors. Next section under the consideration is MDA which stands

for Management's Discussion and Analysis, it covers management's analysis of financial condition and results of operations. In our research, we will outline the effect of each section's negative tone on underpricing. We will attempt to determine whether increased negativity of the entire prospectus, the Management's Discussion and Analysis section or the Risk Factors section has causal effect on the IPO underpricing.

5 Data collection

In our research we have decided to consider the US IPOs that were issued at the period between the beginning of 2014 and the end of 2016, capturing 3 years in total. This period was chosen to take a look at more recent IPOs and at the same time to make sure that we consider the companies that still exist one year after going public to be able to look at their longer-term performance.

Initial sample that consisted of more than 500 IPOs was obtained from Thomson Financial Securities Data Company, to which we refer as SDC. However, among those 500 companies there are many IPOs in which we are not interested for this study, so we excluded companies with offer price of less than 5 US dollars (since they are regarded as too small ones), units, limited partnerships, ADRs, which is consistent with what previous researchers in the area of IPOs did (Arnold, Fishe, North, 2010). As well, we excluded financial companies and heavily regulated industries since they operate under different regulations and quite often do not follow the general economic trends.

From the SDC database for each company we derived the following characteristics: 1st trading day return, as well as 30th, 90th, and 180th days' returns, amount filed, gross spread, high tech code, VC dummy, CIK, total assets before going public, underwriters' names, industry code, type of shares, yesterday's price at the moment of download. Moreover, we used Jay Ritter's databases to derive the number of years firm existed before it went public.

Looking at these characteristics more precisely, 1st trading day return is calculated as a percentage difference between the IPO's offer price and the closing price at the end of

the first trading day. In the literature this measure is often referred to as Underpricing, since it tells how much money the firm left on the table (Arnold, Fishe, North, 2010), and thus in our model it is a dependent variable. As different authors identify underpricing in different ways, we further looked at 30th, 90th, and 180thdays' returns, calculated in a similar way, in order to consider longer term performance of IPOs. We use them as dependent variables in later models.

In our model we decided to control for several factors that may affect underpricing, such as firm specific characteristics. So we looked at the size of the firm in millions of USD, calculated as the total assets of the firm before they issued IPO; firm age in years, calculated as the difference between the issue date and Jay Ritter's reported founding date. As well, we added IPO characteristics like Amount filed (in millions of US dollars) and Gross Spread, as literature suggests controlling for it (Ferris, Hao, Liao, 2012).

Other characteristics were mostly taken to filter out unnecessary IPOs. For example, the type of stock was considered in order to filter out all but common shares, as other types, for example class B shares, have different regulations and it is unreasonable to compare them to common ones. Further, we looked at yesterday's price for the moment of data extraction to avoid IPOs that do not exist anymore. One more characteristic derived from SDC database was CIK as identifier for IPOs, which was essential for integrating different sources, such as Jay Ritter's database, SDC and Securities and Exchange Commission (further referred to as SEC). Further for each IPO we tried to look at leading underwriters, as literature suggests that underwriters' reputation may affect underpricing. However, this data was not used due to lack of information in databases.

Crucial step in our data collection was extraction of IPO prospectuses that are all available at the US SEC website. During this process we faced several difficulties, as manual search and extraction of more than 200 documents was time consuming, so we tried to automate this process. For each IPO we first took the deal number from SDC and then used a link that led to all the documents related to IPO issue. After this point automation was impossible due to the following issue: SEC website presents IPO prospectuses in the S-1 Form (or sometimes F-1), however each firm tends to download

several versions of this form and its amendments as they obtain more data both before and even after going public. While all of them are filed as S-1 Form, it is important to make sure that we derive the form that is most recent for the day of IPO issue, but not later than this date, to make sure we encompass all the information available prior to the day of going public. Another issue was that we needed to make sure that the form we downloaded was the full form, and not just an amendment, as amendments had the same filing (S-1/F-1), but were considerably shorter and contained mostly numbers, legal notes and thus were not of interest for our textual research. The final issue was that even though SEC has strict regulations and requirements in which form prospectuses and documents should be submitted, there were still cases when part of forms were not machine readable due to text being inserted as pictures or special tables that complicated the process.

Finally, the main independent variable, negativity score, was not obtained from any existing source, but was derived using R code that performs textual analysis of the prospectuses, which will be discussed further in this paper.

6 Methods

In this section, we will go over the main algorithm that allowed us to collect the necessary data from IPOs' prospectuses, extract the sections of interest and analyze the sentiment. First step was to read in the text from the prospectus, where we used the function that reads lines of the webpage into a text, since we needed to keep the HTML tags for subsequent steps. In short, the function identifies the beginning of the section via the HTML anchors that precede the section title and captures everything between the HTML anchor of the required section and the beginning of the next section indicated by the next HTML anchor. After obtaining the text in between of two HTML anchors, we proceed to clean it from the HTML tags and save it using the CIK number as an identifier.

In this paragraph, we cover the main cleaning function in more detail. The cleaning function reads in the entire prospectus and as a first step it proceeds to remove the text that comes before the section of interest. It does so via using the regular expression that

consists of an HTML anchor and the section name since the string of text that includes the HTML anchor and the section name in capital letters signals the section's beginning. The function then removes anything that comes before the section's beginning, leaving the portion of the prospectus that begins from the section of interest and goes until the end of the document. Second step is to remove everything that comes after the section of interest. This is done in a similar way, using the regular expression that captures HTML anchor that signals section's beginning. This time, since we have a text that begins with the desired section, function will look for the beginning of the next section (rather than an ending of the current section, since it would be technically complicated to do so) using the same HTML anchors for the section beginning. After finding the beginning of the next section (i.e. "Business", "Legal") we remove everything that comes after it, thus leaving us with the section of interest only. After isolating the section of interest, the function removes anything that is marked as a table and uses one of the HTML specific functions to extract the text from an HTML string.

After obtaining the cleaned text of the necessary sections and of an entire prospectus, we proceed to use the AnalyzeSentiment function, that calculates the aforementioned Negativity score. A dataframe with an identifier and text columns was supplied into the function, and we obtained the negativity scores for our sample: entire prospectus, risk factors section and MDA section. Thus, this dataframe stored main explanatory variable of our model, we refer to it as *negativity scores dataframe*. As a next step, the dataframe consisting of the necessary control variables was merged with the *negativity scores dataframe* according to the CIK identifier. As a result, the merged dataframe served as a regression data source.

7 Descriptive Statistics

We have summarized the statistics about the main IPO variables in Table 1 Summary Statistics. For the short term our main dependent variable, first day underpricing, has an average of 16.2%. This is consistent with post 2000 IPO literature that says the average

first day returns from trade of IPO shares worldwide is around 15% (Lowry, Michaely, Volkova, 2017). At the same time, the median value of 1st day underpricing is lower than the mean, 5.6%. Admittedly, the largest outlier for this variable is 206.67%, which is a huge level of underpricing. As for the longer term models, our dependent variables are 1, 3 and 6 months underpricing. The average values for these variables are around 20%, which is a bit higher than for the 1st day underpricing. Analogous tendency could be observed for the median values that are around 15%, 14% and 11% percent respectively, which is higher than median value for short term underpricing.

Considering the control variables, we have the Amount filed with median USD 86.3 million , and a higher mean around USD 108 million, while the maximum amount is 800 million US dollars. Due to missing data we were not able to find this characteristic for the whole sample, so we have 22 missing observations here. Moving further, firm age varies significantly from the youngest firm being one year old to the oldest firm having 150 years of experience. While the mean age of firms in our sample is 17.2 years, this is due to old outliers as mostly firms are less than 17 years old, which can be seen from 75th percentile. The firm size variable has a big range from smallest firms being less than 1 million dollars to the maximum value being more than 8 billion dollars, while the median value is approximately USD 70 million. We have 7 missing observations for this control. Finally, Gross Spread variable has both mean and median around 7Now we can look at the negativity measures we have derived from prospectus analyses, being our main independent variable. We first look at negativity scores of the entire document for each IPO, seeing that the average score for our sample is slightly less than 5. For the MDA section we observe a similar situation with median score being 4. Admittedly, the Risk Factor Section's median value is significantly higher, 9.27. Due to initial filtering we have not lost any observations for our main explanatory variable.

From the correlation table we can observe that all the underpricing variables are positively correlated with one another and this relation is statistically significant. One can observe that the closer the period of estimation, the higher the correlation coefficient, for example 1st day underpricing is highly correlated to 30 days underpricing, and less

Table 1: Summary Statistics

Statistic	N	Pctl(25)	Median	Pctl(75)	St. Dev.
1-day Underpricing	239	0.000	5.588	23.280	31.683
1-month Underpricing	238	0.000	15.375	40.803	37.608
3-month Underpricing	239	−8.500	14.290	51.145	48.836
6-month Underpricing	239	−22.360	10.830	49.270	59.421
Filed Amount	217	74.800	86.300	100.000	94.299
Firm Age	239	7	10	17	22.102
Firm Size	232	28.600	70.200	353.250	1,263.914
Gross Spread	238	7.000	7.000	7.000	0.650
Full Prospectus Negativity LM	239	4.638	4.982	5.296	0.466
Risk Section Negativity LM	239	8.239	9.270	9.897	1.744
MD&A section Negativity LM	239	3.701	4.032	4.490	0.590

correlated with further period. This could be due to certain price adjustment that takes place over a longer time horizon. Moving further, for the controls we mostly observe no significant correlation. Exception in this case would be the firm size, which is positively correlated to amount filed and age of the firm. This has a logical explanation, as mostly big firms are able to file a greater amount for IPO issue. Moreover, older firms tend to acquire more assets through their long history, thus have bigger size. Gross spread variable has a statistically significant negative correlation with firm characteristics, such as age, size and amount filed. Even though the usual practice is that established underwriters such as Goldman Sachs charge 7% as their spread, it seems that for bigger and experienced firms the spread is usually a bit less. This is probably due to underwriters' belief in successful launch of IPOs for big firms, as they are more experienced and stable, whereas for smaller and younger projects underwriters need to secure themselves by charging a higher spread.

As for the main independent variables, Full text of prospectus (FT) and MDA negativity scores show no significant correlation with any other variables. Risk Section negativity score has a significant negative correlation with longer term underpricing. However, at the moment it is hard to find explanation for such results, as more regressions and more sophisticated analysis is needed.

Table 2: Correlation Table

	1dUp	1mUp	3mUp	6mUp	Amt	Age	Size	Spr	FT	Risk
1dUp										
1mUp	0.65*									
3mUp	0.38*	0.64*								
6mUp	0.26*	0.49*	0.72*							
Amt	-0.02	-0.01	0.07	0.00						
Age	-0.09	-0.09	-0.05	-0.02	0.09					
Size	-0.11	-0.12	-0.03	-0.02	0.25*	0.50*				
Spr	0.06	0.13	0.07	0.04	-0.40*	-0.39*	-0.70*			
FT	0.08	0.12	0.01	0.02	-0.01	-0.03	-0.05	0.09		
Risk	-0.08	-0.08	-0.14(.)	-0.13(.)	0.02	0.07	0.00	-0.02	-0.12	
MDA	-0.09	-0.09	-0.04	0.03	-0.01	0.03	0.01	0.07	0.07	-0.07

Note: $^{(.)}$p<0.1; ; *p<0.01

8 Regression Results

For the regressions, it was decided to use and report two specifications. First specification includes the main explanatory variable and the dependent variable. Further, for the second specification we decided to control for several firms specific characteristics, so it contains main explanatory variable, dependent variable and control variables. Instead of analysing linear relationship for the controls, we used the log measure for the size, filed amount and age variables to decrease the effect of outliers within the sample. Thus, in total six different regressions were performed: two specifications (simple and with control variables) per three sections of interest (entire prospectus, MDA and Risk Factors sections).

From the regressions presented in Table 3 we observe that the entire prospectus negativity score has positive coefficient in simple specification, but the sign flips in the full specification with all the controls. However, none of the coefficients ended up being statistically significant here. Such results could be driven by the fact that the full document reports extensive information and has huge word count, and a number of sections that differ in their purpose (i.e. marketing oriented sections, legal matters), so it is important to consider certain sections that seem to be most valuable for investors in terms of future

performance predictions.

When it comes to the MDA section, we had negative coefficients on the relation between negativity score of the section and 1st day underpricing. In the case of a longer term underpricing, we did not observe any significant explanatory results for both entire prospectus and the MDA section when altering the dependent variable to 1-month, 3-month and 6-month Underpricing. This is not consistent with the previous research on the topic, where MDA section proved to be significant in explaining underpricing in IPOs. However, insignificance of the MDA section can be explained in two different ways. First, due to potential changes in reporting style, significance of the section may have decreased in more recent years, and since our dataset focuses on three most recent years, we might be observing a new trend. Second, although our sample covers recent years, it still might be affected by its size and exclusive focus on US based stocks.

However, when it comes to the Risk Factors sections, we obtained significant results for both regression specifications. We discovered that in both simple and a more complex regressions, Negativity score in the Risk Factors section has a significant positive effect on IPO underpricing, as it is depicted in Table 3. Specifically, one percentage point increase in the proportion of negative words increases the 1st day underpricing by 2.47p.p in the regression specification with control variables. In other words, a more negatively written Risk Factors section would result in higher 1st day Underpricing levels for an issuer. This can be related to the increased uncertainty and severity of the future risks signaled by higher concentration of negative words within the Risk Factors section.

Further step was to look into a longer term underpricing and its relation to the negativity in the Risk Factors section. By changing the dependent variable to a longer term underpricing, we obtain significant relationship between negativity in risk factors section and 6-month underpricing. From this, we can assume that negativity in Risk Factors section partially explains the longer-term underpricing. However, we do not observe any significance between negativity and other underpricing windows: neither 1 month nor the 3 months underpricing is explained by negativity in Risk Factors section.

Table 3: Negative LM Sentiment and Underpricing

	Entire prospectus		MD&A section		Risk section	
Negativity LM	1.999	-4.088	-5.633	-5.237	2.858**	2.466*
	(4.418)	(5.477)	(3.807)	(4.379)	(1.163)	(1.305)
VC binary		11.804*		17.428**		15.657**
		(6.459)		(7.600)		(6.684)
log Amount Filed		-0.616		5.382		0.204
		(5.258)		(6.054)		(5.359)
log Firm Age		-2.526		-2.859		-0.867
		(2.913)		(3.495)		(2.979)
High-Tech binary		5.275		5.308		3.640
		(6.551)		(6.863)		(6.446)
log Firm Size		2.792		4.733**		2.095
		(2.060)		(2.258)		(2.007)
Spread		1.973		9.084		-4.155
		(4.486)		(6.080)		(4.257)
Constant	6.295	8.194	41.322***	-76.333	-9.827	0.703
	(21.994)	(47.745)	(15.789)	(53.797)	(10.414)	(39.414)
Observations	239	211	239	213	239	211
R^2	0.001	0.041	0.009	0.103	0.025	0.066
Adjusted R^2	-0.003	0.008	0.005	0.072	0.021	0.034
Note:						*p<0.1; **p<0.05; ***p<0.01

Interestingly, throughout different specifications and different sections' regression in Table 3, we observed consistent significance of the VC binary's effect on 1st day underpricing. Later in Robustness check section we select the subset of IPOs that were backed by venture capital in attempt to isolate the effect of VC dummy and see if negativity score of the prospectus has any effect on underpricing in the case of VC backed firms.

As a result, our findings suggest that extensive negative tone in the risk factors section can indicate higher IPO underpricing on the 1st trading day and in 6-month post IPO performance. Regression results suggest that one percentage point increase in the proportion of negative words within risks section leads to approximately 2.5 p.p. increase

in IPO underpricing. High proportion of negative words in the risk factors section is probably related to issuer's concern and greater uncertainty about going forward with the offering, thus creating more ambiguity for investors and ultimately leading to a greater underpricing.

Table 4: Negative LM Sentiment and 6-month Underpricing

	Entire prospectus		MD&A section		Risk section	
Negativity LM	7.058	-0.862	5.272	2.250	3.633	5.479**
	(8.277)	(10.316)	(6.879)	(7.838)	(2.211)	(2.449)
VC binary		19.433		45.680***		25.485**
		(12.165)		(13.604)		(12.541)
log Amount Filed		-0.236		-1.915		-2.639
		(9.902)		(10.838)		(10.056)
log Firm Age		-4.050		-13.549**		-4.193
		(5.486)		(6.257)		(5.589)
High-Tech binary		8.071		-11.187		5.990
		(12.337)		(12.285)		(12.095)
log Firm Size		6.770*		8.801**		7.673**
		(3.880)		(4.043)		(3.766)
Spread		6.023		8.438		7.458
		(8.450)		(10.885)		(7.988)
Constant	-14.936	-55.430	-2.939	-68.542	-12.674	-113.564
	(41.204)	(89.919)	(28.533)	(96.303)	(19.794)	(73.958)
Observations	239	211	239	213	239	211
R^2	0.003	0.037	0.002	0.111	0.011	0.072
Adjusted R^2	-0.001	0.004	-0.002	0.081	0.007	0.040
Note:						*p<0.1; **p<0.05; ***p<0.01

9 Robustness

9.1 Subset of Venture Capital backed firms

In order to check if the obtained results are valid in different conditions we decided to perform two robustness checks - subsampling and logarithm of the main explanatory term. Having performed the regression in the sub-sample of Venture Capital backed firms, we observe continued significance of the negativity score in risk factors section in explaining the 1st day underpricing. In both specifications risk factors negativity has bigger effect on underpricing than in the general sample. Table 5 shows that 1 percentage point increase in the negativity of Risk Factors section leads to approximately 3 p.p. increase in underpricing. This suggests that our main explanatory variable is still statistically significant in the subsample of Venture Capital backed firms.

Table 5: Negative LM Sentiment and Underpricing in VC backed firms

	Entire prospectus		MD&A section		Risk section	
Negativity LM	-4.829	0.325	-3.222	-4.373	4.066**	3.135*
	(7.837)	(8.319)	(5.858)	(6.420)	(1.651)	(1.720)
log Amount Filed		6.568		2.701		10.750
		(12.984)		(12.269)		(11.093)
log Firm Age		-3.863		-5.321		-1.782
		(5.153)		(6.660)		(4.746)
High-Tech binary		8.974		12.877		10.224
		(12.235)		(12.340)		(10.882)
log Firm Size		9.335**		11.371**		8.477**
		(4.231)		(4.475)		(3.783)
Spread		-32.684***		-8.957		-32.184***
		(10.682)		(17.587)		(7.508)
Constant	44.886	182.772**	36.674	49.558	-16.007	132.515**
	(40.189)	(85.849)	(24.276)	(131.878)	(14.695)	(60.952)
Observations	143	129	156	144	151	137
R^2	0.003	0.153	0.002	0.114	0.039	0.211
Adjusted R^2	-0.004	0.112	-0.005	0.075	0.033	0.175
Note:						*p<0.1; **p<0.05; ***p<0.01

9.2 Logarithm of the Negativity score

The second robustness check included taking the logarithm of the independent variable – the negativity score. We perform this check in order to decrease the effect of any outliers and extremely high negativity scores. Table 6 depicts the regressions where the risk factors section's negativity score remained as the only significant explanatory variable, as in the case of the main regressions. The coefficients suggest that a 10% increase in risk section's negativity results in 1.8% increase in the 1st day Underpricing. Taking the logarithm of the independent variable resulted in consistent outcome, with negativity in risk section being significant and with the similar positive effect on underpricing.

Overall we can see that negativity score of the risk factors section remains significant and robust throughout different specification and subsamples.

Table 6: Log of Negative LM Sentiment and Underpricing

	Entire prospectus		MD&A section		Risk section	
Negativity LM	10.887	-18.373	-23.403	-21.139	19.926**	18.089**
	(21.505)	(26.596)	(15.655)	(18.389)	(7.840)	(8.611)
VC binary		11.787*		17.547**		15.753**
		(6.475)		(7.593)		(6.671)
log Amount Filed		-0.626		5.407		0.390
		(5.263)		(6.067)		(5.336)
log Firm Age		-2.485		-2.785		-0.691
		(2.910)		(3.494)		(2.966)
High-Tech binary		5.197		5.351		3.881
		(6.546)		(6.870)		(6.418)
log Firm Size		2.802		4.750**		2.069
		(2.061)		(2.260)		(2.002)
Spread		1.950		9.109		-4.178
		(4.487)		(6.082)		(4.245)
Constant	-1.173	17.374	51.007**	-68.814	-27.475	-17.671
	(34.387)	(56.835)	(22.064)	(56.672)	(16.943)	(42.069)
Observations	239	211	239	213	239	211
R^2	0.001	0.040	0.009	0.103	0.027	0.070
Adjusted R^2	-0.003	0.007	0.005	0.072	0.022	0.038
Note:						*p<0.1; **p<0.05; ***p<0.01

10 Discussion

In this section we would like to discuss potential limitations of our research, its general applicability and suggest some areas for further work in this field.

10.1 Dictionaries

Our work is an entry step into a currently very narrow research area that has big potential applications going forward into the future.In our research, we reported the negativity scores using single dictionary: the Loughran – McDonald financial dictionary. Since in the paper by Ferris et al (2012) the use of general language dictionaries in analyzing financial texts was proven inefficient, we decided to omit those dictionaries. Generally, the output of the main function used to calculate the negativity score also includes two additional measures for the negative tone in the text. Different measures arise from different dictionaries in use. In total, the AnalyzeSentiment function uses three different dictionaries: Harvard-IV, Henry's Financial Dictionary and Loughran-McDonald Financial Dictionary. First, the function uses Harvard-IV dictionary as used in the General Inquirer Software, which is a representation of a general language dictionary. Sentimental scores attached to different words in this dictionary are of general purpose and cannot be used in financial line of research. The second dictionary that we did not use in our research is Henry's Financial dictionary. The reason we did not use it is a very low quantity of negative words it has - 85 words, while Loughran-McDonald dictionary identifies 2355 words as negative. Historically, Henry's dictionary was a first attempt of adopting general language dictionaries for financial use in 2008. Loughran-McDonald took it further when they compiled their own dictionary in analyzing annual reports. Therefore, the choice of a dictionary to base our negativity score on was quite obvious and we proceeded with Loughran-McDonald dictionary throughout the paper. One possible way to vastly improve the research on the topic would be compiling an IPO prospectus specific dictionary, which would capture the sentiment in the prospectus text more accurately than the Loughran-McDonald dictionary. This task requires substantial effort and dedication and it was not feasible

within the frames of a master thesis. However, having purposefully designed dictionary would increase the accuracy and drastically improve the results. Most bright example of this is the case of Loughran-McDonald dictionary, authors compiled their own financial dictionary to analyze 10-K annual reports, which then became the industry standard in analyzing financial texts. (Loughran and McDonald, 2011). Another case is the work of Braua, Ciconb and McQueen (2012), who created their own dictionary in order to capture strategic words more precisely and accurately. Overall, textual analysis of any financial items is an emerging research area and there are numerous possible applications that can prove to be useful. For instance, one might use textual sentiment in cases where a prospectus or a report is the only available information. This is the case of a recent phenomenon called ICOs: Initial Coin Offerings. There is even less information and certainty around the ICOs since all the available information is concentrated only within the released whitepaper. Sentiment analysis could be one of the soft information tools for investors to assess the ICOs.

10.2 Limitations and Further Research

In our research we decided to concentrate on the US IPOs only, as reporting requirements and data availability for these IPOs allowed us to conduct textual analysis and research.. Admittedly, up until 2014 the US was a global leader in IPOs by deal volume, as well as by number of IPOs (EY Global IPO trends, Q4 2014). However, as global trends change with time, the US is no longer the sole leader in this area. According to EY Reporting, in 2015 China has replaced the US in terms of IPOs by deal volume, being more than twice as big as the US, as well as by number of deals. Further, 2016 was the first year when the US was not present in the top 10 worldwide deals of the year (EY Global IPO trends, Q4 2016). As a result, we can observe that even though the US has always been in the center of research on IPO topic due to greatest activity and biggest market worldwide, this might not longer be the case. Asia-Pacific, especially China and Japan, is becoming a much bigger and active market for IPOs, raising more than 50% of funds globally (ibid.) Thus, in the future researchers may need to put more emphasis on what is happening in

this newly emerged part of the global market.

Another possible limitation of the conducted research was that data collected was for 3 years only. This resulted in an initial sample size of approximately 500 firms, but due to necessary filtering we ended up with 239 observations only. This could be the reason for insignificance of some variables, for example MDA negativity score. Finally, in order to control for firm specific characteristic we initially planned to include more variables, however due to poor reporting and mismatch of some data (i.e. different CIK identifiers) between databases we were not able to get some of control variables. For example, researchers claim that underwriter's reputation may affect underpricing (Lowry, Michaely, Volkova, 2015), however SDC database did not always report the underwriter for each IPO, thus we were not able to find a reputation score. Even though these were mostly minor variables, they still could have contributed to more accurate results.

Coming to the question of general applicability of our approach, it is possible to claim that it is relatively easy to extend the research over longer horizon, as certain processes have been automated using programming algorithms. For example, text extraction algorithm in R allows to scale up the research for much more than 3 years without major sophistications. This will allow for greater sample size and bigger research period. However, this scalability is most probably applicable to the US IPOs only, as they have same reporting requirements set by the US Securities and Exchange Commission, so general style among prospectuses is similar. As for the global scale, additional algorithms would be needed, as reporting requirements may differ for Europe and Asia, so there would be a need to adjust textual algorithms to specific needs.

All in all, in this section we have identified several dimensions for further research in this area. In addition to this a possible way for improvement could be to analyze the sentiment of other sections and see whether it results in IPO underpricing.

11 Conclusion

Our research aimed at analyzing the issuer's perspective on the post IPO performance through the lense of textual sentiment and how the uncertainty in going forward reflected in negative sentiment of the text can affect the underpricing levels for the issuer. Every step of the research was performed in R statistical software, including: data collection, processing and statistical analysis. Our dataset covered three most recent years of the US IPOs and allowed us to extract the negativity sentiment of around 240 IPO prospectuses. Non-uniformity of the prospectuses' structure and formatting resulted in complications connected with data collection. However, using R-code with appropriate packages and functions allowed us to automate most of the processes and made our research scalable for future implementations. Our work replicates the work by Ferris et al. (2012), however, we are focusing on a more recent data and we use newly developed R-packages.

We controlled for different firm specific factors including firm's size, age, filing amount, underwriter's spread, whether or not it is VC backed and technological level. The regression results uncovered positive relationship between negativity in Risk Factors section and underpricing, leading us to a conclusion that higher negative sentiment in the risks section can lead to increase in both 1st day and 6-month post IPO underpricing. Entire prospectus text as well as second section of interest: the Management's Discussion and Analysis, did not yield significant regression results. These results may be driven by different factors that include sample size, reporting style and changes in trends.

Although the topic of textual analysis in finance has a limited literature at the moment, its relevance is increasing with the development of analytical software and hardware. Our research contributes to the existing literature in two ways. First, we take a look at most recent data available on US stocks. Second, we use statistical software and programming to benefit from quasi-automated data collection and analysis processes . Overall, uncertainty of the issuer reflected in the prospectus' sentiment can have significant effect on IPO underpricing, shifting the paradigm of analysis from investors to issuers can uncover new possible explanations for the phenomenon of underpricing.

Bibliography

Arnold, T., Fishe, R. P. & North, D. (2010), 'The effects of ambiguous information on initial and subsequent ipo returns', *Financial Management* **39**(4), 1497–1519.

Beatty, R. P. & Ritter, J. R. (1986), 'Investment banking, reputation, and the underpricing of initial public offerings', *Journal of financial economics* **15**(1-2), 213–232.

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

EY Global IPO Center of Excellence. EY Global IPO Trends 2014 Q4 (2014).
URL: *http://www.ey.com/Publication/vwLUAssets/ey-q4-14-global-ipo-trends-report/$FILE/ey-q4-14-global-ipo-trends-report.pdf*

EY Global IPO Center of Excellence. EY Global IPO Trends 2015 Q4 (2015).
URL: *http://www.ey.com/Publication/vwLUAssets/EY-global-ipo-trends-2015-q4/$FILE/EY-global-ipo-trends-2015-q4.pdf*

EY Global IPO Center of Excellence. EY Global IPO Trends 2016 Q4 (2016).
URL: *http://www.ey.com/Publication/vwLUAssets/ey-global-ipo-trends-report-4q16/$FILE/ey-global-ipo-trends-report-4q16.pdf*

Ferris, S. P., Hao, Q. & Liao, M.-Y. (2012), 'The effect of issuer conservatism on ipo pricing and performance', *Review of Finance* **17**(3), 993–1027.

Hanley, K. W. & Hoberg, G. (2010), 'The information content of ipo prospectuses', *The Review of Financial Studies* **23**(7), 2821–2864.

Hanley, K. W. & Hoberg, G. (2012), 'Litigation risk, strategic disclosure and the underpricing of initial public offerings', *Journal of Financial Economics* **103**(2), 235–254.

Lee, T., Rice, S., Rock, S. & Willenborg, M. (2010), 'Explaining underpricing through the textual analysis of ipo registration statements', *Utah Winter Workshop* .

Liu, X. & Ritter, J. R. (2011), 'Local underwriter oligopolies and ipo underpricing', *Journal of Financial Economics* **102**(3), 579–601.

Loughran, T. & McDonald, B. (2011), 'When is a liability not a liability? textual analysis, dictionaries, and 10-ks', *The Journal of Finance* **66**(1), 35–65.

Loughran, T. & McDonald, B. (2016), 'Textual analysis in accounting and finance: A survey', *Journal of Accounting Research* **54**(4), 1187–1230.

Loughran, T. & Ritter, J. (2004), 'Why has ipo underpricing changed over time?', *Financial management* pp. 5–37.

Lowry, M., Michaely, R., Volkova, E. et al. (2017), 'Initial public offerings: A synthesis of the literature and directions for future research', *Foundations and Trends® in Finance* **11**(3-4), 154–320.

SDC New Issues Definitions (2007).
URL: *http://mergers.thomsonib.com/td/DealSearch/help/nidef.htm*

R-code Samples

Below is the code used for data collection, extraction and cleaning. This is only the part of the code that extracts and processes the Risk Factors section. The entire prospectus text and the MD&A sections are processed in the similar way, with minor alterations to the code presented below. The code for Robustness checks is also included after the Risk Factors section's code piece.

Risk_Factors.R

```
require(openxlsx)
require(xml2)
require(XML)
require(rvest)
require(pbapply)
links <- read.xlsx("Datasets in Use/Final Dataset.xlsx")

#####################################
# EXTRACT THE RISK SECTION
#####################################
# This is where the extracted risk sections are stored in individual files
risk.section.path <- "Ipo Files/risk_text/"
dir.create(risk.section.path, showWarnings = F)
clean.IPOs <- function(f){
  x <- paste(read_html(f), collapse = " ")
  text <- iconv(x, "utf-8", "ASCII", sub = "")

  # (1) Extract the CIK from file name
  cik <- gsub("https://www.sec.gov/Archives/edgar/data/([[:digit:]]+)/.*.","\\1",f)

  #**** RISK FACTORS
  # (2) Remove everything before setion "RISK FACTORS"
  text.risk <- gsub(".*<A NAME=\".{1,20}\"> *(</a>)? *.{0,100}(RISK +FACTORS *(<.+?> *
    )? *(</p>|</div>))", "\\2 ", text, ignore.case = T)

  # (3) The regex matches the specified section in the IPO only in about 60% of the
```

```
  #      cases. Here we test if the section starts with the right words. If that is
  #      the case, the script continues, otherwise it stops.
  if(grepl("^RISK", text.risk, ignore.case = F) == F){
    return()
  }

  # (4) Remove everything after section "RISK FACTORS" (works only
  #      if there is another section after "RISK FACTORS and if there is no
  #      HTML anchor (ie. "<A NAME...>") in the risk section text itself).
  text.risk <- gsub("(.+?)<A NAME=.*\".{1,20}\"> *(</a>)? *.{0,100}[[:upper:]]{3,}.*", "\\1",
    text.risk, ignore.case = T)

  # (5) Remove everything that is marked as <table> or <script>
  text.risk <- gsub("(<table>.*?</table>)|(<script>.*?</script>)", "", text.risk)
  text.risk <- gsub("<U\\+....>", "",text.risk)

  # (6) Remove html and convert html entities to UTF-8
  text.risk <- xml_text(read_html(text.risk))

  # (7) remove tabs (we want to use tabs later on as seperators in the csv-files)
  text.risk <- gsub("\\t", " ", text.risk)

  # (8) Store CIK and the cleaned text in a data frame
  cleaned <- data.frame(cik=cik,
                        risk_section = text.risk,
                        stringsAsFactors = F)

  # (9) Write CIK and cleaned text in a tab seperated file
  saveRDS(cleaned,
          file =paste0(risk.section.path, cik, "_risk_text.rds"))
  }

# Submit the files in " file .names.raw" to the function "clean.IPOs()"
# This procudes a list of data frames. The function "pblapply"
# prints a progress bar.
```

```
invisible (pblapply(urls, function(x) clean.IPOs(x)))

#####################################
# READING THE CLEANED FILES INTO A DATA FRAME
#####################################
file.names.csv <- list.files(risk.section.path, full.names = T)

require(pbapply)
risk.sections <- pblapply(file.names.csv, FUN = function(x) read.delim(file = x, stringsAsFactors =
    F) )

# rbind the dataframes in the list "risk.sections" into one data frame
library(data.table)
risk.sections.df <- rbindlist(risk.sections)

# reading in the indivudal dataframes with aggregate text
risk.section <- c()
for ( i in 1:213){
  risk.section <- c(risk.section,rbind(readRDS(file.names.csv[i])))
}
# combining everything into one dataframe
df <- data.frame(matrix(unlist(risk.section), nrow=213, byrow=T),stringsAsFactors=FALSE)
colnames(df) <- c("cik", "risk.text")

##### Dealing with not processed urls
z <- data.frame(cik=links$CIK_2,
                url=links$Prospectus,
                stringsAsFactors = F)
not.processed <- (z[z$cik %in% df$X1==F,])
risk.section.path.2 <- "IPO Files/risk_attempt2/"
clean.2 <- function(f) {
  x <- paste(readLines(f), collapse = " ")
  text <- iconv(x, "utf-8", "ASCII", sub = "")

  # (1) Extract the CIK from file name
```

```
  cik <- gsub("https://www.sec.gov/Archives/edgar/data/([[:digit:]]+)/.*.","\\1",f)

  #**** RISK FACTORS
  # (2) Remove everything before setion "RISK FACTORS"
  text.risk <- gsub(".*<a name=\".{1,50}\"> *.{0,100}(RISK +FACTORS *(<.+?> *)? *(</p>|</div>))", "\\2 ", text, ignore.case = T)
  # (3) Remove everything after section "RISK FACTORS" (works only
  #     if there is another section after "RISK FACTORS and if there is no
  #     HTML anchor (ie. "<a name...>") in the risk section text itself).
  text.risk <- gsub("(.+?)<A NAME=.*\".{1,20}\"> *(</a>)? *.{0,100}[[:upper:]]{3,}.*", "\\1", text.risk, ignore.case = T)
  # (4) Remove everything that is marked as <table> or <script>
  text.risk <- gsub("(<table>.*?</table>)|(<script>.*?</script>)", "", text.risk)
  text.risk <- gsub("<U\\+....>", "",text.risk)
  # (5) Remove html and convert html entities to UTF-8
  text.risk <- xml_text(read_html(text.risk))
  # (6) remove tabs (we want to use tabs later on as seperators in the csv-files)
  text.risk <- gsub("\\t", " ", text.risk)
  # (7) Store CIK and the cleaned text in a data frame
  cleaned <- data.frame(cik=cik,
                        risk_section = text.risk,
                        stringsAsFactors = F)
  # (8) Write CIK and cleaned text in a tab seperated file
  saveRDS(cleaned,
          file =paste0(risk.section.path.2, cik, "_risk_text.rds"))
}
invisible(pblapply(not.processed$url, function(x) clean.2(x)))

file.names.csv.2 <- list.files(risk.section.path.2, full.names = T)
# reading in the indivudal dataframes with risks text
risk2 <- c()
for ( i in 1:26){
  risk2 <- c(risk2,rbind(readRDS(file.names.csv[i])))
}
# combining everything into one dataframe
```

```
df.2 <- data.frame(matrix(unlist(risk2), nrow=26, byrow=T),stringsAsFactors=FALSE)
colnames(df.2) <- c("cik", "risk.text")

# binding two dataframes
risk.df <- rbind(df, df.2)
# saving the dataframe for future use
saveRDS(risk.df, file = "risk_sections.rds")

#####################################
# ANALYZING SENTIMENT OF THE TEXT
#####################################
#loading in the dataframe
risk.sections.df <- readRDS("risk_sections.rds")

# Analyzing sentiment of the text, storing it as a data frame
require(SentimentAnalysis)
sentiment.scores <- analyzeSentiment(risk.sections.df$risk.text)*100

# Copying scores to the risks section dataframe: the question here is whether the scores are going to
    be in the right order
risk.sections.df$negativityLM <- sentiment.scores$NegativityLM
risk.sections.df$sentimentLM <- sentiment.scores$SentimentLM

#storing necessary sentiment scores in a separate dataframe
z <- data.frame(cik=risk.sections.df$cik,
                negLM = sentiment.scores$NegativityLM,
                sentLM=sentiment.scores$SentimentLM,
                sentHE=sentiment.scores$SentimentHE,
                negHE=sentiment.scores$NegativityHE,
                sentGI=sentiment.scores$SentimentGI,
                negGI=sentiment.scores$NegativityGI)

saveRDS(z, file= "Risk section scores.rds")

#####################################
```

```
# RUNNING REGRESSIONS
####################################
library("stargazer")
rm(list=ls())
z <- readRDS("Dataframes/Risk section scores.rds")
y <- readRDS("Dataframes/regression_data.rds")
reg.data <- merge(z,y, by = "cik")

#saving the dataframe as xlsx
write.xlsx(reg.data,
           file =paste0( "risk section  scores_underpricing.xlsx"),
           sheetName= "Main",
           row.names = F,
           col.names = T)

#running a regression: underpricing 1 day ~sentiment
lm1.risk <- lm(underpricing.1day ~ negLM, data = reg.data)
lm2.risk <- lm(underpricing.1day ~ negLM + vc.dummy + amt_log+age_log + tech.dummy +size_log
    +spread, data = reg.data)

# Using "stargazer" package to export into LaTex
reg <- stargazer(lm1, lm2, title = "Risk Sections Negative LM Sentiment and Underpricing", align =
    T,
                 covariate.labels = c("Negativity LM", "VC binary", "Amount Filed", "Firm Age",
                                      "High-Tech binary","Firm Size", "Spread"),
                 dep.var.labels = "1st day Underpricing",
                 keep.stat = c("n","rsq", "adj.rsq"))

####################################
# DIFFERENT DEPENDENT VARIABLES
####################################
#running a regression: underpricing 30 days ~sentiment
lm_30.1 <- lm(underp.30days ~ negLM, data = reg.data)
lm_30.2 <- lm(underp.30days ~ negLM + vc.dummy + amt+age + tech.dummy +size +spread, data
    = reg.data)
```

```
# No significance in any of the specifications


#running a regression: underpricing 90 days ~sentiment
lm_90.1 <- lm(underp.90days ~ negLM, data = reg.data)
lm_90.2 <- lm(underp.90days ~ negLM + vc.dummy + amt+age + tech.dummy +size +spread, data
    = reg.data)
# No significance in any of the specifications

#running a regression: underpricing 180 days ~sentiment
lm_180.1 <- lm(underp.180days ~ negLM, data = reg.data)
lm_180.2 <- lm(underp.180days ~ negLM + vc.dummy + amt+age + tech.dummy +size +spread,
    data = reg.data)
# Significant results for the specification with the control variables

# Using "stargazer" package to export into LaTex
stargazer(lm_180.1, lm_180.2, title = "Risk Sections Negative LM Sentiment and Future 6 months
    Performance", align = T,
        covariate.labels = c("Negativity LM", "VC binary", "Amount Filed", "Firm Age",
                             "High-Tech binary","Firm Size", "Spread"),
        dep.var.labels = "1st day Underpricing",
        keep.stat = c("n","rsq", "adj.rsq"))
```

The code below is the sample of the code for Robustness checks.

Robustness_check.R

```
##############################
# ROBUSTNESS CHECK
##############################
require(xlsx)
require(openxlsx)
require(xml2)
require(XML)
require(rvest)
require(pbapply)
require(SentimentAnalysis)
require(stargazer)
```

```
rm(list=ls())
links <- read.xlsx("Datasets in Use/Final Dataset.xlsx")
y <- readRDS("Dataframes/regression_data.rds")
agg.scores <- readRDS("Dataframes/Aggregate scores.rds")
mda.scores <- readRDS("Dataframes/MDA sentiment scores.rds")
risk.scores <- readRDS("Dataframes/Risk section scores.rds")
reg.data.3 <- merge(risk.scores,y, by = "cik")
reg.data.1 <- merge(agg.scores, y, by = "cik")
reg.data.2 <- merge(mda.scores, y, by = "cik")

reg.data.1$negLM_log <- log(reg.data.1$negLM)
reg.data.2$negLM_log <- log(reg.data.2$negLM)
reg.data.3$negLM_log <- log(reg.data.3$negLM)

##########################################
# REGRESSION WITH A LOG INDEPENDENT VARIABLE
##########################################
#running a regression: underpricing~sentiment
lm1.ent <- lm(underpricing.1day ~ negLM_log, data = reg.data.1)
lm2.ent <- lm(underpricing.1day ~ negLM_log + vc.dummy + amt_log + age_log+ tech.dummy +size_
    log + spread , data = reg.data.1)
#running a regression: underpricing~sentiment
lm1.mda <- lm(underpricing.1day ~ negLM_log, data = reg.data.2)
lm2.mda <- lm(underpricing.1day ~ negLM_log + vc.dummy + amt_log + age_log+ tech.dummy +
    size_log + spread , data = reg.data.2)
#running a regression: underpricing 1 day ~sentiment
lm1.risk <- lm(underpricing.1day ~ negLM_log, data = reg.data.3)
lm2.risk <- lm(underpricing.1day ~ negLM_log + vc.dummy + amt_log+age_log + tech.dummy +size
    _log +spread, data = reg.data.3)
# both specifications result in significant coefficient for negativity LM of Risk Factors Section

stargazer(lm1.ent, lm2.ent, lm1.mda, lm2.mda, lm1.risk, lm2.risk,
          title = "Negative LM Sentiment and 6-month Underpricing", align = T,
          covariate.labels = c("Negativity LM", "VC binary", "log Amount Filed", "log Firm Age",
                               "High-Tech binary","log Firm Size", "Spread"),
```

```
            dep.var.labels  = "",
            column.labels   = c("Entire prospectus","MD&A section", "Risk section"),
            column.separate = c(2,2,2),
            model.numbers = F,
            dep.var.caption = "",
            keep.stat  = c("n","rsq", "adj.rsq"))

###################################
#SUB-SAMPLE WITH VC BACKED FIRMS
###################################
sub.sample.1 <- reg.data.1[reg.data.1$vc.dummy==1,]
sub.sample.2 <- reg.data.2[reg.data.2$vc.dummy==1,]
sub.sample.3 <- reg.data.3[reg.data.3$vc.dummy==1,]

lm1 <- lm(underpricing.1day ~ negLM, data = sub.sample.1)
lm2 <- lm(underpricing.1day ~ negLM+amt_log+age_log + tech.dummy +size_log+spread, data =
    sub.sample.1)
lm3 <- lm(underpricing.1day ~ negLM, data = sub.sample.2)
lm4 <- lm(underpricing.1day ~ negLM+amt_log+age_log + tech.dummy +size_log+spread, data =
    sub.sample.2)
lm5 <- lm(underpricing.1day ~ negLM, data = sub.sample.3)
lm6 <- lm(underpricing.1day ~ negLM+amt_log+age_log + tech.dummy +size_log+spread, data =
    sub.sample.3)
# both specifications result in significant coefficient for negativity LM of Risk Factors Section

stargazer(lm1, lm2, lm3, lm4, lm5, lm6,
            title  = "Negative LM Sentiment and Underpricing in VC backed firms", align = T,
            covariate.labels  = c("Negativity LM", "log Amount Filed", "log Firm Age",
                                  "High-Tech binary","log Firm Size", "Spread"),
            dep.var.labels  = "",
            column.labels   = c("Entire prospectus","MD&A section", "Risk section"),
            column.separate = c(2,2,2),
            model.numbers = F,
            dep.var.caption = "",
            keep.stat  = c("n","rsq", "adj.rsq"))
```