von Bodman, Nicolas

Article

The impact of prospectus language on IPO underpricing: A textual analysis of European IPOs

Junior Management Science (JUMS)

Provided in Cooperation with:
Junior Management Science e. V.

Suggested Citation: von Bodman, Nicolas (2024) : The impact of prospectus language on IPO underpricing: A textual analysis of European IPOs, Junior Management Science (JUMS), ISSN 2942-1861, Junior Management Science e. V., Planegg, Vol. 9, Iss. 4, pp. 1934-1963, https://doi.org/10.5282/jums/v9i4pp1934-1963

This Version is available at:
https://hdl.handle.net/10419/308472

Junior Management Science 9(4) (2024) 1934-1963

Junior Management Science

www.jums.academy
ISSN: 2942-1861

The Impact of Prospectus Language on IPO Underpricing: A Textual Analysis of European IPOs

Nicolas von Bodman

Technical University of Munich

Abstract

This study explores the impact of IPO prospectus language on the prevalent underpricing in European IPOs, using natural language processing techniques. Specifically, it investigates whether a relationship exists between litigious, negative, positive, and uncertain language, as well as the degree of document similarity and IPO underpricing. For this purpose, qualitative text data is converted into quantifiable metrics using modern analysis techniques. The study presents new methodological approaches to textual analysis. The results establish a clear relationship between underpricing and multiple dimensions of prospectus language and highlights unique features of European markets. These include specific disclosure obligations of various market segments and the different listing types available to issuing firms. The results of the variables related to sentiment analysis all reveal significant relationships. However, no robust evidence emerges for variables related to document similarity. Overall, the introduced methodological approaches offer enhanced explanatory power over traditional methods, effectively contributing to the explanation of the underpricing phenomenon in European markets.

Keywords: IPO; NLP; prospectus language; textual analysis; underpricing

1. Introduction

The initial public offering (IPO) landscape has experienced a significant shift over the past two decades, with a major decline in activity both in Europe and the United States. In Europe, the number of annual IPOs dropped from 380 per year between 1997 and 2007 to 220 per year between 2008 and 2018 (European IPO Task Force, 2020, p. 11). A similar picture is obtained from the U.S. market, where IPO activity has decreased by 50% since 1997 (Huang et al., 2023, p. 1). The reasons for this decline are manyfold. The observations can be attributed to the lower profitability of firms in the current business environment, leading to more consolidation (Gao et al., 2013). Another explanation lies in the exceptional rise of private markets, which allow companies to stay private by raising late-stage capital. This has the benefit of avoiding the scrutiny and governance regulations imposed on public companies (Ewens & Farre-Mensa, 2020). However, a consistent factor behind many explanations for this declining trend is the underpricing phenomenon, which describes one of the most controversial aspects of IPO research. Underpricing is related to the fact that, on average, companies have very high first-day returns, which can cost them multiple years of operating profits (Loughran & Ritter, 2002).

Numerous studies have attempted to explain this effect. Among the most prominent explanations is the uncertainty theory of Beatty and Ritter (1986), which links underpricing to the high degree of uncertainty surrounding the issuing firm and the valuation of the IPO due to the lack of reliable historical data. Another important theory is that of Benveniste and Spindt (1989), which links the phenomenon to the compensation investors demand for revealing private information about the offering to the issuing firm. The third important theory of underpricing addressed in this thesis is the legal liability theory by Lowry and Shu (2002), which claims that underpricing is used as a protection against lawsuits the firm and its underwriters might face in the aftermarket. Previous research has tended to focus on exploring different explanations for the phenomenon. However, more recently the research community has shifted to finding significant predic-

DOI: https://doi.org/10.5282/jums/v9i4pp1934-1963

tors of underpricing and attributing their findings to prevalent theories. An important research area that has emerged with enhanced analytical capabilities, is the area of textual analysis. In this context, studies have established a link between underpricing and the language in the IPO prospectus. The prospectus is a document containing information required to evaluate the offering, such as a business plan, risk factors and comprehensive financial data. Initial studies were conducted using U.S. samples. Hanley and Hoberg (2010) explored the impact of information revelation on underpricing finding that more unique information disclosure leads to reduced levels of underpricing, while higher document similarity has the opposite effect. In a related study, Hanley and Hoberg (2012) assess document similarity in the context of prospectus revisions during the bookbuilding period and how this affects underpricing. Loughran and McDonald (2013), applied a dictionary-based sentiment analysis, which established a link between underpricing and uncertain sentiment within the prospectus. A similar study of the Chinese market was conducted by Guo et al. (2022), confirming the results for an international market. Ferris et al. (2013) examine the impact of conservative language in a prospectus. Their findings indicate a positive significant relationship between underpricing and document conservatism. These existing studies are based on simple word count methods. However, recent advancements in natural language processing (NLP) have augmented the possibilities of researchers to develop text-based quantitative measures of economic variables. A prominent technology is word embeddings, which are geometric representations of word meanings, that are used to translate textual data into numerical format (Seegmiller et al., 2023, p. 1). The first generation of word embedding models, word2vec, GloVe, fastText, are based on the works of Mikolov et al. (2013), Pennington et al. (2014), and Mikolov et al. (2017), respectively. They represent each word as a fixed vector representation. Recent advancements in model architecture enable context-dependent word and sentence representations, as in the models, BERT of Devlin et al. (2018) and Sentence-BERT (SBERT) of Reimers and Gurevych (2019).

While the relationship between prospectus language and underpricing has been studied previously, the application of modern NLP techniques to this problem remains relatively underexplored. This study addresses this limitation by developing methodologies for sentiment and similarity analysis using novel word embedding techniques. The primary research question is whether the textual analysis of IPO prospectuses can explain the underpricing of European IPOs. Specifically, it is investigated if a relationship can be found between litigious, negative, positive, and uncertain sentiments, as well as the degree of document similarity and underpricing, using both the fastText and SBERT models. The study also evaluates the potential of novel NLP models in analyzing these documents. For robustness, results are benchmarked with traditional word count methods. The idea of the methodology used is to leverage the capabilities of embedding models to better understand text semantics. Therefore, the proposed methodology in this research should provide an improvement over traditional word count methods. In addition, the proposed methodology is easy to implement, unlike many other approaches to sentiment analysis, which typically require large amounts of labeled data. This data is difficult to obtain as it requires expert judgment, which especially for complex financial contexts is rarely available. The suggested approach for sentiment analysis uses the same word lists as Loughran and McDonald (2013), and modifies them for optimized use in combination with word embedding models. The updated methodology for similarity analysis follows the concept presented in Breitung and Müller (2022), which suggests to measure document similarity by comparing pairwise sentence similarities. The obtained results suggest that the methodologies provide valid and coherent insights. Furthermore, this study contributes significantly to the existing body of academic literature in the field of textual analysis and IPO underpricing by corroborating the documented phenomenon using a European dataset. To this end, the dataset initially introduced by Kaserer and Treßel (2023) is utilized. This dataset encompasses the listing documents and important firm and offering characteristics for 745 European listings, including both IPOs and private placements over the period from 2016 to 2022. It is important to note that, in the ensuing paper, the term 'IPO' is frequently employed as an umbrella term for listings, like the term 'prospectus', which is used as a generic term for listing documents.

The findings of this study can be summarized as follows. First, a significant positive relationship between underpricing and the use of uncertain language is evident in the European sample, confirming the findings from previous studies for different markets. Second, through the application of the refined methodology, a significant relationship is observed between litigious sentiment in IPO prospectuses and underpricing. This observation aligns with the legal liability hypothesis. Third, drawing inspiration from the concept of prospectus conservatism, it is found that neutral prospectuses - those that avoid both positive and negative language - are significantly associated with lower levels of underpricing. This finding resulted again from the use of the embedding-based approaches. Fourth, in the similarity analysis, no statistically significant evidence is discovered to suggest that prospectuses, which disclose less new information, affect underpricing. Finally, the analysis demonstrates that the methodologies developed in this study yield superior results in terms of both the number of significant coefficients and the ratio of explained variance. Overall, this study enriches the existing literature by providing new insights into the relationship between textual variables in IPO prospectuses and underpricing, specifically within the European context. The advanced methodologies deployed contribute to a more thorough understanding of these dynamics.

The structure of the paper is organized as follows. Section 2 contains a thorough literature review, delving into the IPO process, theories of underpricing, and relevant research in the area of textual analysis, along with a presentation of the hypotheses derived therefrom. Subsequently, Section 3

describes the dataset and illustrates the development of the methodologies for sentiment and similarity analyses. This is succeeded by Section 4, which presents detailed descriptive statistics. In Section 5, the outcomes of the regression analyses are disclosed and interpreted. Lastly, Section 6 encapsulates a conclusion that recapitulates the findings, highlights the limitations, and outlines the future areas of research.

2. Literature Overview

2.1. IPO Process and Characteristics

2.1.1. The Dynamics of IPOs

IPOs are highly complex financial transactions. The process requires the collaboration of numerous parties, which prepare the issuing firm for its stock market debut and conduct crucial pre-market due diligence. The high complexity and the costs associated with a stock market listing, which usually involve fees of around 7% of gross proceeds charged by investment banks, require careful evaluation of managers (Lowry et al., 2017, p. 193). However, despite the declining trend in IPO activity and the mentioned caveats, going public remains a vital component in the financial strategy of most emerging firms. The primary motivators behind pursuing an IPO include gaining access to capital for investment activities, exploiting attractive valuations and even market inefficiencies (i.e., overvaluation), adjusting more flexibly the firm's capital structure, and providing existing shareholders with an opportunity to sell shares in a liquid secondary market (Lowry et al., 2017, p. 8-10). As a first step after deciding to pursue an IPO, the issuing company typically engages a group of investment banks known as underwriters, who oversee the structuring of the offering. Once the underwriters have been selected and their roles within the syndicate are established, subsequent steps involve determining the types of shares to be issued, the offer size, and the mechanism for selling those shares. The offering might consist of primary shares (i.e., newly issued stock sold for the first time) or secondary shares (i.e., existing shares sold by current investors), but commonly as a combination of both. The total volume of sold stocks is decided based on the company's future investment plans and the liquidation requirements of existing shareholders. In most cases, the issuance is conducted as a firm commitment IPO, where the underwriter guarantees the sale of all the stock at the offer price. The underwriters then purchase the shares from the company at a small discount prior to the offering and subsequently sell them at the offer price on the market. For smaller transactions, IPOs can also be executed on a best-effort basis, in which the underwriters do not guarantee the stock's sale (Berk & DeMarzo, 2019, p. 879). For determining the offer price, underwriters have several options, but the predominant method in international markets is the book building approach. In this approach, underwriters organize meetings with pre-selected institutional investors to engage them in price discovery by submitting bids for IPO shares within a predefined price range. A unique aspect of book building is that underwriters possess the authority to both set the price and determine the allocation of IPO shares (Huibers, 2020, p. 117). Auctions and fixed-price offerings are the remaining pricing mechanisms. Both give underwriters no discretion in determining the allocation of shares, but while auctions provide also no room to determine the offer price, in fixed-price issues the offer price is directly set by the underwriters (Torbira & Oki, 2017, p. 33). The IPO process further necessitates that companies register with their respective national listing authorities. In accordance with market-specific regulations and exemptions, companies are generally required to produce a prospectus containing comprehensive information pertinent to investors interested in the offering, thereby enabling them to conduct an informed assessment prior to participation. These requirements differ greatly depending on the setting of the IPO. The following chapter will explain the different considerations for companies when choosing the best-suited exchange and listing type.

2.1.2. IPO Considerations - Exchange and Listing Types

Exchange Types

Most European stock exchanges consist of a main market and one or more second-tier markets catering to specific firm classes. Historically, these secondary markets included seasoning markets - common before the year 2000. They provided smaller firms with a venue to go public before potentially transitioning to the main market if successful. The "New Markets", which experienced a swift rise and subsequent decline around the year 2000, constituted another market segment. This segment fostered the IPO boom of high-tech firms during the dot-com bubble (Vismara et al., 2012, p. 354). Presently, markets are primarily classified into regulated main markets and exchange-regulated, or in other words "unregulated" secondary markets, which are known under the term multilateral trading facility (MTF). Regulated markets are fully governed by EU law and the respective national legislation and managed by a designated market operator. MTFs can be maintained by market operators, but also investment firms. After recent regulatory changes, MFTs now include a newly created subtype called SME growth market (SME GM), which require a majority of admitted issuers to be classified as SMEs. The intention of regulators behind the introduction of SME GMs is to increase investor appeal and further reduce administrative burdens for SMEs seeking access to capital markets (Kaserer & Treßel, 2023, pp. 5-6). The significance of secondary markets becomes apparent when examining historical transaction volumes. Vismara et al. (2012, p. 353) report that, during the period from 1995 to 2009, approximately 77.5% of IPOs occurred on second-tier markets. For their sample of listings on European exchanges with a registered SME GM market segment, Kaserer and Treßel (2023, p. 6) report a successful start for SME GMs. After the first market was established in 2018, already in 2021 80% of all listings took place on an SME GM. These high transaction volumes can be attributed to the less stringent regulatory requirements imposed on issuers. This be-

comes evident as more than 70% of second-market issuers do not fulfil all the requirements of the respective main market, as found by Vismara et al. (2012, p. 366). In a study focused on the UK Alternative Investment Market (AIM), one of the most notable exchange-regulated markets, Doukas and Hoque (2016, p. 387) investigated the reasons behind firms opting for unregulated markets. Employing a more recent dataset, they demonstrated that only 50.5% of firms listed on AIM failed to satisfy the criteria of the main market, a figure much lower than the 67.2% reported for the identical markets by Vismara et al. (2012, p. 366). Doukas and Hoque (2016, pp. 402-403) further contend that firms primarily select a market platform that aligns with their investment and financing objectives. Companies opting for listings on main markets often exhibit heightened merger and acquisition activity, necessitating liquid share trading and consequently accepting increased regulatory oversight and scrutiny. In contrast, firms selecting secondary markets, such as AIM, often are loss-making, resulting in a greater dependence on seasoned equity offerings for financing. Notably, secondary markets attract smaller and younger companies due to their lower listing and ongoing flotation costs. The mentioned differences between exchanges require prospective issuers to analyze firm characteristics and their needs thoroughly to make the best choice on which exchange type to list. Another important consideration of this evaluation is the right choice of listing type, which will be further explained in the next chapter.

Listing Types

In Europe, the dominant listing types available to issuing companies are IPOs and private placements. The IPO process has already been described in chapter 2.1.1. Private placements differ from IPOs in that they only involve the sale of shares to a select group of qualified investors, which in the case of a secondary market listing can avoid the requirements for extensive regulatory filings. In this context, only an offering memorandum is required, which includes company information and offer details to be distributed to the targeted group of investors. The details of the offer are negotiated on an individual basis with each investor and are finalized with the signing of a purchase agreement. In contrast, IPOs are open to an unlimited number of both institutional and retail investors and consequently require a higher level of regulatory oversight (Geddes, 2003, pp. 129-132). During the negotiation phase of a private placement, the maximum number of qualified investors that can potentially be contacted is capped at 150, as defined in Article 1, Paragraph 4e of the new prospectus regulation.[1] This number may vary depending on member state regulations. Companies often limit the addressees of their offerings in the primary market for several reasons. For the company, the main advantage of this approach is that offerings on MTFs are not regulated according to EU legislation if they do not include a public offering. On secondary markets, the national listing authorities (e.g. BaFin) are not required to approve a prospectus when the listing is without a public offer (Vismara et al., 2012, p. 354). In contrast, in regulated markets, for the admission of securities a registration document has to be provided by the issuers, irrespective of whether the transaction involves an IPO or a private placement (Kaserer & Treßel, 2023, p. 7). By limiting the addressees of their offerings, companies can avoid the extensive regulatory requirements, reduce costs associated with producing and publishing a prospectus, and accelerate the process of raising capital. Additionally, private placements allow for more flexibility in negotiating terms and conditions, catering to the specific needs of both issuers and qualified investors. There are also several motives for initial investors to push against an IPO. According to Torbira and Oki (2017, p. 34) initial investors can better exercise control of the firm thanks to the discrete nature of private placements, but also for the sale of larger quantities of secondary shares during the offering this type of listing is preferred among investors. A pivotal aspect of the decision for the right listing type is the prospectus exemption rule. Thus, the subsequent chapter will elaborate on this critical aspect of the IPO process, discussing the rationale of preparing a prospectus, the requisite contents, and the evolving regulatory landscape surrounding these documents.

2.1.3. Required Disclosure – Listing Documents

An important part of the IPO process is drafting the prospectus, which is a legal document providing information about the offering. A full prospectus is required when securities are offered to the public or when securities are listed on a regulated market, provided that no exemption rules are applicable to the specific offering. The prospectus should provide enough information, allowing investors to make informed decisions regarding their participation in the offering (BaFin, 2023). Omissions of material information or inaccurate statements within the prospectus can lead to shareholder litigation. Such scenarios are particularly prevalent in the United States, where they are governed by the Securities Act of 1933 (Geddes, 2003, p. 140). The European prospectus regulation also sets high information standards for issuers. Yet, the explicit contents of a prospectus are suspectable to the regulation of the respective jurisdictions. However, the International Organization of Securities Commissions has established some guidelines for international harmonization of prospectus content, aiming to enhance comparability across markets. Essential elements of a prospectus are, as described by Geddes (2003, pp. 97-99):

1. A summary of the offering: Presenting business, details of shares being offered, use of proceeds, listing information, and key financial data.

2. A management discussion and analysis: Assessing the company's revenues, expenses, and capital expenditures by comparing the latest year with two prior years.

[1] Regulation (EU) 2017/1129 of the European Parliament and of the Council of 14 June 2017

3. The company's financial statements: Typically featuring three years of audited historical records.
4. A risk factors section: Disclosing potential risks to protect the issuer from investor lawsuits and inform investors about the possible hazards associated with purchasing shares.

In an effort to revive the IPO markets and make them more appealing, particularly for smaller firms, regulators have introduced new reforms in response to the declining IPO activity. In the United States, the Jumpstart Our Business Startups (JOBS) Act was enacted in 2012, while the European Union responded in 2017 with the abovementioned prospectus regulation (Kaserer & Treßel, 2023, p. 13). The new EU law features several key elements, including adjustments to the exemption rule and the introduction of the EU Growth Prospectus. Consequently, the regulation raises the total proceeds threshold, below which a prospectus is not required, to €1 million. Furthermore, it provides member states with the option to increase this threshold up to €8 million (Kaserer & Treßel, 2023, p. 8). The new EU growth prospectus is a simplified version of the full prospectus, applicable for SMEs conducting initial offerings on an unregulated market with planned proceeds above the exemption threshold. Kaserer and Treßel (2023, pp. 11-13) compare content requirements for the EU growth prospectus with those of full prospectuses, and admission documents. The latter are listing documents governed by the discretion of their respective MTF without the influence of EU legislation. The EU growth prospectus requires less content disclosure than the full prospectus, which is consistent with its primary objective of reducing financial burdens and bureaucratic obstacles for companies. However, when compared to admission documents, it imposes more comprehensive content requirements. In summary, prospectus regulation significantly influences a company's decision on which market to go public due to the high associated costs. This is why recent reforms were aimed at balancing reduced financial and bureaucratic burdens with investor protection and global harmonization.

2.2. Underpricing of IPOs

2.2.1. Empirical Evidence and Consequences of Underpricing

Underpricing is one of the most controversial aspects of IPOs. It describes the phenomenon that most IPOs have high positive returns on their first day after floating on the market. This characteristic and its underlying causes have been vastly studied in academic research since the 1970s (Ljungqvist, 2007; Ritter & Welch, 2002). Empirically evidence has been provided for most markets. For the U.S. for example, Hanley (2017) presents an analysis of the IPO market from 1980 to 2015, demonstrating consistent positive underpricing during this period, albeit with extreme fluctuations. The most pronounced underpricing occurred during the dot-com bubble in the years 1999 and 2000, with average first-day returns reaching 71.1% and 56.3%, respectively Hanley (2017, p. 8). Table 1 provides numbers for European markets presented in Ritter (2023), which all show positive average first day returns. The highest underpricing value was observed in Greece, where the average underpricing exceeded 50% between 1976 and 2013. Conversely, the lowest value of 6.2% was recorded in Austria during the period from 1971 to 2018. In recent years, alternative methods of going public have emerged with the aim of circumventing the high costs associated with traditional stock market listings. A remarkable surge in the popularity of special purpose acquisition companies (SPACs) was observed in 2020 and 2021, with nearly 75% (i.e. 861 out of 1157) of all such transactions between 2010 and 2022 occurring within these two years (Huang et al., 2023). A SPAC is a shell-company established by a financial sponsor, which raises funds through an IPO and commits to utilizing the proceeds to acquire a private company, thereby taking it public (Huang et al., 2023, pp. 14-15). As reported by Klausner and Ohlrogge (2023, pp. 112-113) the popularity of SPACs is already fading. Reasons are the substantial costs involved, on average 36% of gross proceeds, which in this case investors are required to bear, and low post-merger performance (on average -62% as of December 2022).

Another alternative to traditional IPOs that emerged as an imminent result to high underpricing, are direct listings. First promoted in the U.S. in 2018 by software company Spotify Technology, direct listings are also an option on European exchanges. A direct listing differs from an IPO in the regard, that no underwriters are involved in the offering process, and that the company directly lists its existing shares without an offer price on an exchange. Direct listings have not yet gained widespread adoption, with a limited number of transactions to date. One potential explanation for this phenomenon is the restricted suitability of direct listings, which primarily cater to companies with strong brand recognition and solid financials with no direct need for raising primary proceeds. Nonetheless, the ongoing public discourse surrounding this form of listing has drawn attention to the role of underwriters in the underpricing of IPOs (Huang et al., 2023). The following chapter will explore established theories, explaining the persistent occurrence of underpricing in IPOs, which has been a subject of extensive research and debate in the field of finance. Several theories have been proposed to explain this phenomenon, with varying degrees of emphasis on information asymmetry, institutional aspects, and behavioral factors. The following chapter will focus on three prominent theories of underpricing, namely the winner's curse, the information revelation, and the lawsuit avoidance hypothesis. It should be noted that other theories, such as the signaling theory and behavioral theories like the principal-agent theory, might also provide additional insights into the underpricing phenomenon.

Table 1: Average IPO underpricing in European countries. Taken from Ritter (2023)

Country	Sample Size	Time Period	Avg. Initial Return
Austria	106	1971-2018	6.2%
Belgium	154	1984-2017	11.0%
Bulgaria	9	2004-2007	36.5%
Cyprus	73	1997-2012	20.3%
Denmark	190	1984-2021	7.6%
Finland	244	1971-2021	14.5%
France	904	1983-2021	9.4%
Germany	840	1978-2020	21.8%
Greece	373	1976-2013	50.8%
Ireland	38	1991-2013	21.6%
Italy	413	1985-2018	13.1%
Netherlands	245	1983-2021	12.0%
Norway	368	1984-2021	10.3%
Poland	359	1991-2022	12.4%
Portugal	33	1992-2017	11.5%
Spain	204	1986-2021	9.5%
Sweden	442	1980-2021	28.2%
Switzerland	173	1983-2021	24.6%
United Kingdom	5,309	1959-2020	15.7%
		Average Europe	**16.9%**

2.2.2. Theories of Underpricing

Information Asymmetry and Uncertainty

Fundamental research on information asymmetries as a reason for underpricing was conducted by Rock (1986, pp. 188-189). He assumes the existence of two types of investors: informed and uninformed. Uninformed investors participate in an issue without any private information, whereas informed investors possess perfect information and solely subscribe to underpriced issues. This means that underpriced issues get rationed more often than overpriced issues. Therefore, the probability of receiving an allocation of underpriced shares is inversely related to the degree of underpricing. This implies that for issues with substantial underpricing uninformed investors are crowed out by informed demand, whereas for overpriced issues, uninformed investors obtain a full allocation. However, in general, the demand from uninformed investors is essential, as otherwise, total demand would be insufficient to fill many offerings. Thus, to attract uninformed investors, who are requiring a positive conditional return from their investment, "the issuer must price the shares at a discount, which can be interpreted as compensation for receiving a disproportionate number of overpriced stocks" (Rock, 1986, p. 193). On the other hand, informed investors require returns that equal the costs of becoming informed in the first place (Ljungqvist, 2007, p. 389). Therefore the central aspect of underpricing in Rock's (1986) model is the heterogeneity of investors participating in the offering. Michaely and Shaw (1994, pp. 289-290) investigate this theory by examining a unique type of IPOs, namely master limited partnerships (MLPs). MLPs are of particular interest because institutional investors tend to avoid this market due to unfavorable tax implications. Thus, creating a homogeneous market comprised predominantly of uninformed retail investors. In their study, the researchers discover marginally negative levels of average underpricing for MLP issues, while disclosing positive first day returns for regular IPOs. A finding that supports the winner's curse theory of Rock (1986). Beatty and Ritter (1986, pp. 215-217) argue that the winner's curse problem is at least partially caused by ex-ante uncertainty, which makes the expected payout for informed investors less predictable and thus requires higher underpricing to attract them. The researchers further argue that it is an important role of investment banks to enforce an underpricing equilibrium. Beatty and Ritter (1986, p. 229) compare the decision to invest in information production by investors to investing in a call option on the issuing firm that would be exercised if the estimated price is higher than the quoted offer price. The price of a call option increases with implied volatility, i.e. ex-ante uncertainty. Consequently, they argue that as uncertainty increases, more investors choose to become informed and thus the underpricing must be higher. They use the number of uses of proceeds (regulation-specific indicator) and the inverse of gross proceeds (firm size indicator) as proxies for ex-ante uncertainty. A positive relationship between both variables and underpricing confirms the theory of ex-ante uncertainty (Beatty & Ritter, 1986, p. 223). According to Loughran and McDonald (2013, p. 13), the ex-ante uncertainty proxies are due to regulatory changes and different investor sentiment no longer relevant indicators. The content of IPO prospec-

tuses can help to derive new proxies for this purpose. In the next chapter, underpricing will be discussed in the context of price discovery and costs of information production.

Information Revelation and Disclosure

Another central theory of underpricing was developed by Benveniste and Spindt (1989). Similarly, to the winner's curse model, the information revelation theory assumes that information asymmetries exist. However, in this case, asymmetries exist between the informed investors and the uninformed issuer. Investors have private information about the offering, which they without additional incentives would keep to themselves to profit from in the post-IPO period. As a consequence, underwriters, who conduct the bookbuilding process, use underpricing to incentivize investors to disclose this information truthfully. The discretion to allocate shares to those investors regularly participating in the underwriters' transactions is used to effectively minimize the required underpricing, through preferred allocation to those investors (Benveniste & Spindt, 1989, pp. 344-345). Another study focusing on the process of information revelation during bookbuilding comes from Sherman and Titman (2002). The researchers argue that issuers with a great need for pricing accuracy, which often are riskier, smaller firms with a great need for subsequent seasonal offerings, are paying for price discovery with higher underpricing. In this regard, firms must weigh the costs of engaging investors in information production or choosing to disclose more information by themselves. Depending on the degree of outsourcing of information production, the required level of underpricing to compensate investors will vary (Hanley, 2017; Hanley & Hoberg, 2010). The default option to disclose information to investors is through listing documents. Depending on the type of offering this is the prospectus or the respective admission document. The issuing firm and its underwriters have the chance to engage in premarket due diligence and disclose this information in the prospectus, thereby reducing the levels of underpricing. Hanley and Hoberg (2010) examine the consequences of informative disclosures within prospectuses. Their findings indicate an inverse relationship between underpricing and unique prospectus contents. This underscores the trade-off between the expensive production of information during due diligence and the revelation of information by investors during the bookbuilding process. In a recent study, Jenkinson et al. (2018, pp. 2305-2309) provided empirical support for the information revelation hypothesis by investigating underwriter practices using typically undisclosed data. The researchers analyzed private bookbuilding data for 410 IPOs managed by UK-based banks from January 2010 to May 2015. Their research examined investor behavior and its influence on allocations, focusing on three bid characteristics – price limitations, early submission, and revisions during bookbuilding – along with investor participation in pre-bookbuilding meetings and any potential investor-bank relationships. The findings suggest that price-sensitive bids generally received higher allocations. Furthermore, investors who participated in bookbuilding meetings often received greater allocations of underpriced shares. This supports the notion that investors, who choose to disclose information are rewarded with underpriced shares. This process, however, also has faced criticism due to persistent allegations of reciprocal arrangements between bankers and investors, which are based on a lack of transparency inherent in the bookbuilding process. Nevertheless, it can be acknowledged that using underpricing might serve as a useful tool for companies with high costs of information production. Hence, a viable proxy for underpricing could be the quantity of new information presented in the prospectuses, as it mirrors the balance between self-produced information and information sourced from investors. So far theories of underpricing dealt primarily with information asymmetries. This might not fully explain the pervasive underpricing phenomenon observed in the market. Thus, the next chapter will introduce a notable theory – the lawsuit avoidance theory. This theory explores the concept of using underpricing as a form of insurance against potential legal risks.

Lawsuit Avoidance

A fundamental component in theories explaining underpricing beyond the view of information asymmetry is the lawsuit avoidance hypothesis, which goes back to a study by Tinic (1988). It states that underpricing is used by underwriters to protect themselves from lawsuits in the post-IPO period. This litigation risk stems from the fact that investors can bring legal claims against the issuers for wrongful statements or omissions in IPO prospectuses. Underpricing is an effective tool to reduce the expected value of potential legal liabilities thanks to two aspects. First by decreasing the likelihood that investors will lose money on their investment, it helps to reduce the probability of investors engaging in legal action against the issuers (i.e., lawsuit deterrence). Second, underpricing affects the maximum recoverable damages (i.e., lawsuit insurance), as possible claims for reparations are usually limited to the IPO offer price (Tinic, 1988, pp. 797-800). Drake and Vetsuypens (1993, pp. 68-70) analyze data from 93 IPOs between 1969 and 1990, for which issuers were sued for inadequate prospectus disclosure. To create an artificial experiment setting, they compare these IPOs with a sample of 1,114 IPOs from 1983 to 1987 used in Muscarella and Vetsuypens (1990). After combing both data sets, they group the total sample into three categories – overpriced, underpriced and more than 10% underpriced. In contrast to the lawsuit avoidance theory, no discernible differences were identified within the subsets related to lawsuit probability. The researchers further argue that it is rather medium to long-term share price declines that affect the probability of a lawsuit. Referring to the fact that investors who buy the stock in the secondary market also have the possibility to invoke legal actions, Drake and Vetsuypens (1993, p. 72) conclude that initial underpricing plays no decisive role for the litigation risk and thus revoke the lawsuit avoidance hypothesis. Concerns about the endogeneity of the study were

raised by the authors themselves. The true relationship between underpricing and litigation might not be adequately addressed because the ex-ante risk of a company being sued is neglected by the sample selection design. Addressing these concerns, Lowry and Shu (2002, pp. 322-326) examine the relationship between underpricing and litigation risk. The measurement of litigation risk is complex, as it is influenced by two interrelated factors: the level of insurance and the firm's intrinsic litigation risk. Specifically, underpricing the offering can lead to a decreased likelihood of lawsuits, while higher intrinsic risk tends to correlate with stronger underpricing. This dual relationship can distort results when analyzed through standard Ordinary Least Squares (OLS) methods. To address this issue, the researchers have proposed a two-stage estimation technique. In the first stage, initial return and litigation risk are regressed on exogenous predictors. In the second stage, the predicted values from the initial stage are used as explanatory variables to predict the other respective variable. They find that firms with higher litigation risk underprice their IPOs by greater amounts, providing support for the insurance effect component of the litigation-risk hypothesis. They also find that firms that engage in more underpricing significantly lower their litigation risks, especially for lawsuits occurring closer to the IPO dates, providing support for the deterrence effect of underpricing. Their results emphasize the importance of controlling for endogeneity in studying the relationship between underpricing and litigation risk.

Geographically, studies of litigation risk and underpricing are largely focused on the U.S. There, investors can bring legal action in the form of class action lawsuits against underwriters in relation to Section 11 of the Securities Act of 1933, which requires issuers to reveal only accurate information without material omissions (Drake & Vetsuypens, 1993, p. 65). In Europe, shareholder litigation through class action lawsuits is less common, despite many countries having laws that allow for collective shareholder action. In practice, only a few such lawsuits have been brought to court (Allianz Global Corporate & Specialty SE, 2020). However, recent regulatory initiatives indicate a shift towards a more progressive direction. Article 11 of the prospectus regulation mandates comparable standards for information disclosure to be implemented by national regulators, and a new directive on representative actions has been introduced.[2] Lin et al. (2013) conducted a comprehensive cross-country study to investigate the relationship between litigation risk and IPO underpricing, and how it is influenced by a country's legal environment. By using a 40-country sample, they adopted legal classifications from La Porta et al. (2008) and considered various country-level variables to measure litigation risk. Their findings supported the lawsuit avoidance hypothesis in a cross-country setting, revealing that IPOs in countries with higher litigation risk experience higher underpricing levels. The study further indicated that IPOs in common law countries, which are representative of the Anglo-Saxon regions and show more litigation risk, exhibit significantly greater underpricing compared to those in civil law countries, which are representative of the law systems of Continental European countries (Lin et al., 2013, pp. 66-69). In conclusion, the lawsuit avoidance hypothesis provides a compelling explanation for underpricing in IPOs across various legal environments. A notable critique against the lawsuit avoidance theory comes from Abrahamson et al. (2011, p. 2073), who argue that litigation costs account for only a minor fraction of 0.58% of gross proceeds in a large sample of IPOs. This limited magnitude of legal risk, they contend, cannot fully explain the underpricing phenomenon. However, this argument overlooks the endogeneity concern of actual litigation costs if underpricing were not employed and fails to consider potential reputational and organizational damages resulting from a lawsuit. The underpricing phenomenon remains a dominant topic in IPO literature. Its widespread occurrence across various settings complicates the attribution of a full explanation to a single theory. The subsequent chapters will discuss a set of methodologies, grouped under the term textual analysis, which can be utilized to extract information from IPO prospectuses for predicting underpricing.

2.3. Textual Analysis of Financial Documents

2.3.1. Sentiment Analysis

Traditionally, financial research has focused predominantly on numerical data. However, with the emergence of NLP and related analytical methods, the scope of studies has expanded to include more qualitative information. This field of research is labelled as textual analysis and encompasses methods, which are used to transform qualitative data into quantifiable metrics. Efficient tools for this purpose are domain-specific word lists. For the financial domain, the most important examples are the Loughran and McDonald (2011) word lists (LM word lists). The LM word lists include negative, positive, uncertainty, litigious, strong modal, and weak modal sentiment word lists. This method allows comparing word counts or percentage-of-words (POW) of different documents to gauge document tone. In a subsequent study Loughran and McDonald (2013), use these word lists to study the relationship between first day returns and language in S1 filings (i.e., the first SEC filing in the IPO process, also containing the initial version of the prospectus). Loughran and McDonald (2013, p. 2) further argue that IPOs characterized by a higher frequency of uncertain words are more challenging for investors to value. Consequently, the ratio of words with uncertain sentiment should act as a direct proxy for ex-ante uncertainty, which is linked to higher underpricing according to the theory of Beatty and Ritter (1986). In a sample of 1,887 U.S. IPOs that occurred between 1997-2010, the researchers find evidence supporting their initial conjecture, demonstrating a significant positive relationship between underpricing and document tone. Significant relationships are also discovered for uncertain, weak modal and negative word frequencies (Loughran & McDonald, 2013,

[2] Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020

pp. 4-8). A related study by Ferris et al. (2013) examines the impact of conservative language in a prospectus. They hypothesize that a prospectus filled with conservative language may appear more credible and reduces the risk of litigation from dissatisfied investors. On the other hand, excessive use of conservative language reduces the interest of investors in the offering and thus requires more underpricing as an incentive. To research this conjecture, they apply different negative sentiment word lists, including the respective LM word list. The findings indicate a positive significant relationship between underpricing and document conservatism, which is more pronounced for technology firms (Ferris et al., 2013, p. 995). For a similar study of the Chinese market, Guo et al. (2022, p. 2) compiled Chinese versions of the LM word lists to test the relationship between underpricing and prospectus sentiment. In addition to the ex-ante uncertainty proxy, the researchers establish litigious words as an ex-ante litigation risk proxy. Based on a dataset with 1917 IPOs from 2009 to 2018, their findings are similar to those of Loughran and McDonald (2013) - negative and uncertain sentiments show a positive and statistically significant relationship with underpricing (Guo et al., 2022, pp. 6-7). Brau et al. (2016, pp. 3-5) analyze the relative frequency of positive and negative strategic words in prospectuses. To compile these two word lists, they first collected a set of strategy-related words. Then, they conducted a survey among MBA students to rate the words as positive or negative in terms of business strategy. By employing these word lists, they discovered that a higher frequency of positive words (and a lower frequency of negative words) was associated with a larger first-day return. Since the process of manually generating word lists can be quite cumbersome, new technologies can be utilized. Das et al. (2022) demonstrate how pre-trained fastText word embeddings, can be employed to create financial lexicons for sentiment analysis tasks. The suggested approach extends the possibilities for researchers, which are bound to a limited set of sentiments defined by published conventional word lists. They find that the generated word lists consist of words that are notably relevant for the respective financial concepts and show similar performance to the manually selected word lists, such as the LM word lists. Sehrawat (2019) suggests using word embedding vectors to compute the similarity between financial documents and the LM word list sentiment categories. In the following chapter, another important component of textual analysis will be examined: the analysis of similarities between documents, which is used to describe the amount of new information that is contained in the document.

2.3.2. Similarity Analysis

Extracting document similarities may serve many functions in the financial context. It can be utilized to convey crucial information, such as document sentiment but also the degree of novel information contained within the text. While humans can easily discern whether two texts are similar, this task is more complex for machines. It involves the translation of textual data into numerical representation for automatic processing (Breitung & Müller, 2022, p. 1). Numerical representation can be obtained through established methods like bag-of-words (BOW), or through the use of word embeddings (e.g., fastText and BERT), which are more recent technological advancements in the field of NLP.

Hanley and Hoberg (2010) conducted an important study in the context of IPO underpricing and document similarity. Aligning with the information revelation theory proposed by Benveniste and Spindt (1989), they argue that higher levels of information production and pre-market due diligence by underwriters contribute to reduced underpricing. Issuers investing more resources in information production will thus produce a prospectus with more distinct and informative content. The degree of underpricing is linked to the ratio between informative content (passages containing new information) and standard content (passages mirroring peer prospectuses). To decompose a document into informative and standard content, the researchers developed a novel methodology. The process begins with the creation of a BOW representation for each prospectus. Comparable IPOs are then identified, comprised of two groups: Those that took place within the previous 90 days (recent IPOs), and those within the same industry occurring between 90 to 365 days prior to the issue date (same-industry IPOs). Lastly, an OLS regression without intercept is run. The reference document's BOW vector is used as the dependent variable. Two independent variables – the average BOW vector of recent IPOs and the average BOW vector of same-industry IPOs – are used. Finally, standard content is defined as the sum of both regression coefficients, while informative content is equal to the absolute value of the residuals, representing the content not explained by the two predictors (Hanley & Hoberg, 2010, pp. 2837-2839). The results show a significant positive (negative) relationship between standard content (informative content) and underpricing. In a subsequent study, Hanley and Hoberg (2012, pp. 236-239) apply BOW analysis to a set of IPO prospectuses to analyze the relationship between litigation risk, underpricing and strategic disclosure. The researchers hypothesize that firms may choose between disclosing additional information or using higher levels of underpricing to protect against potential lawsuits. For each given IPO, they compare the similarity of BOW vectors between the initial version of the prospectus and later revised versions. A high degree of similarity leads the researchers to anticipate an increased likelihood of omitting value-relevant information, that was disclosed during bookbuilding. This hypothesis is conversely related to the ex-ante litigation risk proxy mentioned in Guo et al. (2022), which links more legal disclosure to higher underpricing. In a sample of 1,623 US IPOs between 1997 and 2005, they find robust support for the idea that firms may choose underpricing over disclosure as a strategy to hedge against litigation risk. They further argue that the deterrence effect of underpricing is mainly to reduce the likelihood of a Section 11 lawsuit, which would include the names of the underwriters, and the resulting reputational damage to the investment banks. This explains why underwriters continue to opt for underpricing

even when positive information is revealed at investor meetings. A methodologically analogous approach is employed in a study by Hoberg and Phillips (2010, p. 1425), who research asset complementarities as a success factor for mergers and acquisition deals. Therefore, they utilize the BOW technique to evaluate the similarity of product descriptions in 10-K reports of acquirer and target. A recent study making use of advancements in NLP was conducted by Breitung and Müller (2022, pp. 2-3). They propose a new methodology to assess document similarity, using context-dependent sentence embeddings to develop a new metric indicating how similar the current annual report of a firm is to the one from the past year. Therefore, they calculate pairwise cosine similarities between sentence embeddings to identify for each sentence the most similar sentence in the compared document. The final similarity score, labeled as simBERT, is the average of the maximum cosine similarities. High simBERT scores indicate that the company did not disclose much new information. The advantage of this approach compared to conventional BOW methods is that they capture the full semantic meaning of sentences, which means that no exact word overlap is required to show that sentences are similar and negated sentences have dissimilar meanings. This approach is also supported by the findings of Meden (2022, p. 4), showing that sentence embeddings demonstrate high accuracy for semantic similarity tasks. The author further argues that fastText embeddings are also a valid tool for the exploration of similarities, although these output embeddings are not context dependent. The evolution of technologies in sentiment and similarity analysis has greatly expanded the potential for researchers to enhance the accuracy and efficiency of their results. In the following chapter, the development of the hypotheses for this study will be described, which will then be followed by a new section including a detailed description of the methodological approach.

2.4. Hypothesis Development

In order to investigate the research question – if textual analysis of IPO prospectuses can contribute to explaining underpricing in European IPOs – two propositions and a set of hypotheses are developed in accordance with the cited literature. These hypotheses address the probability that a certain amount and type of information in a prospectus will affect the underpricing level and evaluate the performance of different NLP methodologies in conducting textual analysis.

Proposition 1: Textual information contained in IPO prospectuses explains the level of underpricing.

Ex-ante uncertainty hypothesis (H1): Uncertain sentiment within a prospectus is indicative of higher levels of ex-ante risk for investors (Beatty & Ritter, 1986; Loughran & McDonald, 2013; Rock, 1986).

H1: Uncertain sentiment is positively related to underpricing.

Legal liability hypothesis (H2): The presence of legal sentiment in a prospectus is indicative of a firm's higher ex-ante risk of litigation (Guo et al., 2022; Hanley & Hoberg, 2012; Lowry & Shu, 2002)

H2: Litigious sentiment is positively related to underpricing.

Neutral language hypothesis (H3): Neutrality (i.e., neither positive nor negative sentiment) in prospectus language should improve the trust in the firm's disclosure. This hypothesis is inspired by the idea in Ferris et al. (2013), that conservative language increases credibility among investors. Expanding on the notion of greater credibility, this study hypothesizes that language with more negative, but also positive sentiment increases skepticism and results in greater underpricing.

H3: Positive and negative sentiments are positively related to underpricing.

Information revelation hypothesis (H4): Firms can invest in price discovery or engage investors in information production. Investors require compensation for information revelation (Benveniste & Spindt, 1989; Hanley & Hoberg, 2010).

H4: Prospectus similarity is positively related to underpricing.

Proposition 2: Performance of textual analysis influenced by selected NLP tool.

Methodological approach hypothesis (H5): The efficiency of semantic content analysis in prospectus documents can be enhanced through state-of-the art NLP techniques.

Neural network hypothesis (H5.1): The use of neural network-based word embeddings in textual analysis will yield superior results compared to traditional word count methods, due to their ability to interpret meaning without relying on exact word overlap (Mikolov et al., 2013, 2017; Seegmiller et al., 2023).

H5.1: FastText variables explain more variance than word count methods.

Transformer model hypothesis (H5.2): Transformer-based word embeddings are expected to yield the highest performance in textual analysis, given their inherent capacity to understand the contextual relationship of words (Breitung & Müller, 2022; Devlin et al., 2018).

H5.2: BERT-based variables explain more variance than fastText variables.

3. Data and Methodology

3.1. Data and Sample

The dataset employed in this study is derived from a subset of the sample originally utilized by Kaserer and Treßel

(2023, pp. 3-4). It includes listings from Denmark, Finland, France, Norway, and Sweden, with issue date between January 2016 and September 2022. Only exchanges with a registered SME growth market are included. This applies to Euronext, NASDAQ Nordic, Nordic Growth Market (NGM) and Spotlight Stock Market (SSM). The original study considers only operating companies and therefore, excludes listings from SPACs, REITs and closed-end funds. Further, it excludes market transfers, relistings, and secondary listings to limit the scope of the study to initial listings. Merger-based transactions (mergers, demergers, reverse takeovers) are neither considered. Furthermore, the listing must include the sale of either primary or secondary shares, which does not apply to direct listings. Since the goal of this study is to research the effect of textual variables on IPO underpricing, only reliable underpricing data should be considered. Thus, listings with underpricing above 200% are excluded to reduce the impact of noise in the data. After applying these filters, a final sample of 745 offerings are included. As shown in Table 2, of the 745 listings 632 are classified as IPOs, while only 113 are classified as private placements. The majority of listings in the sample are from Sweden, followed by Norway, France, Finland and Denmark having the fewest.

The dataset contains the listing documents in a machine-readable format for each transaction. In the original study, these documents were translated into English using Google Translate (Kaserer & Treßel, 2023, pp. 17-18). To ensure that only English vocabulary words are retained, in this study the listing documents are filtered using the vocabularies from the Python packages NLTK corpus and Pyspellchecker. Words not found in at least one of these dictionaries were discarded. The original dataset contains variables for firm and offering characteristics, including many hand-collected variables from prospectuses, which are further augmented in this study with selected variables from Refinitv Eikon.

3.2. Techniques for Textual Analysis

The proposed methodologies use different NLP approaches for textual analysis, with the objective to link underpricing with prospectus sentiment and prospectus similarity. The primary tool to extract semantic meanings from unstructured textual data is the implementation of word or sentence embeddings. The chosen methodologies are adopted according to Seegmiller et al. (2023) and Breitung and Müller (2022). Word embeddings are a novel technology, employing algorithms to map words into vector space. Their technological advantage, compared to traditional methods, is that they allow for words to be similar, without requiring exact overlap. This makes them a more suitable choice for extracting meaning from a document than traditional word count, such as POW and BOW methods. Introduced by Mikolov et al. (2013, pp. 1-5), word2vec was the first widely available embedding model based on neural networks. To train the model, Continuous Bag-of-Words (CBOW) and Continuous Skip-gram algorithms are used. CBOW predicts a target word based on its surrounding context and Skip-gram does the opposite, predicting the context words from a given target word. More recent models of this type include GloVe, which learns efficient word representations by training on word-word co-occurrence statistics (Pennington et al., 2014, p. 1532) and fastText (Mikolov et al., 2017, p. 2), which is trained on n-grams and has the ability it to handle out-of-vocabulary words. The introduction of transformer architecture (Vaswani et al., 2017) has spurred considerable advancements in the NLP domain. In contrast to earlier models, current methods can generate contextualized word embeddings that effectively capture the semantics of words within their unique contexts. A prominent model is BERT, short for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018, pp. 1-3). BERT, structured as a stack of multiple Transformer encoder layers, incorporates several self-attention heads in each layer. These self-attention mechanisms facilitate the computation of context-aware embeddings for any given sequence of input tokens. A shortcoming, is that no independent sentence or document embedding is computed and thus cosine similarity measures are not well suited for document comparisons. Reimers and Gurevych (2019) address this issue and suggest SBERT, which is a modification of the conventional BERT model fine-tuned for such semantic textual similarity tasks.

FastText word embeddings can be synthesized to document vector representation by aggregating the set of word vectors using their term-frequency-inverse-document-frequency (tf-idf) weights. Tf-idf weighting is a statistical measure that quantifies the importance of a word to a specific document relative to the entire corpus. It emphasizes terms that are frequent within a particular document but not common across all documents (Seegmiller et al., 2023, p. 6). This resulting document embedding can be used for tasks such as sentiment and similarity analysis. However, when it comes to SBERT embeddings, as per the approach by Reimers and Gurevych (2019) and Breitung and Müller (2022), such weighting is not required. Unlike fastText, where a single document vector is obtained, SBERT characterizes a document through a list of sentence embeddings. Cosine similarity, which is derived by comparing the cosine of the angle between two vectors, is used in both cases to obtain sentiment and similarity scores. This technique interprets the similarity in terms of semantic meaning of two embedding vectors. Specifically, the measure is computed as the normalized dot product of the two vectors, which provides a robust measure of their relative orientation. Cosine similarity scores range between [-1,1]. A value of -1 signifies oppositely directed vectors, a value of 0 denotes orthogonal vectors, and a value of 1 represents identically orientated vectors. These values analogously express the semantic associations between words or documents. Equation 1 describes the formula for cosine similarity of two vectors A and B (Seegmiller et al., 2023, p. 2):

$$similarity = \cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} \tag{1}$$

Table 2: Initial offerings by exchange country and type

Country	Exchange Name	Exchange Type	IPOs	Private placements	Total
Denmark	NASDAQ	Regulated market	9	0	9
	Nordic	MTF	13	0	13
		SME growth market	31	2	33
	Total		**53**	**2**	**55**
Finland	NASDAQ	Regulated market	17	0	17
	Nordic	MTF	19	1	20
		SME growth market	27	1	28
	Total		**63**	**2**	**65**
France	Euronext	Regulated market	38	0	38
		MTF	6	7	13
		SME growth market	39	2	41
	Total		**83**	**9**	**92**
Norway	Euronext	Regulated market	19	3	22
		MTF	1	86	87
	Total		**20**	**89**	**109**
Sweden	NASDAQ	Regulated market	65	0	65
	Nordic	MTF	108	5	113
		SME growth market	124	3	127
	NGM	MTF	30	1	31
		SME growth market	11	0	11
	SSM	MTF	46	2	48
		SME growth market	29	0	29
	Total		**413**	**11**	**424**
Total			**632**	**113**	**745**

This table shows the number of transactions split by country, exchange name, exchange type and listing type for the period between 2016 to 2022.

Cosine similarity is employed to generate similarity and sentiment variables. In the following chapter, a detailed explanation of the derivation process for these textual variables is provided.

3.3. Sentiment Analysis of IPO Prospectuses

3.3.1. Percentage-of-Words Sentiment Analysis

The POW approach is based on the study by Loughran and McDonald (2013). It is a straightforward method, of determining document sentiment through word list frequencies within a document. Critical for the success of this approach is the use of a word list that is specific to the applicable domain. Particularly in the field of accounting and finance, words often have different meanings than in general language. Thus, using a "discipline-specific word list can reduce measurement error" (Loughran & McDonald, 2011, p. 44). Similar to the study by Loughran and McDonald (2013), the LM word lists are used to determine the POW variables. A Python module to download the word lists can be found at the Notre Dame Software Repository for Accounting and Finance, which is maintained by Bill McDonald, who co-authored the referenced study[3]. Following the developed set of hypotheses, this study makes use of the litigious, negative, positive, and uncertainty word lists. Loughran and McDonald (2011, pp. 44-45) provide a detailed description of the word lists, which is summarized in the following paragraph.

The litigious sentiment list consists of 731 words that are not necessarily directly related to lawsuits but are common in a litigious environment and often are related to legislation and regulation. These words include terms like 'allegation', 'claimant', 'deposition' and 'hereupon'. The negative sentiment list, which consists of 2,337 words, describes undesirable financial situations for a firm. Examples of these words include 'bankruptcy, 'decline', 'difficult' and 'loss'. These words are often associated with negative implications and are reflective of adverse conditions or outcomes. The positive sentiment list is significantly more compact, containing 353 words that are usually linked with favorable circumstances in finance. This list encompasses terms like 'achieve', 'efficient', 'improve' and 'profitable', all of which convey a sense of success, strength, or beneficial attributes. The authors highlight

[3] Notre Dame Software Repository for Accounting and Finance: https://sraf.nd.edu/loughranmcdonald-master-dictionary/

that underwriters who write the prospectus are conscious of investors' use of textual analysis to evaluate the document. As a result, they try to avoid negative words and instead use negated positive words. Consequently, the researchers emphasized selecting terms with unilateral meanings that, when used, clearly express the intended sentiment. Lastly, the uncertainty list focuses on the sentiment of ambiguity, doubt, and imprecision, containing a total of 285 words. Words such as 'approximate', 'contingency', 'uncertain' and 'sometimes' are included. Some words may appear in multiple lists. The overlaps exist mainly between litigious, negative and uncertainty lists (Loughran & McDonald, 2011, p. 45). Each list contains several topics that are related to the specific sentiment and describe it holistically. In the process of obtaining the POW variables, no pre-processing operations such as stop-word removal or lemmatization are applied to word lists and documents. As a result, the lists often include multiple words from the same word group. For example, the negative word list contains words such as 'defend', 'defendant', 'defendants', 'defended', 'defending', and 'defends'. Additionally, words that may lack standalone semantic meaning but intend to create specific sentiments in certain contexts are also included in the list. Examples from the litigious list are 'herefor', 'herefrom', 'insofar', 'moreover', 'therefrom'. This shows that the selection of words is very approach-specific. In the next chapter, which will discuss the creation of a set of embedding-based variables, a new method addressing these characteristics will be suggested.

3.3.2. FastText Sentiment Analysis

Similar to the POW approach, fastText variables are calculated for the same categories, using LM word lists as the foundation for sentiment scoring. The extraction of sentiment through word embeddings is a multi-stage process inspired by the study of Seegmiller et al. (2023). The process consists of data preprocessing, document vectorization, and finally, document sentiment scoring for each category. The first step involves selecting the most suitable embedding model, and in this study, fastText has been chosen. Developed by Mikolov et al. (2017), fastText is the latest model in conventional word embeddings. It utilizes n-grams to handle out-of-vocabulary words and thereby to optimize the amount of information captured. The default version of fastText represents embeddings in a 300-dimensional vector space. Pretrained word vectors are available for download in multiple languages, having been trained on Wikipedia and other web data sources.[4] Before computing the word embeddings, the documents must be preprocessed, to facilitate accurate sentiment analysis. The pre-processing steps are adopted from the BOW approach described in Kaserer and Treßel (2023, p. 19). The prospectuses are subjected to several pre-processing steps. First, they are tokenized, and stop words, named entities, and punctuation marks are removed. Subsequently, the resulting tokens are filtered based on their parts of speech tag, with pronouns, proper nouns, conjunctions, determiners, adpositions, interjections, symbols, and other text parts being discarded. The remaining tokens undergo lemmatization, simplifying them to their base or root form. This normalization allows for simplified comparisons within word groups by establishing a standard form for the words. In addition, non-ASCII characters are removed and tokens with less than three and more than 45 characters are removed to minimize noise. These pre-processing steps are executed using the Python packages Spacy, NLTK and Unidecode.

The next phase involves computing a tf-idf weighted embedding vector for each document. This includes first applying the tf-idf vectorization to each document, a method that is similar to the BOW approach. It creates a vector, representing all unique words in the corpus of documents. Using the tf-idf vectorizer from the Python package Scikit-learn, for the given dataset, a vector of length 25,684 is obtained. Each of the 745 documents is then characterized by a vector of the same length, storing the individual tf-idf weights for each word. The formula for calculating the tf-idf weight for word t in document d is outlined in equations 2 to 4. Equation 2 describes the synthesized tf-idf formula. Equation 3 describes the calculation for the first component of the full formula, the term frequency (tf), and equation 4 describes the second component, the inverse document frequency (idf). The representation is based on Seegmiller et al. (2023, pp. 6-7) and Mandal et al. (2021, p. 435):

$$tf - idf(t,d) = tf(t,d) \cdot idf(t) \quad (2)$$

$$tf(t,d) = \frac{\text{frequency of term } t \text{ in document } d}{\text{total number of terms in document } d} \quad (3)$$

$$idf(t) = \log\left(\frac{\text{total number of documents}}{\text{total number of documents with term } t + 1}\right) \quad (4)$$

The computations yield a 745 × 25,684 matrix representing the tf-idf weights of the documents. Subsequently, the vector containing all words is passed into the fastText model, generating a 25,684 × 300 matrix that includes the embedding representations of each word. To obtain a single vector for each document, the dot product between a document's tf-idf weights and the word embedding matrix is computed, effectively combining the importance of individual words with their semantic representations. Especially for prospectuses that contain large amounts of standardized words, which are present across all documents, this weighting technique optimizes the information content of the document embedding. The resulting outputs are 300-dimensional document arrays characterizing the prospectuses. Compared to the traditional BOW methodology, this approach provides a significantly reduced dimensionality, which makes computation more efficient (Seegmiller et al., 2023, pp. 7-8). To further refine the analysis, several modifications are applied to the LM word lists. These adjustments address limitations arising from the narrow application scope (limited to POW). Unaltered LM

[4] fastText library: https://fasttext.cc/docs/en/crawl-vectors.html

word lists may introduce measurement errors due to varying word group lengths and the inclusion of words without standalone meanings. A subset of the LM word lists is created by applying similar part-of-speech filters as for the prospectus pre-processing. Additional filters are added for auxiliary words and existential adverbs (e.g., there, therefrom, etc.) because these word types often do not express a real sentiment. Subsequently, the remaining words are lemmatized, and resulting duplicates are discarded. FastText embeddings are then calculated for each word in the reduced word lists. For optimized mapping, these vectors are further reduced to two dimensions using principal component analysis. Subsequently, for each word list, the two-dimensional word vectors are clustered into five groups using the k-means algorithm. These operations are executed utilizing the Scikit-learn package. From each cluster, two words are randomly drawn to form a robust new word list, yielding a total of 10 different words. This results in four lists of identical length with a robust exposure to distinct topics within each sentiment list. The reduced word lists are included in Table 11, in the appendix. Finally, sentiment scores for each document are calculated using the different word lists. This is achieved by computing the cosine similarities between the document embedding vector and each of word list item's embedding vectors. The document's score for a given category is represented as the average of these cosine similarities. This measure quantifies the relationship between the sentiment in the prospectus and the sentiment embodied by the wordlist. Consequently, each prospectus is associated with multiple sentiment scores. Each score is corresponding to how close the semantic meanings between prospectus and respective wordlists are. A shortcoming of this approach is that it neglects the contexts of words, resulting in a word with different meanings for a given context being wrongfully represented by the same embedding vector. Additionally, the order of words and the semantic meaning of negated expressions is not captured effectively. To address these limitations, the next chapter presents a method based on state-of-the-art transformer models, which enables the computation of contextualized sentence embeddings.

3.3.3. Sentence-BERT Sentiment Analysis

The introduction of the transformer architecture, which is the basis for BERT, has revolutionized the field of NLP and many connected domains, as it has established a new benchmark for numerous language-specific tasks with state-of-the-art results. (Devlin et al., 2018, p. 2). The inspiration for the SBERT sentiment variables comes from Seegmiller et al. (2023, p. 9), who propose to extend the methodology described in the previous chapter with contextualized word embeddings. For this purpose, numerous models, pre-trained on extensive corpora and fine-tuned for various downstream tasks, can be readily accessed through the Hugging Face Transformers[5] and Sentence Transformers[6] libraries. The broad availability of these resources has significantly catalyzed the adoption of transformer-based models amongst the research community (Wolf et al., 2020). Previous studies have demonstrated the superior performance of contextualized embeddings in terms of word similarity tasks compared to traditional models (Rogers et al., 2021, p. 845).

In this study, the SBERT model by Reimers and Gurevych (2019) is used. As shown in Figure 1, the basis for the sentence embeddings is the classical BERT model introduced by Devlin et al. (2018, p. 3). The BERT base model is equipped with 12 layers, and 12 attention heads and contains 110M parameters. The model has a hidden size of 768, which represents the dimensions of the embedding vectors. SBERT is an extension of BERT, trained on classified sentence pairs, it produces a fixed-sized average vector representation for a given input sentence. The fine-tuning process optimizes the model to combine individual word vectors in a way that the resulting sentence embedding is semantically meaningful. For each sentence an embedding with a hidden size of 768 is produced (Reimers & Gurevych, 2019, pp. 3-4). SBERT requires fewer preprocessing steps than the fastText method. The only requirement is to split the documents into lists of sentences. For this purpose, the Sentencizer method from the Spacy package is used, which employs grammar-based sentence-boundary detection. Next, the list of sentences from each of the 745 documents is passed into the SBERT model. The resulting output is a list of the same length, with each entry being a 768-dimensional sentence embedding vector. Although SBERT can handle single-word inputs, as found in the sentiment word lists, the model is fine-tuned for sentence inputs. Therefore, it is suggested to modify the word lists to optimize the sentiment scores for this approach. The previously introduced reduced word lists are expanded by incorporating each word into a short, exemplary sentence that conveys its potential meaning within a prospectus document, consistent with the given sentiment category. The framed sentences are intentionally unspecific so as not to limit the semantic meaning of the original word list. The sentence dictionaries are listed in Table 11 in the Appendix. Except for the fact that the sentiment lists now contain sentences instead of words, the remaining process to compute sentiment scores is analogous to the process presented in the previous chapter. The sentences from the sentiment list are encoded using the SBERT model. Next, by calculating the average pairwise cosine similarity between the document's and the dictionary's sentence embeddings, the final sentiment scores for each category are computed. To demonstrate the validity of the newly designed sentiment scoring method, two exemplary documents are provided:

[5] Hugging Face Transformers library: https://huggingface.co/docs/transformers/index
[6] Hugging Face Sentence Transformers library: https://huggingface.co/sentence-transformers

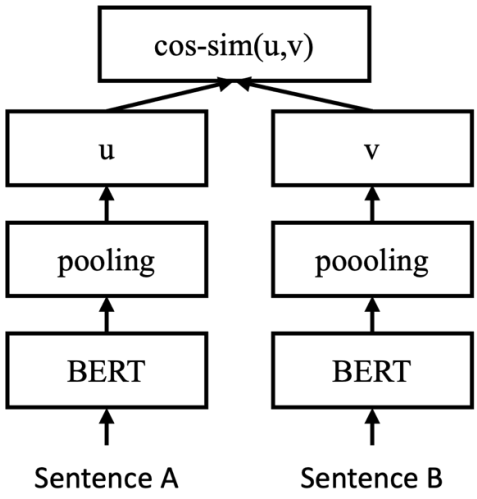

Figure 1: SBERT architecture to compute similarity scores. Taken from Reimers and Gurevych (2019, p. 3)

> **Document 1:** ["We are not able to limit losses related to tax penalties."; "Economic conditions might be worse than expected."; "Competitive disadvantages could be result of criminal proceedings."]
>
> **Document 2:** ["We managed to exceed quoted profit guidance thanks to strong demand."; "The annual targets for the company will be exceeded." "We have successfully averted any legal problems."]

The sentiment scores logically align with the anticipated direction: Document 1 contains negative, litigious, and uncertain language, while Document 2 expresses positive, certain, and non-litigious language. The scores for these categories are notably distinct. Because prospectuses are long documents with thousands of sentences, the differences between sentiment scores for full documents will be more nuanced, and the scores will generally be lower due to the different sentence structure (i.e., prospectuses contain more complex sentences than sample documents). In addition to sentiment analysis, this study explores the degree of information revelation in a prospectus document. To achieve this, a similarity analysis of comparable documents is conducted. The methodology for this will be detailed in the subsequent chapter.

3.4. Similarity Analysis of IPO Prospectuses

3.4.1. Bag-of-Words Similarity Analysis

In order to study the impact of information disclosure on underpricing, the BOW methodology is adopted according to the studies by Hanley and Hoberg (2010) and Hanley and Hoberg (2012). The objective is to establish a measure for the information revelation present in an IPO prospectus. Information revelation refers to the level of information generation during pre-market due diligence, carried out by the issuing company and its underwriters. Since the process of generating information involves substantial costs, underwriters might choose to outsource these activities to investors. If a larger portion of the price discovery process is outsourced, it may result in the prospectus containing less information specific to the given offering. Consequently, the prospectus might resemble those from prior related offerings. Therefore, prospectus similarity should describe the degree of information revelation with an inverse relationship. Based on the BOW approaches referenced in the above-mentioned studies, a new measure for similarity of a given prospectus compared to those of its peer group is introduced to test this assumption.

The process consists of document pre-processing, vectorization, and similarity analysis. For preprocessing, the identical steps, as described in Chapter 3.3.2 are utilized. Subsequently, to convert the prospectuses into a numerical format, the CountVectorizer function from the Scikit-learn package is employed. To refine the content and eliminate unwanted elements not removed during preprocessing, only words that appear in a minimum of 2.5% of the documents are included. This step is necessary with regard to similarity analysis since higher dimensional vectors might lead to biased results (Breitung & Müller, 2022, p. 8). The output of the vectorization is a BOW vector containing 6505 elements, which is notably smaller than the tf-idf vector associated with the fastText approach. For each of the 745 documents the vector stores the word count of the respective item. Normalization is applied, so that vectors store relative word frequencies which sum up to 1 for every document, independent of its length (Hanley & Hoberg, 2010; Kaserer & Treßel, 2023). In the next step, the reference documents are identified from recent IPOs and same-industry IPOs, as outlined by Hanley and Hoberg (2010). Recent IPOs occurred in the previous 90-day period and same-industry IPOs are identified by their Fama-French 12 industry code (FF12) and are limited to offerings that occurred between 90 days to 1 year before the respective issue

Table 3: Document sentiment scores with SBERT

	Litigious-SBERT	Negative-SBERT	Positive-SBERT	Uncertainty-SBERT
Document 1	0.408	0.520	0.325	0.503
Document 2	0.318	0.318	0.507	0.292

This table presents the SBERT scores for Document 1 and Document 2 for sentiment categories litigious, negative, positive, and uncertainty.

date. The similarity of a given prospectus to the comparable documents is then determined using cosine similarity, as recommended by Hanley and Hoberg (2012, p. 253). For each of the two reference document groups, the average cosine similarity is computed between the BOW vector of the given prospectus and the BOW vector of the documents in the respective group. Subsequently, the similarity score is expressed as the average of the cosine similarities of both groups (in case one of the groups has no elements, the similarity score defaults to the value of the other group). The approach detailed in Hanley and Hoberg (2010), which determines standard and informative content using regression coefficients and residuals respectively, was also explored. The findings mirrored those in the original study. Yet, due to concerns over interpretability and consistency, the cosine similarity method was favored. For clarity, this specific approach is labelled as BOW (cosine similarity). The next chapter will present a novel similarity measure utilizing contextualized sentence embeddings.

3.4.2. SimBERT Similarity Analysis

The simBERT document similarity measure addresses a significant limitation of the BOW method. Namely, BOW approaches cannot capture the semantic meaning of texts, which largely depends on structure and use of words. Moreover, different words often express identical concepts, for which simple word count techniques fail to capture the similarities. Depending on the author's writing style, two documents can be characterized by very different vector representations, even when the message conveyed is similar. SimBERT, a novel method for extracting semantic similarity, that addresses these concerns, is proposed by Breitung and Müller (2022). The proposed document similarity measure makes use of SBERT sentence embeddings, which are obtained identically as in Chapter 3.3.3, for the sentiment scoring method. SimBERT assesses the similarity between two documents on the sentence level. For every sentence in a particular document, its closest counterpart in the second document is determined by computing pairwise cosine similarities between the sentence embeddings from both documents. The similarity score between documents is derived by averaging the maximum cosine similarities across all sentence pairs (Breitung & Müller, 2022, p. 2). Finally, to determine the degree of information revelation, each prospectus is assessed in comparison to documents from similar offerings. A high degree of similarity indicates a low level of information revelation, which is anticipated to relate to more pronounced underpricing. To identify the relevant offerings for comparison, the methodology outlined in the preceding chapter is used. For each category – recent IPOs and same-industry IPOs – the average similarity scores are calculated between the prospectus and every document within the respective group. The final simBERT score is derived by averaging the similarity scores obtained from each of the two document categories. Again, if one group contains no documents, the value from the other group is adopted as the final simBERT score. To illustrate this methodology, the example from the SBERT chapter is augmented with a third document, with a similar meaning as Document 1, but different use of vocabulary (Document 1 and Document 2 remain identical as in the previous example with opposing meanings):

> **Document 3:** ["We cannot mitigate losses from tax-related fines.", "The economic climate may deteriorate beyond our predictions.", "Legal issues could lead to competitive setbacks."]

The outcomes validate the efficacy of the simBERT approach. They distinctly differentiate between semantically divergent documents (Document 1 - Document 2 and Document 2 - Document 3) while demonstrating a high similarity score for related documents (Document 1 - Document 3). A traditional BOW method would have failed to identify the similarity, since word overlaps were intentionally avoided. Consequently, simBERT can be used as a robust measure of textual similarity. This concludes the present chapter and the section detailing methodology. The subsequent chapter will provide summary statistics encompassing the most critical document, firm- and offering-specific, and textual variables.

4. Descriptive Statistics

4.1. Listing Documents

This study explores the relationship between textual information and underpricing, therefore using IPO prospectuses as the main source of information for the study. Given the different content requirements of the prospectus types outlined in previous chapters, it is important to illustrate the effect of this regulatory aspect on the actual listing document. For this purpose, Table 11, in the Appendix, provides detailed information on the word count, sentence count, average sentence length, and number of unique words for each of the document types. It shows that the average length of documents varies substantially among the groups. The full prospectus stands out with the highest average length, containing 58,331.384 words distributed over 2,216.858

Table 4: Document similarity scores with simBERT

	Document 1	**Document 2**	**Document 3**
Document 1	1	0.388	0.837
Document 2	0.388	1	0.468
Document 3	0.837	0.468	1

This table presents pairwise simBERT scores between Document 1, Document 2 and Document 3.

sentences. Following that, the growth prospectus averages 33,506.486 words, divided into 1,252.171 sentences. Admission documents have the shortest content, with an average of 19,950.344 words and 827.466 sentences. These disparities in content volume align with the stricter regulatory requirements imposed on growth prospectuses and full prospectuses. Figure 2 visually represents this relationship, highlighting the sentence count associated with the different listing documents.

In terms of sentence length, full prospectuses and growth prospectuses are quite similar, averaging 25.669 and 25.697 words per sentence, respectively. Admission documents present a noticeably shorter average sentence length, with 24.203 words. Sentence length is a commonly utilized metric to gauge complexity and readability (Loughran & McDonald, 2016). Longer sentences often denote lower readability. This suggests that the stringent content prerequisites for full and growth prospectuses while indicating more information disclosure, might simultaneously heighten complexity for investors. The number of unique words is distributed analogously to the document length. The influence of different listing types on underpricing is further examined in the following chapter.

4.2. Transaction Volume and Underpricing

The underpricing phenomenon proves to be consistent within the analyzed sample. Figure 3 presents the yearly average underpricing across the different exchange types during the sample period. The values are mostly positive throughout the years. Listings on regulated markets consistently display an underpricing of around 10%. Listings on MTFs, in contrast, displayed greater variation. While they showed average underpricing of over 20% in 2019 and 2020, in 2018 the average underpricing was negative. The newly introduced SME growth markets, which recorded the first transactions in 2019, demonstrated positive underpricing for the years 2019 through 2021, but shift to negative underpricing in 2022.

Figures 6 and 7 in the Appendix provide a more comprehensive view of transaction volume and mean underpricing across different category breakdowns. As shown in subfigure 6(a), the periods with the highest market activity, are spanning from Q3 2020 to Q4 2021. Subfigure 7(a) reveals notably elevated levels of underpricing during Q2 and Q3 2020, coinciding with the onset of the Covid-19 pandemic in Europe. According to the theory of Beatty and Ritter (1986), an explanation for this observation might be that the pandemic has raised underlying risk and risk perception, which caused the higher levels of underpricing.

The data in subfigure 7(b) underscores the disparities between listing types. Private placements, with a sample size of 113, show an average underpricing of 14,336%, nearly twice the average underpricing of the 632 IPOs, which stands at 7,380%. Further breakdowns by exchange and prospectus type can be found in subfigures 6(c), 6(d), 7(c), and 7(d). Surprisingly, the listings on a regulated market display the highest underpricing average at 10,265%, followed by MTFs at 8,431% and SME GMs at 7,431%. Contrary to expectations that listings without a prospectus requirement would experience higher underpricing, the data shows listings with a full prospectus average of 9.118% underpricing almost identical to admission documents, which have a slightly higher value of 9.749%. Only growth prospectuses deviate notably with a mean underpricing of 2.766%. The levels of underpricing vary significantly throughout the sample period. The irregular behavior observed in certain categories may be partly due to regulatory changes and external shocks, such as the Covid-19 pandemic, which affected European markets during the sample period. Given the limited timeframe of the sample, potential biases in the data cannot be ruled out. To obtain reliable results from the study, it is crucial to select a robust set of control variables for subsequent regressions. The following chapter will provide summary statistics for these variables.

4.3. Control Variables

In this chapter, the set of variables, later used for regression analysis is introduced. The selection of these variables is based on related studies on IPO underpricing and textual analysis by Guo et al. (2022), Hanley and Hoberg (2010), and Loughran and McDonald (2013). The subsequent analysis is based on Table 12 in the Appendix, which contains the descriptive statistics.

The primary variable of interest in this study is underpricing. This is defined as the change between the offer price and the first day closing price, indicated in decimal format. As highlighted in the preceding chapter, differences exist between IPOs and private placements, with the latter displaying substantially greater underpricing. The average underpricing across the entire sample is 0.085, which, when compared with the European averages shown in Table 1, appears relatively low. The values reported by Loughran and McDonald

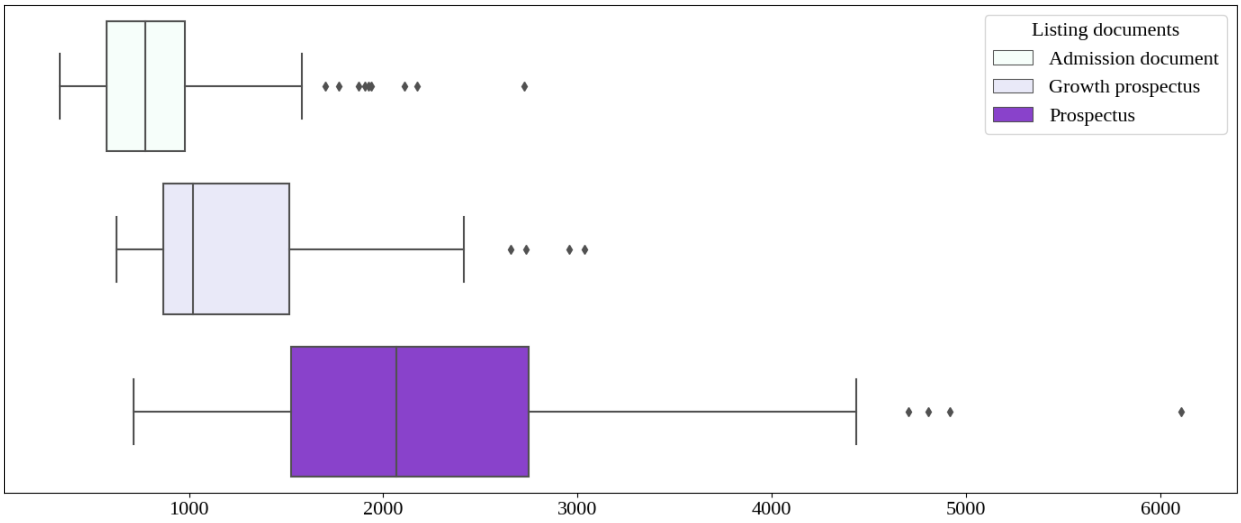

Figure 2: Boxplot of sentence counts in listing documents.

The figure presents boxplots of document lengths proxied by number of sentences for different listing types: full prospectus, growth prospectus, and admission document.

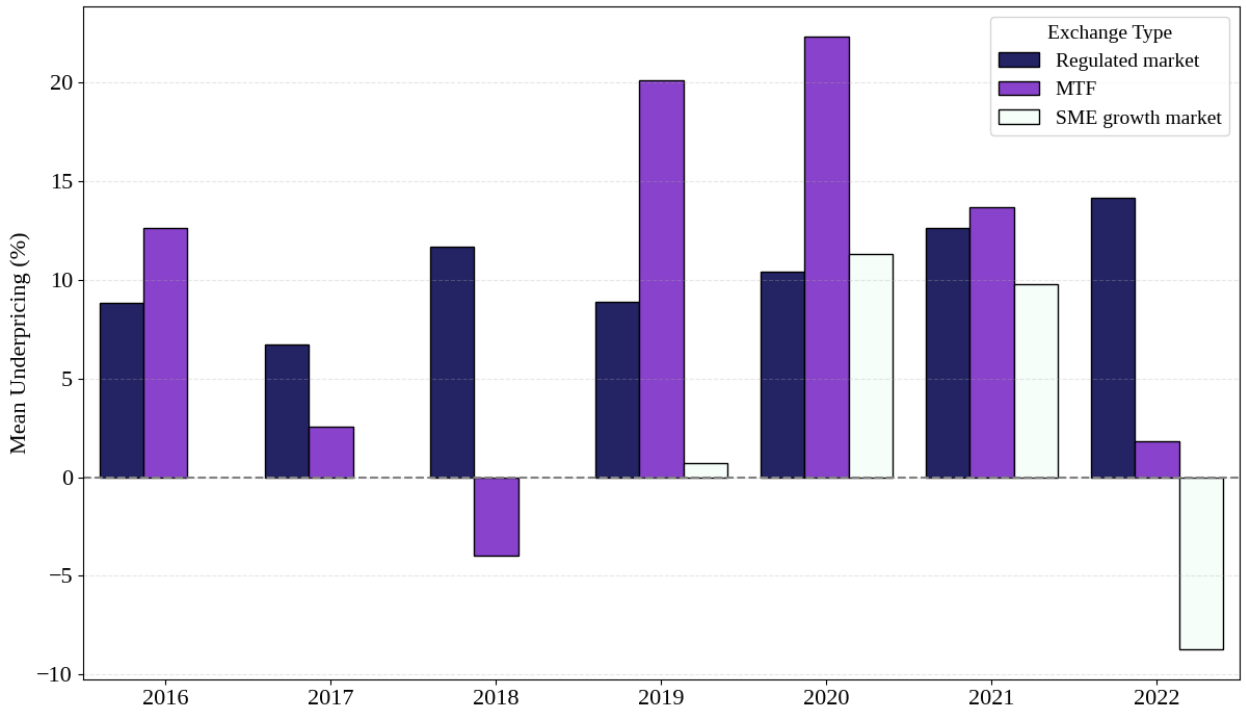

Figure 3: Time-series bar charts of yearly average underpricing

For the period from 2016 to 2022 yearly values for mean underpricing are represented across the different exchange types: Regulated Market, MTF, and SME GM. Each bar represents the mean underpricing for a specific year and exchange type.

(2013, p. 6) and Guo et al. (2022, p. 5) also show much higher average values for underpricing of 0.348 and 0.404, respectively. Table 12 further includes summary statistics for offering and company characteristics used as control variables in the regression models. The sales variable is represented by the issuing firm's sales figure for the year preceding the listing. While sales data is an important indicator of a firm's size and economic performance, used as a control variable in Loughran and McDonald (2013), it was not included in the regression models. This exclusion was due to the multicollinearity issues it posed in combination with textual variables. Instead of sales, the study employs the num-

ber of employees as an indicator of firm size (similarly embodied by the value from the year preceding the IPO). As depicted in Table 12, a parallel trend is observable between sales and the number of employees. Typically, firms opting for IPOs show higher sales and a larger workforce, than those choosing private placements to go public. Yet, it is notable that both metrics are significantly skewed by outliers. This skewness becomes apparent when comparing the mean and the median: while firms, on average, have more than 319 employees, the median is just 24. The stark contrast is evident when examining the maximum value for employees, which stands at 40,131. Thus, to limit the impact of outliers, log transformation is used for this variable for regression analyses. The next control variable to be used is the pre-file NASDAQ return, which serves as a market sentiment proxy, similarly as adopted by Hanley and Hoberg (2010). The variable is determined by the 30-day return preceding the issue date for the specific listing. The average pre-file NASDAQ return stood at 0.016 and is notably higher for private placements than for IPOs. This difference can be attributed to the relaxed regulatory requirements on private placements, allowing underwriters to move more swiftly and react to periods of heightened market sentiment. Compared to the sample used in Hanley and Hoberg (2010, p. 2833) from 1996 to 2005, which shows prior NASDAQ returns of 0.049, it seems that either the market sentiment was generally less favorable over the sample period of this study or the variable became a less important factor for the timing of listings. Market sentiment is also included as a control variable in the study by Guo et al. (2022), however there it is defined as the return between offering date and listing date, which is usually a much shorter period. The remaining control variables are dummy variables. The tech dummy variable, similar to the one found in Hanley and Hoberg (2010), is based on the identified SIC code classification by Loughran and Ritter (2004, p. 35). Nearly one-third of the entire dataset is classified as a technology firm, with a more significant proportion being observed among IPOs. In Hanley and Hoberg (2010, p. 2833) almost 50% of issuances come from tech firms. The subsequent two variables, the regulated market, and the IPO dummy, serve to control for the specific characteristics of European markets and to account for differences between listing types. The regulated market variable shows, that only 2.7% of private placements are listed on regulated markets, a notably smaller fraction compared to the 23.4% of IPOs. This can be explained by considering a main motivator for companies to opt for private placements, which is the prospectus exemption rule. However, prospectus exemption applies only to exchange-regulated markets. As a result, companies choosing private placements rarely decide to list on regulated markets (Kaserer & Treßel, 2023, p. 6).

Table 5 presents the pairwise correlation of the mentioned variables. The pronounced and statistically significant correlation between the number of employees and sales supports the decision to substitute the sales variable in the regression analysis. The only variables that demonstrate a significant correlation with underpricing are the pre-file NASDAQ return and the IPO dummy. Overall, the observed correlations are relatively low, mitigating the risk of multicollinearity in the subsequent regression analyses. In the next chapter summary statistics for the textual variables, which constitute the second component of the regression analyses will be presented.

4.4. Textual Variables

4.4.1. Sentiment Analysis

This chapter explores the sentiment variables highlighted in Section 3.3. Table 13, in the Appendix, offers a comprehensive set of summary statistics for these variables, with a dedicated breakdown by listing type and will be used as a reference throughout this section. First, POW variables are analyzed, which are also considered in the studies by Guo et al. (2022) and Loughran and McDonald (2013). For the POW variables, the mean values marginally exceed the median values across all word lists, though the discrepancies are not substantial. The recorded mean values for litigious, negative, positive, and uncertainty sentiments are 0.648%, 1.147%, 0.816%, and 1.325%, respectively. As such, listing documents display a stronger negative and uncertain sentiment compared to litigious and positive sentiments. These findings align closely with those of Loughran and McDonald (2013, p. 6), who documented values of 0.72%, 1.41%, 0.94%, and 1.28% for the same categories in the final prospectus. However, the values obtained for the Chinese IPO market by Guo et al. (2022, p. 5) differ markedly, indicating 7.06%, 3.60%, 5.22%, and 1.51% for litigious, negative, positive, and uncertainty sentiments respectively. One potential explanation for this variance is that Guo et al. (2022, p. 3) expanded the translated LM wordlists by adding 207 positive words, 53 negative words, 28 uncertainty words, and 51 litigious words, which might appear with high frequency in their sample of prospectuses. Another consideration could be fundamental differences between Chinese listing documents and those from Europe or the U.S. For the sample documents analyzed in this study the standard deviation ranges between 0.202 for positive-POW and 0.341 for uncertainty-POW variable. Comparing IPOs to private placements, the data suggests that private placement documents have a higher frequency of litigious, positive, and uncertain words. In contrast, IPO prospectuses contain more positive terms. For example, the occurrence of litigious words in private placements is nearly double that in IPOs. The differences between litigious and negative sentiments are also substantial, whereas the values for positive words show only a marginal variation. The heightened litigious, negative, and uncertain sentiments observed in IPOs align logically with the company characteristics discussed in the preceding chapter. Typically, private placements are favored by smaller firms with lower sales figures and a smaller workforce. As such, smaller firms often have an inherently higher risk profile. This relationship offers one plausible connection between textual information and the anticipated ex-ante risk.

The subsequent variables were derived using the fastText word embeddings. Higher values signify a more pronounced

Table 5: Pairwise correlations of the control variables

	Underpricing	Sales	Employees	Pre-file NASDAQ return	Tech company (D)	Regulated market (D)	IPO (D)
Underpricing	1						
Sales	0.01	1					
Employees	0.017	0.878***	1				
Pre-fileNASDAQ return	0.094**	0.023	0.024	1			
Tech company (D)	-0.01	-0.059	-0.075**	0.002	1		
Regulated market (D)	0.034	0.189***	0.270***	-0.004	-0.145***	1	
IPO (D)	-0.077**	0.021	0.032	-0.042	0.096***	0.188***	1

This table presents pairwise correlations of control variables. In the table, (D) is used to indicate that the respective variable is a dummy variable. Statistical significance at the 1%, 5%, and 10% levels is denoted by ***, **, and *, respectively.

similarity between the document and the corresponding sentiment word list. Surprisingly, when compared with the POW variables, the highest mean value emerged for the positive sentiment at 0.333, paired with a standard deviation of 0.014. It is followed by the uncertainty variable, with a mean value of 0.300 and a standard deviation of 0.013. The next highest score is attributed to negative sentiment with a mean of 0.264 and the least standard deviation of 0.011. The litigious sentiment records the lowest score with a mean value of 0.245 and a standard deviation of 0.014. Similar to the POW set, the differences between mean and median are neglectable. When looking at the standard deviations for the fastText variables, it becomes evident that the standard deviations are considerably smaller than those of the POW variables. One plausible explanation for this might lie in the computation methods of the variables. Whereas the POW methodology considers solely the words in the LM lists, the embedding method considers all words. Since the maximum similarity score between any two words is set at 1, differences between entire documents containing a vast array of unique words will be denoted by more subtle variations. The SBERT scores are similar to those obtained from fastText in terms of interpretation. Both groups are based on cosine similarities, where a maximum value of 1 signifies two identical sentences. The findings present another unique relative ranking of sentiment detected in the prospectuses when compared to POW and fastText methodologies. Among the SBERT variables, the litigious sentiment shows the highest mean value (0.308) followed by negative (0.299), positive (0.287) and uncertainty (0.282). The respective standard deviations for the variables are 0.012; 0.016; 0.018 and 0.015. Although standard deviations are still small it appears that SBERT variables express more variation within documents. Noteworthy, uncertainty which was most pronounced for the previous variable categories records in this case the lowest mean. The evident differences between variable groups might be attributed to the better processing power of newer NLP techniques. The POW metrics are limited to the terms included in the LM word lists, thus offering a narrower perspective. In contrast, fastText has the capability to discover similarities between related words, which are not considered in the predefined word lists. SBERT marks a further enhancement by including context in its computations, a feature especially useful for negated expressions. While there are pronounced differences in sentiment scores across the various methods, the relationship between IPOs and private placements remains consistent for each sentiment. Across all three methodologies, private placements consistently exhibit higher values for litigious, negative, and uncertain sentiments. Conversely, IPOs consistently display a greater prevalence of positive language.

Furthermore, Figure 4 reveals strict positive correlations when examining pairwise relationships between variables representing each sentiment. The elevated correlation of fastText and BERT with POW acts as a validation of the chosen methodological approach, suggesting that all three techniques capture the same underlying construct. Given that not only NLP techniques but also utilized dictionaries varied for the different approaches, it is anticipated that correlations will still leave some room for variations. This explains why most correlation values are below 0. 500. The distributions and relationships between underpricing and the collection of sentiment variables for each of the described methods are depicted in Figures 8 through 10 in the appendix. An important observation is the high, positive correlation among negative, litigious, and uncertain sentiments, which is strongest for SBERT variables. For the POW methodology, a distinctive clustering can be observed between IPOs and private placements. However, this clustering becomes less apparent in the subsequent methods.

In summary, the descriptive statistics presented offer consistent values across the various sets of variables, laying the groundwork for further analysis in the regression models. The following chapter will provide a parallel analysis of the similarity scores.

4.4.2. Similarity Analysis

In addition to sentiment analysis, this study explores the information content of listing documents. High information content is characterized by a low degree of document simi-

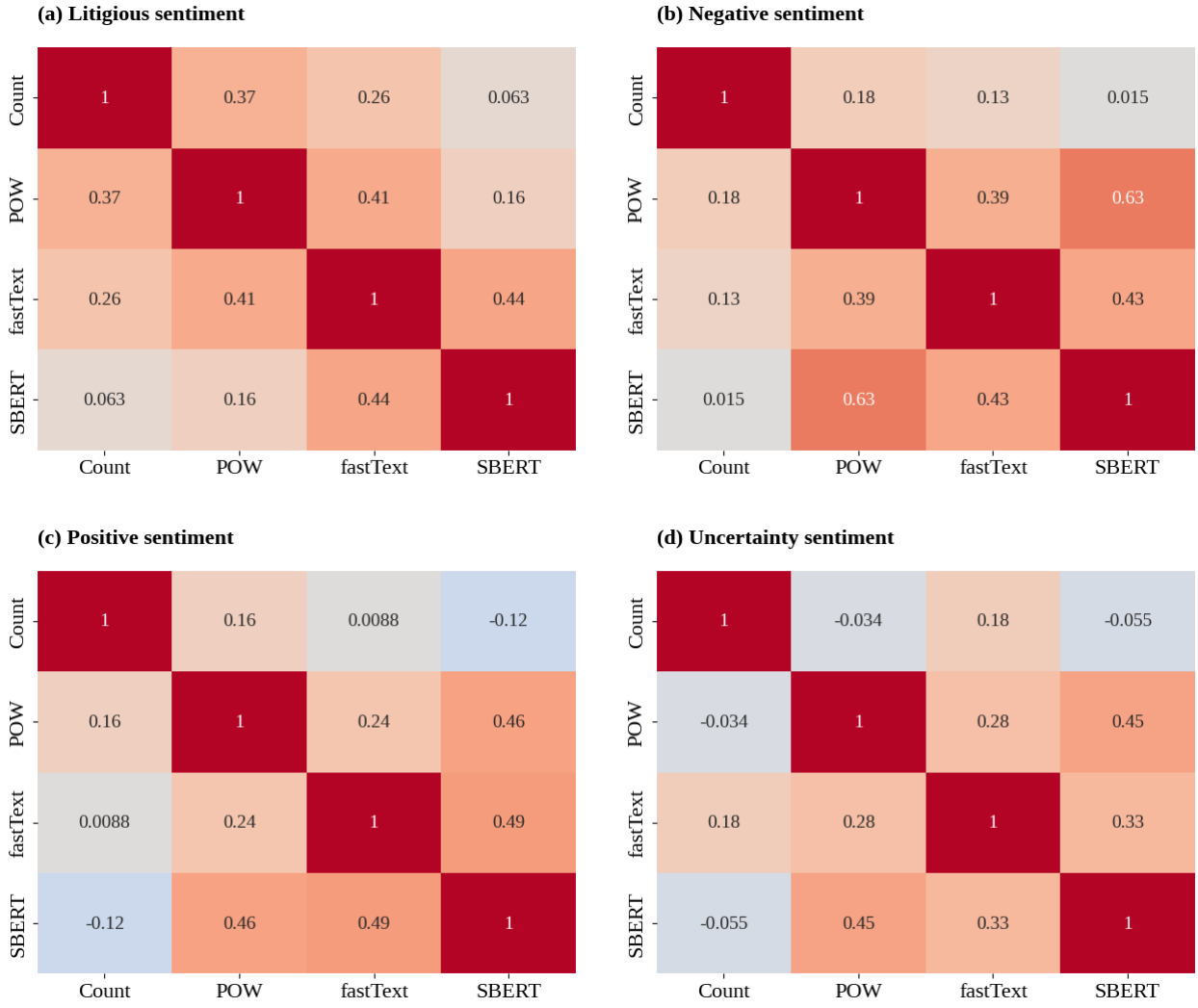

Figure 4: Correlation heatmaps for the sentiment variable grouped by sentiment

This figure presents the pairwise correlation of the different variables for a given sentiment. The heatmap shows the correlations between the different types of variables used to extract this sentiment. "Count" represents the word count of the LM word list of the respective sentiment.

larity, as explained in Section 3.4. To measure document similarity, two methodologies are employed: BOW (cosine similarity) and simBERT. Both metrics are based on cosine similarities, comparing document vectors and sentence vectors respectively. Consequently, the maximum similarity value is capped at 1, indicative of two identical entities. As both metrics depend on a reference set of documents from a specific period preceding the listing for comparison, not every document in the sample was assigned a similarity value. Therefore, the total sample size is reduced to 695 observations, as shown in Table 14 in the Appendix, which is used as a reference in this chapter.

The BOW variable is characterized by a mean value of 0.697 a median of 0.706 and a standard deviation of 0.055. Differences between mean and median are caused by outliers, with the minimum as low as 0.392 and the maximum as high as 0.815. IPOs show higher similarity scores, with mean and median values at 0.700 and 0.707, respectively. In contrast, private placements register values of 0.685 (mean) and 0.697 (median). The second metric, simBERT records a mean of 0.773, a median of 0.774, and a notably smaller standard deviation of 0.015 compared to BOW (cosine similarity). When analyzing the relationship between IPOs and private placements for simBERT, it can be observed that IPOs demonstrate slightly lower similarities. This is reflected in the mean and median, both with values of 0.773 for IPOs, as opposed to 0.777 and 0.779 for private placements, respectively. Based on the different ranking of both listing types, it can be concluded that BOW (cosine similarity) and simBERT show differing effectiveness in measuring document similarity. A possible explanation is the higher accuracy of contextualized embedding techniques in determining the semantic meaning of text, which is more important predictor than usage of words. Breitung and Müller (2022, p. 28) provide statistics for both variables in their research. The simBERT score they report is in a close range, averaging at 0.79. How-

ever, the BOW measure yields a substantially higher average document similarity, registering a mean of 0.93. Since Breitung and Müller (2022, p. 11), compare two consecutive annual reports from the same company, it is probable that the report was authored by the same accounting firm or even by the same individuals. This would account for the recurrence of similar words across the documents, which explains high BOW similarities. In general, in their analysis, similarities are expected to be higher as numerous items are expected to remain constant across two reports from the same firm.

Figure 5 shows the correlations between document size and discussed similarity variables. While not a feature in the regression models, document size is used as a proxy for information content, as noted by Loughran and McDonald (2016, p. 1223) and Guo et al. (2022, p. 2). The findings align logically, showcasing inverse correlations between document size and both similarity metrics. BOW and simBERT exhibit a positive correlation of 0.354.

As noted in Hanley and Hoberg (2010, p. 2849) documents with richer information content correlate with significantly higher listing expenses, highlighting the costs of information production. Figure 11 in the Appendix visually supports this trend, plotting log-transformed listing expenses against similarity measures, BOW and simBERT. The observed relationship is coherent, indicating that higher similarity scores are negatively related to listing expenses. These similarity variables, thus, offer valid insights for exploring the relationship between underpricing and information revelation. The subsequent chapter will present the regression models including a discussion of the results.

5. Empirical results

5.1. Model Design

The structure of the regression models used in this study is derived from related research in the field of textual analysis and IPO underpricing. For the sentiment analysis, the empirical models of Loughran and McDonald (2013) as well as Guo et al. (2022) serve as the most relevant benchmarks. Both studies explore the relationship between sentiment word lists and underpricing. For the similarity analysis, the outlined model for the study of information revelation and underpricing by Hanley and Hoberg (2010) is used as a reference.

The OLS regression models presented in the subsequent chapters all follow a consistent structure. The dependent variable in each model is underpricing. Each model incorporates an intercept, a single textual variable, a consistent set of base predictors, fixed effects (FE), and employs clustered standard errors. The regression tables include below the coefficients in parentheses the t-statistics for the respective variable. The set of base predictors remains constant across all models and encompasses the following control variables: Log(Employees), Pre-file NASDAQ return, Tech company dummy, Regulated market dummy, and IPO offering dummy. These variables are defined in detail in Chapter 4.3. Other predictors used in Loughran and McDonald (2013, p. 7) are not included in this study because of limited data availability. Examples are share overhang and upward revision of the pricing range, which are defined as the number of retained shares divided by the number of issued shares and the percentage upward revision from the mid-point of the filing range, respectively. Particularly, upward revisions account for a large part of the explained variance in the models of the reference study. However, in contrast to the U.S., in Europe offer price revisions are much less common (Jenkinson et al., 2006). Thus, in this study, it is anticipated that the impact of upward revisions would be less pronounced. To ensure that the estimates are not biased due to trends in specific variables, FE are included for the IPO year, FF12, and the financial market authority (FMA). FMA introduces country-specific fixed effects based on the country of origin of the FMA responsible for regulating the issuing company. Furthermore, to account for phenomena like hot- or cold-issue periods or sector-specific market dynamics, the displayed standard errors are clustered by IPO year and FF12. The same fixed effects and robust standard errors are applied in all the conducted regression models. The following section will describe the results of the regression models for POW, fastText, SBERT and similarity variables.

5.2. Regression Results

5.2.1. Percentage-of-Word Variables

This chapter examines the relationship between the POW sentiment variables and IPO underpricing. Table 6 displays the results of the multivariate regressions for each sentiment word list. The observed values for R^2 are lowest for litigious and negative sentiments, in columns (1) and (2), both with values of 0.076. Positive-POW in column (3) has a value for R^2 of 0.077. The highest level of explained variance can be attributed to column (4), which includes the uncertainty word list variable and shows a value of 0.080. These values are considerably lower than those reported in the studies by Loughran and McDonald (2013, p. 7) and Guo et al. (2022, p. 7). The lower R^2 values can be attributed to predictors that were used in the referenced studies but were omitted in this study. Among the base predictors, only the tech company and the regulated market dummy variables consistently display significant coefficients across the columns. Both variables are significant at the 1% level in each regression. The tech company dummy variable has significant positive coefficients, ranging from 0.118 in column (3) to 0.126 in column (4). Similarly, the regulated market dummy variable has positive coefficients, with values ranging from 0.065 in column (1) to 0.069 in column (4). The firm size proxy represented by Log(employees), the market sentiment indicated by the Pre-file NASDAQ return, and the IPO dummy variable all show insignificant coefficients. This aligns partly with Loughran and McDonald (2013, p. 7) findings, where size proxy, expressed as Log(sales), does not consistently display significant values. In contrast, the market sentiment variables are consistently significant in their study and in that of Guo et al. (2022, p. 6). This underscores the conjecture

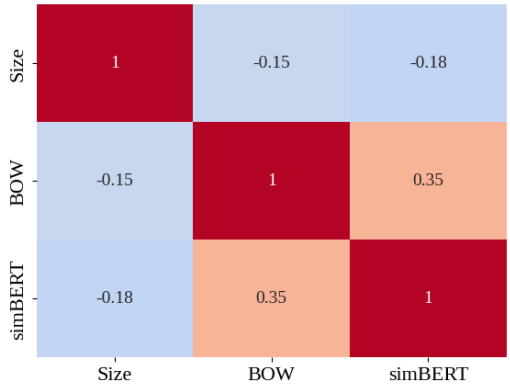

Figure 5: Correlation heatmaps for the similarity variables

Size is the document size measured as number of words within a listing document. The heatmap shows pairwise correlations between the set of variables.

of the limited importance of this variable for European IPOs or the current business environment. The last finding is surprising, as the average underpricing for IPOs is considerably lower for private placements, the impact of the IPO dummy variable remains insignificant. This could potentially be attributed to the mediating influence of one or more other control variables.

Regarding the POW variables, columns (1) through (3) yield insignificant results for litigious, negative, and positive sentiments. It is only in column (4) that the uncertainty sentiment variable becomes statistically significant with a t-statistic of 1.820. A one-standard deviation increase in the uncertainty variable is associated with an economically significant increase in underpricing by 0.031 (derived from multiplying the standard deviation of 0.341 with a coefficient of 0.091). This effect is nearly equivalent to the value presented by Loughran and McDonald (2013, p. 8), which is documented at 0.033. In contrast, Guo et al. (2022, p. 7) report a one-standard deviation increase in the percentage of uncertain words leads only to an increase of 0.012 in underpricing. Moreover, the findings of Loughran and McDonald exhibit statistically significant positive relationship for uncertainty, weak modal, and negative word lists. A finding echoed by Guo et al. (2022), albeit only for negative and uncertain sentiments. The results presented in Table 6, however, do not validate the positive effect of higher frequencies of negative words on underpricing. Given that no significant coefficients are reported, the results from the POW variables do not provide support for the conjectured impact of litigious, positive, and negative sentiments on underpricing. Nonetheless, the findings from the reference studies regarding the relationship between uncertain sentiment and underpricing can be confirmed. The subsequent chapters explore the hypothesis that the updated methodologies might better capture textual semantics and provide a clearer picture between prospectus content and underpricing.

5.2.2. FastText Variables

To evaluate the potential of the revised methodology, using fastText word embeddings, the regression models from the previous chapter are repeated, replacing POW with fastText variables. Table 7 presents the results of this set of multivariate regressions. On the basis of the R^2 values, the explanatory power of the models in Table 7 is slightly higher than in Table 6. Specifically, the R^2 value for litigious sentiment remains consistent across both tables, registering at 0.076. The R^2 value for negative sentiment, with a value of 0.077, records a marginal increase. Notably, the variable accounting for negative sentiment records the most substantial R^2 value of 0.088, representing an increase of 0.011 relative to the values observed in Table 6. This is subsequently followed by the value of uncertainty-POW in column (4), which stands at 0.081. The results for the base predictors mostly align with the findings from Table 6. The tech company and regulated market dummy variables continues to display significant positive associations with underpricing. Log(Employees), the pre-file NASDAQ return, and the IPO dummy variable again yield insignificant coefficients. In the presence of these control variables, it is observed that half of the fastText variables exhibit coefficients that are statistically insignificant. This includes the litigious sentiment variable illustrated in column (1) and the negative sentiment variable depicted in column (2). In column (3), contrary to the results from Table 6, the coefficient for the positive-POW variable is statistically significant (t-statistic of 5.757). A one-standard deviation increase in this variable corresponds to a 0.043 rise in underpricing (0.014 standard deviation, 3.097 regression coefficient). The economic magnitude is similar to those of the significant variables found in the study by Loughran and McDonald (2013, p. 8). Consistent with Table 6, the uncertainty variable remains statistically significant, with a t-statistic of 1.816. The economic significance stands at 0.030, derived from multiplying the standard deviation of 0.013 by the positive regression coefficient of 2.205. This outcome closely mirrors those from the POW approach and the study by Loughran and McDonald (2013).

Table 6: Multivariate regressions models for POW variables

Underpricing	**(1)**	**(2)**	**(3)**	**(4)**
Litigious-POW	0.057			
	(0.983)			
Negative-POW		-0.019		
		(-0.380)		
Positive-POW			0.073	
			(0.961)	
Uncertainty-POW				0.091*
				(1.820)
Log(Employees)	−0.009	−0.008	−0.009	−0.008
	(−1.125)	(−1.000)	(−1.125)	(−1.333)
Pre-file NASDAQ return	0.476	0.477	0.479	0.472
	(1.021)	(0.996)	(1.002)	(1.011)
Tech company (D)	0.119***	0.119***	0.118***	0.126***
	(4.958)	(5.174)	(5.130)	(6.300)
Regulated market (D)	0.065***	0.067***	0.067***	0.069***
	(3.250)	(3.526)	(3.722)	(3.450)
IPO offering (D)	−0.036	−0.056	−0.048	−0.021
	(−0.529)	(−0.836)	(−0.750)	(−0.412)
Year FE	Included	Included	Included	Included
FF12 FE	Included	Included	Included	Included
FMA FE	Included	Included	Included	Included
Observations	707	707	707	707
R2	0.076	0.076	0.077	0.080

This table shows the regression models of underpricing as explanatory variables with the standard set of control variables and one POW sentiment variable in the respective columns (1) to (4). The t-statistics are presented in parentheses. ***, ** and * indicate significance on the 1%-, 5%- and 10%-levels, respectively.

As a result, the findings displayed in Table 7 demonstrate that word embeddings can serve as an effective tool to enhance the analysis of document tone. In the following chapter, the technologically advanced, SBERT model, is employed to further investigate the relationship between sentiment variables and underpricing.

5.2.3. Sentence-BERT Variables

The usage of fastText variables has already proven to be more effective in terms of the number of significant variables and explained variance in the regression model. This chapter now explores the relationship between SBERT variables, using technologically more advanced contextualized word embeddings, and underpricing.

The significant set of predictors has a positive impact on the R^2 values of the models, all of which demonstrate a higher explanatory power compared to those in Tables 6 and 7. Specifically, the R^2 values sequentially for columns (1) through (4) are 0.088; 0.087; 0.090; and 0.090, respectively. The results for the base predictors align closely with previous tables. Examining the coefficients of the SBERT variables, notable differences in comparison to Tables 6 and 7 can be observed. In column (1), the litigious-SBERT has a coefficient of 3.481 with a t-statistic of 3.430. A one-standard deviation increase is associated with an 0.042 increase in underpricing (based on a standard deviation of 0.012). For the negative sentiment variable, the reported coefficient is 2.588, with an associated t-statistic of 3.034. A one-standard deviation increase for this predictor corresponds to an approximate 0.041 rise in underpricing (based on a standard deviation of 0.016). Contrary to previous models, both litigious and negative sentiments demonstrate a positive and statistically significant relationship with underpricing. For SBERT, the positive sentiment variable remains significant with a t-statistic of 2.125 and exhibits a coefficient of 2.901. Given its standard deviation of 0.018, a one-standard deviation higher variable value translates into an economically significant rise in underpricing by 0.052. Column (4) assesses uncertainty-SBERT, which has a coefficient of 2.546 (t-statistic = 2.714). An increase of one standard deviation in the uncertainty sentiment is associated with an increase in underpricing by about 0.039 (2.546 multiplied by a standard deviation of 0.015). The presented results match some of the findings from the previous chapters. Similar as in the context of fastText-based regressions, positive and uncertain sentiment are significant predictors, with positive sentiment showing the most pronounced effect. Contrary, through the adoption of contextualized embeddings both litigious and negative variables turned significant. This demonstrates a novel insight compared to the

Table 7: Multivariate regressions models for fastText variables

Underpricing	**(1)**	**(2)**	**(3)**	**(4)**
Litigious-fastText	0.473			
	(0.423)			
Negative-fastText		1.399		
		(1.345)		
Positive -fastText			3.097***	
			(5.757)	
Uncertainty-fastText				2.205*
				(1.816)
Log(Employees)	−0.008	−0.008	−0.013	−0.006
	(−1.000)	(−1.143)	(−1.625)	(−0.857)
Pre-file NASDAQ return	0.476	0.476	0.497	0.477
	(1.004)	(1.026)	(1.096)	(1.053)
Tech company (D)	0.119***	0.119***	0.117***	0.122***
	(6.263)	(5.409)	(4.680)	(6.100)
Regulated market (D)	0.066***	0.067***	0.079***	0.061***
	(3.667)	(3.190)	(4.158)	(3.813)
IPO offering (D)	−0.051	−0.050	−0.059	−0.053
	(−0.813)	(−0.850)	(−0.952)	(−0.917)
Year FE	Included	Included	Included	Included
FF12 FE	Included	Included	Included	Included
FMA FE	Included	Included	Included	Included
Observations	707	707	707	707
R2	0.077	0.077	0.088	0.081

This table shows the regression models of underpricing as explanatory variables with the standard set of control variables and one fastText sentiment variable in the respective columns (1) to (4). The t-statistics are presented in parentheses. ***, ** and * indicate significance on the 1%-, 5%- and 10%-levels, respectively.

findings by Loughran and McDonald (2013) and Guo et al. (2022). The final chapter of this section with regression results describes the relationship of document similarities on underpricing.

5.2.4. Similarity Variables

The similarity analysis aims to understand how information revelation in the prospectus impacts underpricing levels. Both the BOW (cosine similarity) and simBERT methods assess the similarity of a given prospectus compared to those of prior comparable listings. Considering the cost associated with information production, it is suggested that higher similarity levels are associated with more underpricing. The regression models follow the identical structure as those for the similarity analysis. The findings are described in Table 9. Column (1) displays the results of BOW (cosine similarity) and column (2) includes the model with simBERT as the respective textual variable.

Neither BOW (cosine similarity) nor simBERT exhibits a significant coefficient. The BOW variable has a coefficient of 0.213, while simBERT has a coefficient of 1.150, which are associated with insignificant t-statistics of 0.653 and 1.067. The results from this study fail to confirm the findings by Hanley and Hoberg (2010, p. 2848). The researchers reported a positive and significant relationship between standard content (i.e., content that is similar across prospectuses) and underpricing. This relationship is represented by a one-standard deviation increase in standard content, which translates into a 6% increase in underpricing. The variable construction used in their study follows a different methodology, however, the measured construct is expected to be similar, as previous analysis has indicated.

While the predictors are not statistically significant, the R^2 values for BOW (cosine similarity) and simBERT are comparable to those observed for SBERT sentiment variables, registering at 0.086 and 0.087, respectively. This suggests that some explanatory power of the variables is present. However, due to the insignificance of the predictors no meaningful conclusion can be drawn. It is possible that a larger sample size might reveal a significant relationship between document similarity and underpricing in European IPOs. However, for the given context no relationship between information revelation and underpricing can be established. The subsequent chapter elaborates the interpretations of the findings from this and previous chapters, contextualizing them within the framework of the developed hypotheses.

5.3. Interpretation of Regression Results

This section links the results from our regression analyses with the hypotheses outlined in Chapter 2.4, relating

Table 8: Multivariate regressions models for SBERT variables

Underpricing	**(1)**	**(2)**	**(3)**	**(4)**
Litigious-SBERT	3.481***			
	(3.430)			
Negative-SBERT		2.588***		
		(3.034)		
Positive-SBERT			2.901**	
			(2.125)	
Uncertainty-SBERT				2.546***
				(2.714)
Log(Employees)	−0.008	−0.007	−0.010	−0.006
	(−1.143)	(−1.000)	(−1.111)	(−1.000)
Pre-file NASDAQ return	0.477	0.473	0.504	0.472
	(1.089)	(1.068)	(1.086)	(1.068)
Tech company (D)	0.112***	0.120***	0.105***	0.119***
	(4.870)	(5.455)	(4.375)	(5.174)
Regulated market (D)	0.073***	0.072***	0.084***	0.071***
	(3.650)	(3.600)	(4.941)	(3.737)
IPO offering (D)	−0.056	−0.036	−0.060	−0.047
	(−1.018)	(−0.667)	(−1.034)	(−0.870)
Year FE	Included	Included	Included	Included
FF12 FE	Included	Included	Included	Included
FMA FE	Included	Included	Included	Included
Observations	707	707	707	707
R2	0.088	0.087	0.090	0.087

This table shows the regression models of underpricing as explanatory variables with the standard set of control variables and one SBERT sentiment variable in the respective columns (1) to (4). The t-statistics are presented in parentheses. ***, ** and * indicate significance on the 1%-, 5%- and 10%-levels, respectively.

underpricing to key academic theories. The ex-ante uncertainty hypothesis (H1) is supported by our findings. A clear link between uncertain sentiment in IPO prospectuses and anticipated underpricing is observable. Every methodological approach yielded a positive, statistically significant outcome, with the economic magnitude paralleling the findings of Loughran and McDonald (2013). This result is notable when considering that underpricing in the reference study was significantly higher. It is important, however, to note that some variables with high explanatory power from the reference study were omitted. If included, these variables could influence the effect's magnitude. Nevertheless, the results robustly support the theory that uncertain language in a prospectus, is a good proxy for ex-ante uncertainty. In line with the theory of Beatty and Ritter (1986), the uncertainty explains a notable portion of the underpricing observed in our sample highlighting the heightened risk and valuation uncertainty for investors associated with the listing.

The legal liability hypothesis (H2) is validated through the SBERT-based sentiment variable, revealing a positive connection between ex-ante litigation risk and underpricing. This suggests that legal language in the prospectus can effectively represent a company's ex-ante litigation risk. Importantly, it appears that this risk is not simply recognized though word counts or individual word meanings (as obtained from classical word embeddings) but is represented through specific contexts. This explains, that only the SBERT methodology can establish a significant relationship with underpricing. The results emphasize the theory that companies with elevated levels of ex-ante litigation risks employ higher underpricing as a strategy to deter lawsuits, a finding emphasized by Lowry and Shu (2002). It is important to acknowledge that scores for litigious sentiment are strongly correlated with those of negative and uncertain sentiments. It appears that the selected methodologies do not allow to make a clear distinction between legal risk and general firm uncertainty. Determining whether the results can be definitively attributed to either the ex-ante uncertainty or the legal liability theory is challenging, as both appear plausible. To gain a clearer understanding, data concerning post-IPO litigation would be essential.

The neutral language hypothesis (H3) is confirmed in this study. The positive sentiment variable shows significance for both fastText and SBERT models, with the negative sentiment variable significant only when based on the SBERT methodology. The original hypothesis of Ferris et al. (2013), is extended with the idea of investors' skepticism towards excessive positive language in prospectuses. The link between un-

Table 9: Multivariate regressions models for similarity variables

Underpricing	(1)	(2)
BOW (cosine similarity)	0.213	
	(0.653)	
simBERT		1.150
		(1.067)
Log(Employees)	−0.009	−0.011
	(−1.125)	(−1.222)
Pre-file NASDAQ return	0.472	0.475
	(0.925)	(0.942)
Tech company (D)	0.112***	0.108***
	(5.600)	(5.143)
Regulated market (D)	0.076***	0.079***
	(4.000)	(3.591)
IPO offering (D)	−0.062	−0.058
	(−1.107)	(−1.160)
Year FE	Included	Included
FF12 FE	Included	Included
FMA FE	Included	Included
Observations	671	671
R2	0.086	0.087

This table shows the regression models of underpricing as explanatory variables with the standard set of control variables and one similarity variable in the respective columns (1) to (2). The t-statistics are presented in parentheses. ***, ** and * indicate significance on the 1%-, 5%- and 10%-levels, respectively.

derpricing and the positive sentiment variables of fastText and SBERT suggest that classical and contextual word embeddings can accurately detect positive sentiment in documents. An aspect that, as noted by Guo et al. (2022, p. 7), citing Tetlock (2007), is difficult to be captured by conventional word count methods. The analysis corroborates the association between neutral language and underpricing. In this context, the findings support the conjecture that investors favor neutral language as a trust-building mechanism.

The information revelation hypothesis (H4) could not be confirmed. This suggests that underwriters do not have the ability to influence underpricing by revealing more information in the IPO prospectus. It appears to be the case that this inconsistency is not related to measurement, since both similarity metrics align with patterns observed in Hanley and Hoberg (2010), as shown in Figure 11. Moreover, it can be concluded that the anticipated positive impact of higher document similarities on underpricing is not evident in this sample. The findings suggest that underpricing is more reflective of specific firm characteristics than of the effort issuers put into pre-market due diligence and information disclosure. As a result, this finding strengthens hypotheses H1, H2, and H3, indicating that the contents of the prospectus are predominantly utilized to evaluate the risk profile of a firm and the credibility of this information, which in turn impacts the discount required on the expected firm value. In contrast, the results contradict the findings of Hanley and Hoberg (2010) and Hanley and Hoberg (2012), who found that increased disclosure has a negative effect on the level of underpricing. This conclusion resonates with observations on the financial magnitude of underpricing for certain listings, which can exhibit first-day returns as high as 200% in extreme cases. It appears unrealistic that firms willingly leave that much money on the table if they would have the option to reveal more information in the IPO prospectus.

The methodological approach hypothesis (H5) is supported by the findings in Tables 7 and 8. For the neural network hypothesis (H5.1), the fastText sentiment variables consistently exhibit higher values of R^2 in comparison to the POW variables in Table 6. While uncertain sentiment is significant in both models, positive sentiment also demonstrates a significant correlation with underpricing for the fastText variable. This underscores the constraints of word lists. They are frequently defined too restrictively and miss the semantic essence of words not covered in the dictionary. Similarly for the transformer model hypothesis (H5.2), which is supported again by higher levels of explained variance in Table 8 and significant coefficients for all textual variables. However, the similarity analysis results in Table 9 fail to validate the hypothesis. Here, both the BOW, rooted in word counts, and simBERT produce insignificant outcomes. It is probable that this is not a limitation of the methodology but could rather be attributed to sample characteristics or a lack of causal linkage among the variables. The impact of negative language on underpricing, which was found to be statistically significant in the studies by Loughran and McDonald (2013) and Guo et al. (2022), could only be confirmed by the SBERT methodology. As previously mentioned, the challenges faced

by word count methods in capturing positive sentiment were also addressed by fastText and SBERT. The elevated correlations of these newly developed approaches with the POW method also acts as a sanity check, ensuring the accurate measurement of the intended concepts. These results support the validity and efficacy of both methods developed for this study.

6. Conclusion

This study establishes a clear link between the language used in prospectuses of European IPOs and underpricing. Therefore, traditional word count-based methods for sentiment and similarity analysis, as conceptualized by Loughran and McDonald (2013) and Hanley and Hoberg (2010), are augmented and refined. This is achieved by leveraging recent advances in the field of NLP and making use of neural network-based word embeddings and transformer-based language models. FastText and SBERT, respectively, are chosen as the most suited models for this purpose. For the sentiment analysis, this study relies on a subset of the sentiment categories defined in Loughran and McDonald (2011), namely litigious, negative, positive and uncertainty sentiment. The POW approach measures document tone by simply totaling the frequencies of words from the corresponding sentiment word list. In contrast, for the fastText and SBERT methodologies, cosine similarity is employed to ascertain how closely the sentiment of a given document aligns with the sentiments from the LM word lists. Given that these word lists were originally developed for word count methods, they often include multiple words for the same word group or connecting words that lack standalone meanings. This study introduces an updated version, refining these lists to a concise dictionary that captures the most important topics within each word list. Notably, for SBERT – which exhibits optimized results with full sentence inputs – the reduced word lists are augmented into a dictionary of short sentences. Based on the geometrical representations obtained from both document and word or sentence dictionary, cosine similarity is used to compute the sentiment score. The resulting scores express how similar the text of a prospectus is compared to the dictionary. The results of the newly developed sentiment measures are benchmarked against the results of the POW approach. The findings support the validity of the extended methodologies.

The similarity analysis is based on the BOW approach of Hanley and Hoberg (2010) and the simBERT methodology of Breitung and Müller (2022). For both approaches, the prospectus of a given company is compared to related documents from comparable listings. Date and industry filters are used to identify the comparable items. Similarity scores are again determined using cosine similarity. The BOW (cosine similarity) measure makes use of normalized vectors of word frequencies, while the simBERT score determines document similarity by averaging the maximum cosine similarities across all sentence pairs. The values obtained from the simBERT methodology are coherent with those of the reference study and correlate with BOW (cosine similarity). The findings show that simBERT can be used as a robust alternative to the traditional BOW method.

The results of the regression analyses can be used to explain several dimensions of IPO underpricing. Evidently, the relationship between uncertain sentiment in IPO prospectuses and underpricing is substantive. The findings mirror those of Loughran and McDonald (2013) and are consistent across all variable groups. The significant relationship between legal language and underpricing allows to successfully link prospectus sentiment to the legal liability hypothesis of Lowry and Shu (2002). To properly detect legal sentiment, the document's context is important which requires the application of SBERT-based sentiment scoring. Inspired by the theory proposed by Ferris et al. (2013), a link between the neutrality of prospectus language (impacting the perceived credibility) and underpricing is identified. In this study, neutrality is defined as avoiding negative and positive language and is anticipated to reduce levels of underpricing. The neutrality theory is corroborated as there is a statistically significant positive relationship observed between both negative and positive sentiments and underpricing. This indicates that investors are more inclined to trust information when it is conveyed in a neutral and objective manner, without hyperbolic expressions. However, for the similarity analysis, neither the BOW (cosine similarity) nor the simBERT scores deliver significant results. Consequently, the information revelation hypothesis, which suggests that underpricing serves as a compensation mechanism for investors in exchange for revealing information, based on the theory of Benveniste and Spindt (1989), cannot be confirmed in this study. This suggests that underwriters do not have the control to make the choice between committing more resources to information production or employing more underpricing as an incentive for investors to reveal their private information truthfully.

As a critical analysis of the methodologies employed, it is essential to recognize that interpreting sentiment scores can prove difficult. The utilization of neural network-based word embeddings and transformer-based language models adds a layer of complexity that makes interpretation in some parts impossible. These models represent text in high-dimensional vector spaces. Thus, attributing a specific meaning to the individual values contained within such a vector is impossible from a human perspective. Although the developed metrics show a positive correlation with conventional word count approaches, subjective evaluations cannot rule out the possibility that the improved sentiment measurement is due to another confounding factor present in both the vector representations of the sentiment dictionary and the prospectus. Additionally, Kaserer and Treßel (2023, p. 18) note that the translation of documents using Google Translate might distort the intended meaning of the prospectus passages. This could have a detrimental impact on the effectiveness of our methodologies and introduce biases, as they rely on semantic interpretations of words rather than do traditional word count methods. Furthermore, in contrast to the reference studies on textual analysis and underpricing, the

current research faces certain data constraints. The listings from exchange-regulated markets in this study display less comprehensive data coverage in financial databases. Therefore, a considerable portion of the utilized variables were hand-selected for the study conducted by Kaserer and Treßel (2023). Since the scope of this research was limited, it was not possible to extend the dataset with the missing control variables used in the reference studies. Another limitation based on the missing variables, is the overall explanatory power of the models which all show limited levels of explained variance. Therefore, it is essential to acknowledge that the results obtained in this study should be interpreted accordingly and may not fully capture the complexity of the relationships between textual variables and underpricing in European IPOs.

In synthesizing the findings and insights from this research, several directions for future research are presented, serving as a conclusion to this study. Firstly, the results of this study can be replicated by using language models that have been specially trained on financial domain-specific language. This could reduce potential biases from misinterpretations of words that convey a different meaning in general than in financial language. This suggestion follows the idea of LM word lists, which, as mentioned in Loughran and McDonald (2011), were introduced specifically to avoid this type of bias. In addition, it could be beneficial to extend the sentence dictionaries used for sentiment analysis with SBERT with different sentence examples. Here, it might be useful to take advantage of recent advances in generative AI to automate the generation process. Furthermore, to affirm the robustness of the findings in this study, it would be valuable to replicate it using a more extensive dataset. This replication should ideally account for variables that were omitted in the present study. Lastly, drawing inspiration from the study conducted by Cao et al. (2023), it would be compelling to explore how the increased adoption of textual analysis methods has prompted underwriters to modify the language used in prospectuses. In this context, it could be investigated if a link between adopted language and underpricing exists. The publication dates of word list methods or the release dates of language models could be used to create an experimental setting.

References

Abrahamson, M., Jenkinson, T., & Jones, H. (2011). Why Don't US Issuers Demand European Fees for IPOs? *The Journal of Finance*, *66*(6), 2055–2082.

Allianz Global Corporate & Specialty SE. (2020). Collective Actions and Litigation Funding and the Impact on Securities Claims: A Global Snapshot.

BaFin. (2023). Prospectus Requirement. Retrieved April 5, 2023, from https://www.bafin.de/EN/Aufsicht/Prospekte/Wertpapiere/Prospektpflicht/prospektpflicht_node_en.html;jsessionid=141A9E291A85A19245F799B4B2D7785C.1_cid500

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.

Benveniste, L. M., & Spindt, P. A. (1989). How Investment Bankers Determine the Offer Price and Allocation of New Issues. *Journal of Financial Economics*, *24*(2), 343–361.

Berk, J., & DeMarzo, P. (2019). *Corporate Finance, Global Edition*. Pearson Education, Limited.

Brau, J. C., Cicon, J., & McQueen, G. (2016). Soft Strategic Information and IPO Underpricing. *Journal of Behavioral Finance*, *17*(1), 1–17.

Breitung, C., & Müller, S. (2022). *When Firms Open Up: Identifying Value Relevant Textual Disclosure Using simBERT* (Working Paper). TUM School of Management.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to Talk When a Machine is Listening?: Corporate Disclosure in the Age of AI. *The Review of Financial Studies*, *36*(9), 3603–3642.

Das, S. R., Donini, M., Zafar, M. B., He, J., & Kenthapadi, K. (2022). FinLex: An Effective Use of Word Embeddings for Financial Lexicon Generation. *The Journal of Finance and Data Science*, *8*, 1–11.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Doukas, J. A., & Hoque, H. (2016). Why Firms Favour the AIM When They Can List on Main Market? *Journal of International Money and Finance*, *60*, 378–404.

Drake, P. D., & Vetsuypens, M. R. (1993). IPO Underpricing and Insurance Against Legal Liability. *Financial Management*, 64–73.

European IPO Task Force. (2020). *European IPO Report 2020*.

Ewens, M., & Farre-Mensa, J. (2020). The Deregulation of the Private Equity Markets and the Decline in IPOs. *The Review of Financial Studies*, *33*(12), 5463–5509.

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

Gao, X., Ritter, J. R., & Zhu, Z. (2013). Where Have All the IPOs Gone? *Journal of Financial and Quantitative Analysis*, *48*(6), 1663–1692.

Geddes, R. (2003). *IPOs and Equity Offerings*. Butterworth-Heinemann.

Guo, H., Wang, Y., Wang, B., & Ge, Y. (2022). Does Prospectus AE Affect IPO Underpricing? A Content Analysis of the Chinese Stock Market. *International Review of Economics & Finance*, *82*, 1–12.

Hanley, K. W. (2017). *The Economics of Primary Markets* (Working Paper). New Special Study of the Securities Markets.

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.

Hoberg, G., & Phillips, G. (2010). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-based Analysis. *The Review of Financial Studies*, *23*(10), 3773–3811.

Huang, R., Ritter, J. R., & Zhang, D. (2023). IPOs and SPACs: Recent Developments [Forthcoming]. *Annual Review of Financial Economics*.

Huibers, F. E. (2020). Towards an Optimal IPO Mechanism. *Journal of Risk and Financial Management*, *13*(6), 115–129.

Jenkinson, T., Jones, H., & Suntheim, F. (2018). Quid Pro Quo? What Factors Influence IPO Allocations to Investors? *The Journal of Finance*, *73*(5), 2303–2341.

Jenkinson, T., Morrison, A. D., & Wilhelm Jr, W. J. (2006). Why Are European IPOs So Rarely Priced Outside the Indicative Price Range? *Journal of Financial Economics*, *80*(1), 185–209.

Kaserer, C., & Treßel, V. (2023). *The EU Prospectus Regulation and its Impact on SME Listings* (Working Paper). TUM School of Management.

Klausner, M., & Ohlrogge, M. (2023). Was the SPAC Crash Predictable? [Forthcoming]. *Yale Journal on Regulation Bulletin*.

La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The Economic Consequences of Legal Origins. *Journal of Economic Literature*, *46*(2), 285–332.

Lin, H. L., Pukthuanthong, K., & Walker, T. J. (2013). An International Look at the Lawsuit Avoidance Hypothesis of IPO Underpricing. *Journal of Corporate Finance*, *19*, 56–77.

Ljungqvist, A. (2007). Handbook of Empirical Corporate Finance. In *Handbooks in Finance* (pp. 375–422). Elsevier Science.

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

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*, 5–37.

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*, *15*(2), 413–444.

Lowry, M., Michaely, R., & Volkova, E. (2017). Initial Public Offerings: A Synthesis of the Literature and Directions for Future Research. *Foundations and Trends® in Finance*, *11*(3-4), 154–320.

Lowry, M., & Shu, S. (2002). Litigation Risk and IPO Underpricing. *Journal of Financial Economics*, *65*(3), 309–335.

Mandal, A., Ghosh, K., Ghosh, S., & Mandal, S. (2021). Unsupervised Approaches for Measuring Textual Similarity between Legal Court Case Reports. *Artificial Intelligence and Law*, 1–35.

Meden, K. (2022). *Semantic Similarity of Parliamentary Speech using BERT Language Models & fastText Word Embeddings* (Working Paper). Jožef Stefan Institute.

Michaely, R., & Shaw, W. H. (1994). The Pricing of Initial Public Offerings: Tests of Adverse-Selection and Signaling Theories. *The Review of Financial Studies*, *7*(2), 279–319.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in Pre-training Distributed Word Representations. *arXiv preprint arXiv:1712.09405*. https://arxiv.org/abs/1712.09405

Muscarella, C., & Vetsuypens, M. (1990). *Firm Age, Uncertainty, and IPO Underpricing: Some New Empirical Evidence* (Working Paper). Southern Methodist University.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://www.aclweb.org/anthology/D14-1162

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*. https://arxiv.org/abs/1908.10084

Ritter, J. R. (2023). Initial Public Offerings: International Insights. Retrieved August 5, 2023, from https://site.warrington.ufl.edu/ritter/files/IPOs-International.pdf

Ritter, J. R., & Welch, I. (2002). A Review of IPO Activity, Pricing, and Allocations. *The Journal of Finance*, *57*(4), 1795–1828.

Rock, K. (1986). Why New Issues Are Underpriced. *Journal of Financial Economics*, *15*(1-2), 187–212.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Seegmiller, B., Papanikolaou, D., & Schmidt, L. D. (2023). Measuring Document Similarity with Weighted Averages of Word Embeddings. *Explorations in Economic History*, *87*, 101494.

Sehrawat, S. (2019). Learning Word Embeddings from 10-K Filings for Financial NLP Tasks.

Sherman, A. E., & Titman, S. (2002). Building the IPO Order Book: Underpricing and Participation Limits with Costly Information. *Journal of Financial Economics*, *65*(1), 3–29.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

Tinic, S. M. (1988). Anatomy of Initial Public Offerings of Common Stock. *The Journal of Finance*, *43*(4), 789–822.

Torbira, L. L., & Oki, J. (2017). Determinants of Initial Public Offer Underpricing in the United Kingdom: Pre and Post Financial Crisis Evidence. *Research Journal of Finance and Accounting*, *8*(17), 31–59.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*.

Vismara, S., Paleari, S., & Ritter, J. R. (2012). Europe's Second Markets for Small Companies. *European Financial Management*, *18*(3), 352–388.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.