Dirk Laschat

Analyzing IPO-Prospectus Text Data during COVID-19

Term Paper

PhD course „Machine Learning“

Institute for Econometrics and Economic Statistics
(Westfälische Wilhelms-Universität, Münster)

Lecturer: Prof. Dr. Rainer Alexander Schüssler

Submission Date: 18.02.2022

Submitted by: Dirk Laschat

Institute for Strategic Management,
University of Münster

Tel.: +49 179 5999033
laschat@ism.uni-muenster.de

Matriculation number: 430765

Introduction

The IPO (initial public offering) prospectus, represents the most important document issued during the process of going public. It can be regarded as a kind of sales promotion towards potential investors. It is key for companies entering capital markets to overcome information asymmetries as far as possible to achieve the best pricing of their stock and collect as much capital as possible (Ljungqvist, 2009). However, it is important to note that in the majority of countries the document is filed with the local regulation authority and must meet certain guidelines. Thus, information disclosed in the IPO prospectus must be true, impartial and complete. On this way the document may reduce uncertainty and information asymmetries between issuer and investors (op. cit.).

Regarding these facts, the words used in an IPO prospectus can provide much information about a company's condition. Does the prospectus use a rather positive or negative language? Are there many words conveying uncertainty or constraints, or do opportunity and strength prevail? Which topics are discussed and to which extent? Is it rather easy or hard to read? The answers to all these questions can influence investor's attitude about the IPO and thus as well its success.

With COVID-19 disrupting societies as well as economies, our world entered a state of a new normal in many parts of our lives and uncertainty persists about when we can return to our familiar habits. There is the possibility that COVID also created a new normal of language used in IPO prospectuses. An intuitive assumption is, that the economic downturn created by the pandemic results in a more negative tone (sentiment) of prospectus documents and that there are more words conveying uncertainty and constraints. However, COVID could also rather replace former uncertainty instead of creating an additional one.

Thus, this term paper's objective is to find out more about the possible change in the language of IPO prospectuses due to COVID-19 and how language adapts to the new uncertainties created by the pandemic. Therefore, I first build language topic models

for the pre-COVID and COVID-period, to evaluate if topics discussed changed. Secondly, negative, positive, uncertain und constraining words as well as COVID-19 related words are identified and correlations are calculated to gain insights if these groups of words are rather complementary or substitutes. Lastly, sentiment scores are constructed with a dictionary approach and alternatively with the recently published FinBERT model to evaluate if both methods are consistent in the realm of COVID-period prospectus risk factors.

Related Literature and Contribution

Many scientific papers already examined the language of IPO prospectuses, most of them evaluating the phenomenon of IPO underpricing. Underpricing describes the fact that firms and their underwriters leave money on the table because they set the initial offering price too low, such that the price immediately jumps higher when trading starts. Literature finds four groups of reasons for underpricing: asymmetric information, institutional reasons, control considerations and behavioral approaches (Eckbo, 2009). The natural language processing papers mainly use the asymmetric information argument to evaluate underpricing, since investors can gain information through evaluating the way a prospectus is written.

The classical approach of these papers is to create one or more sentiment measures from the prospectus text data and examine its influence on the stock's first trading days returns. The earliest and most common approach is the work by Loughran and McDonald (2013), who classify words into uncertain, weak modal, negative, positive, legal, and strong modal categories and measure their influence on underpricing. Still, recent work is using this approach, like Ly and Nguyen (2020), however, they use the data to train different machine learning models and then forecast underpricing.

But also, text readability and text similarity approaches are used to explain IPO underpricing: Ding (2016), examines especially Australian IPO's and uses among others a variable for uniqueness of content. Their findings are in line with information asymmetry theory, which suggests that more disclosure decreases underpricing. Zhou et al.

(2020) also find that lower quality of an IPO's prospectus, as proxied by low readability and high similarity, produces higher IPO underpricing in Chinese IPOs from 2014-2018.

Another rather small topic is prospectuses' language influence on stock analysts. Here, the paper by Balakrishnan and Barov (2011) shows that while future earnings are correlated with the qualitative downside earnings risk information in the risk factors section, analyst earnings forecasts are not. This means that analysts do not incorporate the downside earnings risk into their evaluation of IPO prices.

By examining the language of prospectuses during COVID-19 descriptively and comparing sentiment scores of older and newer approaches, I contribute to the literature by identifying hints for changes in language due to COVID. Is there any need to adjust research methods or interpret results differently when looking at prospectus data from the COVID-periods? Are there signs of inconsistencies between newer sentiment measurements like BERT and the usual dictionary approaches, when examining the risk factor section of prospectuses?

Methodology

In a first step two topic models are created, to get an overview of the content discussed in the risk sections. Here, the data is divided into a pre-COVID and a COVID-period, to identify possible changes in topics discussed. The models are built via the LDA (Latent Dirichlet Allocation) method, which assumes documents are random mixtures over a predefined number of latent topics, where each topic is characterized by a distribution over words. The process of capturing the topics can be illustrated by Figure 1.

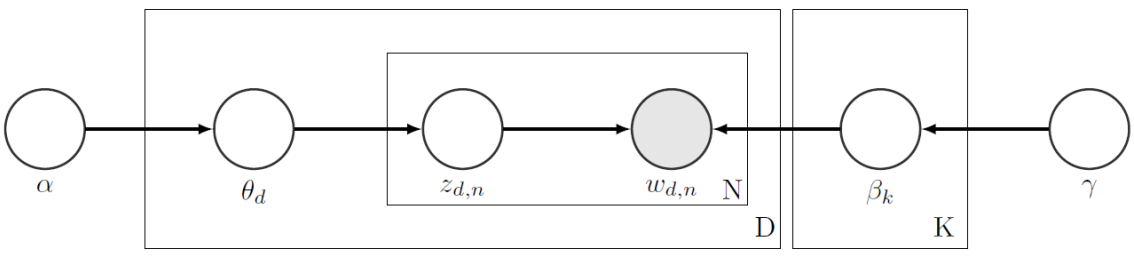

Figure 1: Schematic illustration of parameters and their effects in LDA models, Blei et al. (2003)

Here, α and γ are the parameters for the Dirichlet distributions. More precisely, α describes the prior distribution over topic weights in each document and γ describes the prior distribution over word weights in each topic. The betas, thus can be interpreted as topic probability vectors because each topic (beta) is a distribution over the observed words w. θ_d can be described as per-document topic mixing parameter, while $z_{d,n}$ is the topic assignment for the n^{th} word in document d (Blei et al., 2003).

The generative process for LDA corresponds to the following joint distribution of the hidden and observed variables (op. cit.):

$$p(\beta, \theta, z, w, \gamma, \alpha) = \prod_{k=1}^{K} p(\beta_k|\gamma) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)\, p(w_{d,n}|\, \beta_K, z_{d,n}, \beta_{k:K}) \right) \quad (1)$$

For the topic models to convey interpretable results, I use some additional clean up steps on top of the usual text pre-processing, like removing bullet points, double characters (like spaces, points and commas) and non-Unicode characters. These additional steps include the removal of punctuation, numbers, currencies and money related terms as well as country and country's institution names. Additionally, words are stemmed to remove pre- and suffixes and obtain better results. Finally, prospectus-specific frequently used words (like "prospectus", "section", "ordinary", "class", "stock" and "shares") are removed, because they do not contain a lot of information and could distort the topic models.

As a second part of analyzing IPO prospectuses' language words related to uncertainty, constraints, the COVID-19 pandemic as well as words with positive/negative connotation are counted via dictionary approach. The aim of this counting is to identify whether the pandemic had influence on the usage of these subgroups of words and the overall sentiment of the document's language.

To determine the pandemic words, I use IATE's (Interactive Terminology for Europe) "COVID-19 terminology" list, which contains around 1.700 words related to the current pandemic (IATE, 2021). The list is adjusted in such a way, that longer words, which overlap with already listed shorter words are removed to avoid double matching. Additionally, financial terms which might be used in context with the pandemic but are

also frequently used in pre-pandemic financial documents, are removed. To identify the remaining words, I use 2020's edition of Loughran & McDonald's Master Dictionary, which is the leading dictionary for financial sentiment analysis.

In a next step the groups of pandemic, uncertain and constraining words, are divided by a document's overall word number. With the help of positive and negative words a sentiment score for each document d is built in the following manner:

$$LMD\ Sentiment\ Ratio_d = \frac{Positive\ Words_d - Negative\ Words_d}{All\ Words_d} \quad (2)$$

For the third and last part of analyzing prospectuses' language, I calculate an alternative sentiment score via the pre-trained NLP model "FinBERT" by Yang et al. (2020). It is built by further training the BERT language model in the finance domain, using a large financial corpus and thereby fine-tuning it for financial sentiment classification. The basic BERT (Bidirectional Encoder Representations from Transformers) model is a deep learning model developed by Google. It is able to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2018). Meaning BERT and FinBERT analyze the sentiment of whole sentences and thus coherences between words are included into the calculation of a sentence's sentiment, which is calculated as follows:

$$Finbert\ Sentiment\ Ratio_d = \frac{Positive\ Sentences_d^{FinBERT} - Negative\ Sentences_d^{FinBERT}}{All\ Sentences_d} \quad (3)$$

Afterwards I perform a correlation analysis between all measures, to evaluate COVID-19's effect on word subgroup usage and the consistency of both sentiment measures during the pandemic.

Data

Since there is no single database for IPO prospectuses, I scraped the EU, UK databases and downloaded prospectuses manually for IPOs on selected Asian exchanges as well as the Australian ASX during the defined time period. On this way I was able to gather prospectuses for 221 IPOs from 36 countries, for this term paper.

Results

First, I evaluate the topic model's results to determine possible changes in topics discussed. To choose the optimal number of topics and learning decay as parameters for the topic model, the Log-Likelihood Scores for a combination of a hand full of intuitive parameters and both subperiods are compared. The Log-Likelihood measures how probable new data is for the model. It thus shows how well the model represents or reproduces the statistics of the held-out test set (Chen & Wang Yufei, 2016).

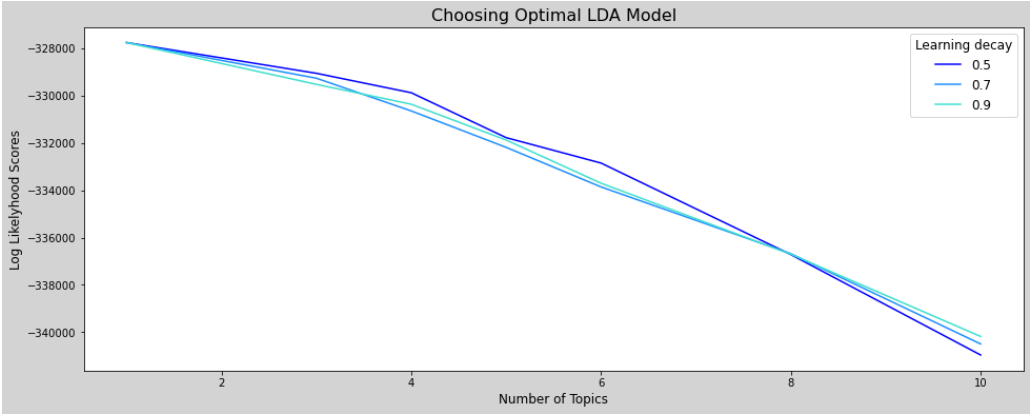

Figure 2: Evaluation of parameter choice for pre-COVID-period

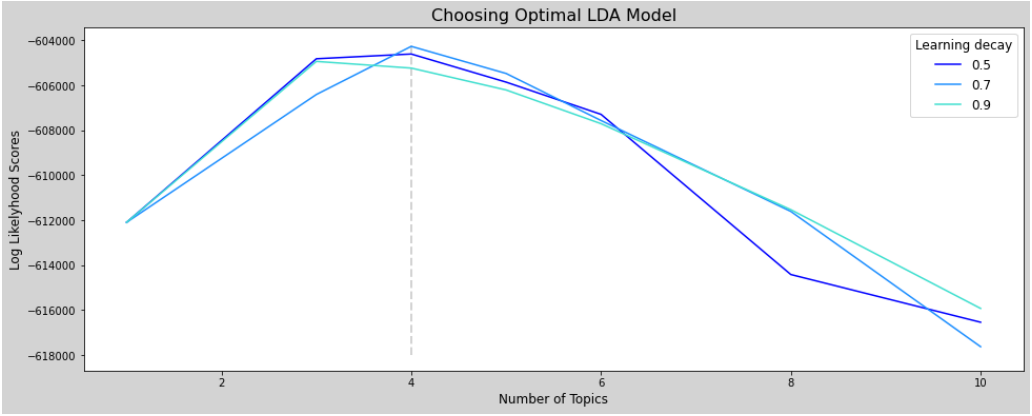

Figure 3: Evaluation of parameter choice for COVID-period

The results are shown in Figures 2 and 3, and suggest to use 4 topics with a learning decay of 0.7 for the COVID-period and only one topic for the pre COVID-period. Literature however shows that a high likelihood does not necessarily yield semantically coherent topics (Chang et al., 2009). That is why I evaluate different topic models with parameters around the optimum manually. After running the LDA model for topic parameters between three and six with a learning decay of 0.7 for both periods, I

decided to use five as number of topics for both periods due to its better interpretability with regards to the COVID-period. For the pre-period none of the tested models can be preferred interpretability wise and therefore the same number of topics as for the COVID-period was used. The corresponding words for both models are shown in Table 1 and 2 respectively.

	Pre-Topic 1	**Pre-Topic 2**	**Pre-Topic 3**	**Pre-Topic 4**	**Pre-Topic 5**
0	composit	liabl	composit	rapid	liabl
1	attack	progress	placement	defer	length
2	possess	qualiti	discov	defin	forecast
3	concern	length	pipelin	code	clean
4	go	oversea	deed	gdpr	monetari
5	faith	cancel	possess	monetari	go
6	destroy	defin	opinion	intang	attack
7	burden	magnitud	hack	invalid	defer
8	lesser	accru	bankruptci	liabl	amongst
9	interact	misappropri	clearanc	familiar	privaci
10	memorandum	equal	loyalti	concern	composit
...	...	...	...	...	...

Table 1: Topic model pre-COVID-period

	COVID-Topic 1	**COVID-Topic 2**	**COVID-Topic 3**	**COVID-Topic 4**	**COVID-Topic 5**
1	custom	properti	candid	project	combin
2	bank	custom	patent	portfolio	warrant
3	partner	enterpris	clinic	infrastructur	target
4	manufactur	data	trial	energi	sponsor
5	credit	reput	manufactur	net	holder
6	supplier	brand	licens	issuer	entiti
7	facil	resid	intellectu	shar	redempt
8	suppli	consum	properti	power	vote
9	pandem	intellectu	data	properti	per
10	flow	platfo	collabor	acquir	acquir
11	disrupt	supplier	patient	renew	unit
12	reput	record	enterpris	construct	consumm
13	action	subsidiari	enforc	note	privat
14	end	administr	medic	land	opportun
15	data	qualiti	research	trust	affili
16	loan	enforc	test	facil	founder
17	fiscal	entiti	administr	advis	trust
18	capac	fee	program	electr	member
19	litig	contractu	safe	flow	offic
20	raw	social	resid	realis	amend

Table 2: Topic model COVID-period

Regarding the pre-period, it is hard to label the topics the model yields. For the COVID-period however, topics seem to form more clearly. Especially Topic 1 seems to gather words related to the pandemic, because the stems 'disrupt' and 'pandem' are represented here. Supply chain related stems like 'suppli', 'facil', 'capac' and 'raw' are present just as financial terms like 'bank', 'credit', 'loan' and 'fiscal'. Additionally, stakeholders of the firm seem to be addressed many times in this 'pandemic' topic (see high ranking of the stems 'custom', 'bank', 'partner', 'manufactur', 'supplier'). Likewise, the other topics can be labelled: E.g. topic 3 builds around medical topics which makes sense regarding the high number of biotech IPOs in 2020 and 2021 and topic 4 lists many words from the field of construction or buildings.

Thus, COVID definitely had an impact on the topics discussed in the risk sections, since it composes a whole new topic. Next, I will evaluate if there is a change in sentiment and sentiment related words. Figure 2 shows the ratios for the groups of constraining words, uncertain words and both sentiment measures over time as well as the corresponding histograms for these measures for both subperiods. Since the data is interval-scaled, it is not possible to infer about what is a high or low value. Therefore, only differences between the subsamples' distributions can be interpreted with regard to a possible decline in sentiment.

There is no clear time trend observable for any of the variables. However, slight differences between pre- and COVID-period can be seen in the scatterplots, when highlighting parts of the data. The corresponding histograms confirm this impression: a skew to the right is observable for the constraining and uncertainty words ratios, and a skew to the left is observable for the FinBERT sentiment ratio in the COVID-period. Regarding the LMD Sentiment Ratio, changes of the pre-period data distribution to both directions are observable in the COVID-period. This means that constraining and uncertain words are used slightly more in the COVID-period and sentiment declined to a certain degree, when using FinBERT sentiment. When using the LDM sentiment, there are mixed results: the distribution, shows some increase in higher sentiments as well as in lower sentiment scores.

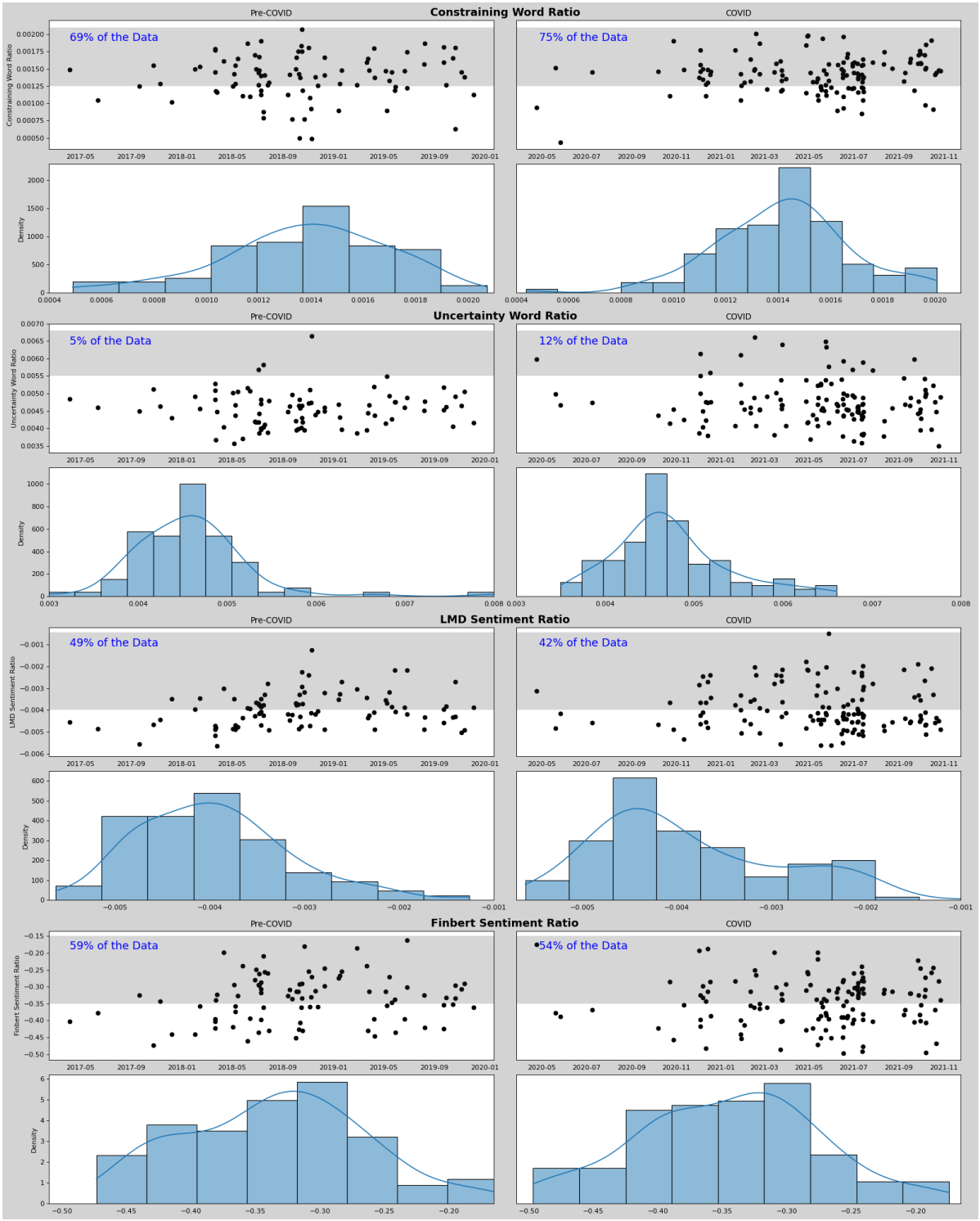

Figure 4: Data distributions

To gain more insights about the pandemic's influence on prospectus language, words related to COVID-19 are counted for all IPOs in the COVID-period and correlations between all variables are evaluated via regression plots. Figure 3 shows the relation between the constraining and uncertainty ratios and the pandemic word ratio. There is evidence, that words associated with the pandemic are substitutes for constraining

words, since constraining words usage decreases with more pandemic words used. For uncertain words by contrast, there is rather the impression of a complementary (or even independent[1]) usage of both word groups during the pandemic.

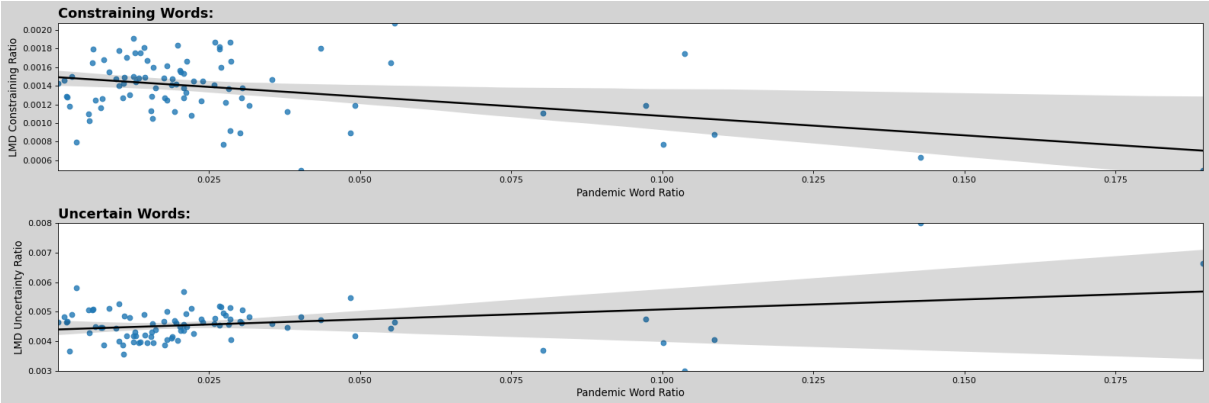

Figure 5: Regression plots regarding pandemic words

In a last step the consistency of sentiment measures will be evaluated. For this reason, I compare the sentiment variables' correlations to all remaining variables. If the sentiment measures are consistent, their correlations to the remaining variables should have the same sign and similar strength. There is however a significant difference between those correlations. The correlation of the FinBERT Sentiment ratio with the pandemic word ratio is slightly negative to neutral (when including confidence bands), while for the LDM Sentiment Ratio it is positive. Additionally, there is a negative correlation between FinBERT sentiment and uncertainty words ratio, while it is positive to neutral for the LMD sentiment. Both measures only show aligned results regarding the correlations to constraining words and are of course positively correlated with each other, since they try to measure the same construct.

[1] When considering confidence intervals.

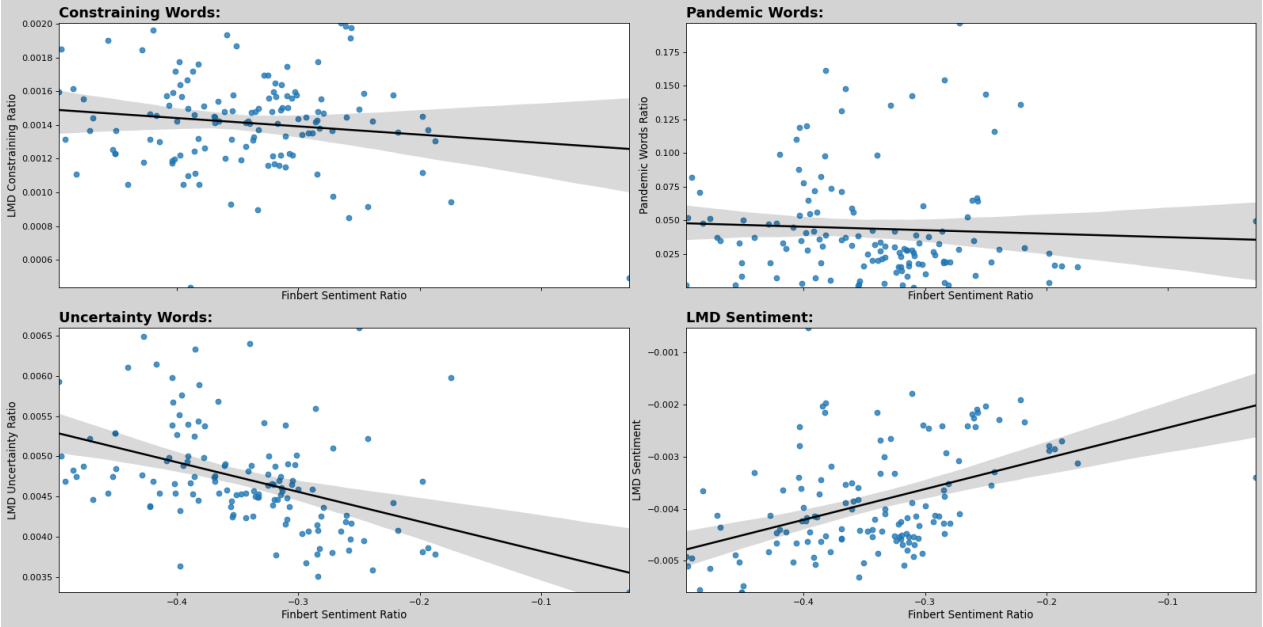

Figure 6: Regression plots regarding FinBERT sentiment

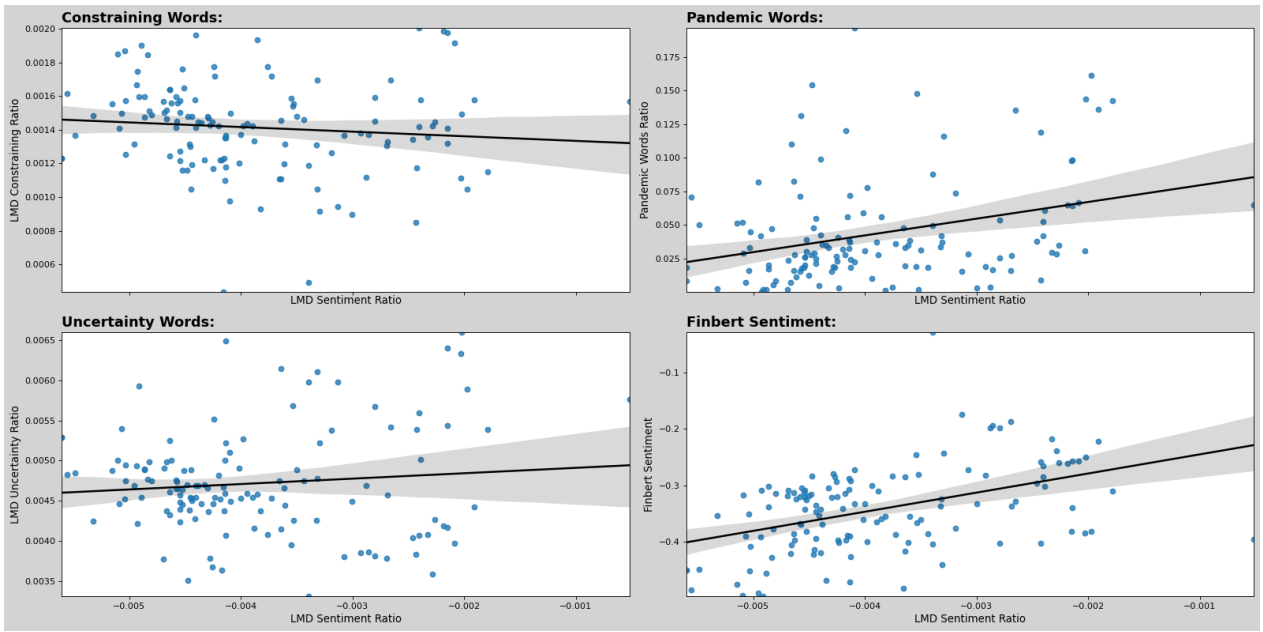

Figure 7: Regression plots regarding LMD sentiment

Interpretation of Results and Concluding Remarks

Regarding the topic models, as expected, COVID created a new important topic around supply chain, finance and stakeholder related terms. However, the topic model of the pre COVID-period does not yield results that are easy to interpret. Thus, it is hard to make a statement about whether and how strong topics discussed in the risk section changed, because the remaining topics of the COVID-period cannot be compared to the pre-period.

Through analyzing COVID related terms I detect hints of a possible substitution between constraining and pandemic words. This is intuitive in the sense, that words like 'lockdown', 'quarantine' or 'travel ban' imply constraints and thus might replace constraining words used in the pre-period. Regarding uncertainty words, I detect a positive to neutral relationship with pandemic words. Uncertainty words seem to be used alongside pandemic words. Many pandemic words like 'outbreak', 'super spreader' or 'panic buying' might need to be elaborated on, in order to explain consequences and thus uncertainty words are used in the further explanation of consequences.

When evaluating the sentiment measures' correlation with the remaining variables, overall, the results with regard to the FinBERT sentiment are more intuitive. More uncertain words are supposed to lead to a worse sentiment and COVID related words are supposed to spread uncertainty which should lead to a lower sentiment score.[2] The inconsistencies between LMD sentiment and FinBERT sentiment might stem from the fact that FinBERT evaluates the sentiment based on whole sentences. One sentence could e.g. contain three positive words and will be evaluated as positive, while another one only contains one negative word and will be evaluated negative. On this way LMD ratio might be distorted when using word counts based on the whole document. An additional advantage of FinBERT is its bidirectional component, saying that it can set

[2] An explanation for a possible neutral effect of pandemic words on sentiment could be, that COVID related terms are a new group of words. They do not affect the usage of other word groups like uncertain words and thus have no impact on the sentiment. This would also address a possible missing effect of pandemic words on uncertain words in Figure 3.

words into context. For those reasons, FinBERT sentiment should be preferred to the LDM dictionary sentiment when examining prospectus risk sections.

Limitation of my work and its findings might be the sample size, which was kept rather small to limit sample construction time and computation durations. However, there are some interesting approaches for further research. By building topic models over a significantly larger sample, the model for the pre-period might get easier to interpret and the unfeasible comparison between topics discussed in both periods might get feasible with interesting outcomes. Additionally, a more thorough examination of how COVID words influence the usage of other words in the risk section of IPO prospectuses, might bring helpful insights for detailed sentiment analyses as well as other methods. Finally, it might be interesting to implement COVID terminology usage into existing explanation approaches for IPO underpricing.

References

Balakrishnan, K., & Barov, E. (2011). *Analysts' use of qualitative earnings information: Evidence from the IPO prospectus' risk factors section.* University of Pennsylvania; New York University. https://www.researchgate.net/publication/228968901_Analysts'_use_of_qualitative_earnings_information_Evidence_from_the_ipo_prospectus'_risk_factors_section

Blei, D. M., Andrew Y. Ng, & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research.*

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems (Vol. 22)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Chen, S., & Wang Yufei. (2016). *Latent Dirichlet Allocation.* University of California San Diego. https://acsweb.ucsd.edu/~yuw176/report/lda.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* http://arxiv.org/pdf/1810.04805v2

Ding, R. (2016). Disclosure of Downside Risk and Investors' Use of Qualitative Information: Evidence from the IPO Prospectus's Risk Factor Section. *International Review of Finance, 16(1)*, 73–126. https://doi.org/10.1111/irfi.12066

Eckbo, B. E. (Ed.). (2009). *Handbooks in Finance: Vol. 1. Handbook of corporate finance: Empirical corporate finance* (Reprinted.). Elsevier North-Holland. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=294165

IATE. (2021). *COVID-19 terminology available in IATE.* https://data.europa.eu/euodp/repository/CDT/OP_Covid19_IATE_11032021.xlsx.gz

Ljungqvist, A. (2009). Ipo Underpricing: A Survey. In B. E. Eckbo (Ed.), *Handbooks in Finance: Vol. 1. Handbook of corporate finance: Empirical corporate finance.* Elsevier North-Holland. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=609422

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. https://doi.org/10.1016/j.jfineco.2013.02.017

Ly, T. H., & Nguyen, K. (2020). Do Words Matter: Predicting IPO Performance from Prospectus Sentiment. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)* (pp. 307–310). IEEE. https://doi.org/10.1109/ICSC.2020.00061

Yang, Y., Uy, M. C. S., & Huang, A. (2020, June 15). FinBERT: A Pretrained Language Model for Financial Communications. https://arxiv.org/pdf/2006.08097

Zhou, K. U., Zhou, B., & Liu, H. (2020). IPO Underpricing and Information Quality of Prospectuses. *The Singapore Economic Review, 65*(06), 1559–1577. https://doi.org/10.1142/S0217590820500289