Prospectus Language and IPO Performance

Jared Sharpe
Institute for Financial Services Analytics
University of Delaware
Newark, Delaware, USA
jaredws@udel.edu

Keith Decker
Department of Computer &
Information Sciences
University of Delaware
Newark, Delaware, USA
decker@udel.edu

Abstract

Pricing a firm's Initial Public Offering (IPO) has historically been very difficult, with high average returns on the first-day of trading. Furthermore, IPO withdrawal, the event in which companies who file to go public ultimately rescind the application before the offering, is an equally challenging prediction problem. This research utilizes word embedding techniques to evaluate existing theories concerning firm sentiment on first-day trading performance and the probability of withdrawal, which has not yet been explored empirically. The results suggest that firms attempting to go public experience a decreased probability of withdrawal with the increased presence of positive, litigious, and uncertain language in their initial prospectus, while the increased presence of strong modular language leads to an increased probability of withdrawal. The results also suggest that frequent or large adjustments in the strong modular language of subsequent filings leads to smaller first-day returns.

1 Introduction

Underpricing, the high average return over a stock's Initial Public Offering (IPO) price on the first-day of trading, is a subject of great financial research (Ritter and Welch, 2002; Huibers, 2020). While the price of the offering is believed to be the best efforts of the underwriter, the individual or firm who assigns the final price, the average first-day return tends to be between 10–14% (Ritter and Welch, 2002). The difference between the first-day closing price and the IPO price, magnified by the number of shares sold, is referred to as 'money left on the table' since it would seem that the firm whose stock is being sold at a premium was undervalued by the underwriter (Ritter and Welch, 2002). A great wealth of literature has attempted explanations using industry, year, momentum, and an assortment of variables and incentive theories (Sherman and Titman, 2000; Lowry and Schwert, 2004; Loughran and Ritter, 2004; Quintana et al., 2005; King and Banderet, 2014; Tao et al., 2018; Seth et al., 2019; Moran and Pandes, 2019), to name a few.

The Efficient Market Hypothesis (EMH) (Fama, 1970) theorizes that current stock prices incorporate all present market information, including the previous price for which a stock traded. IPOs do not have this luxury; market information that would be contained in its price must be discovered through other means, such as within the documents that must be filed with the SEC in order to conduct the IPO: the prospectus filings. Incorporating information from text sources, such as news articles and financial documents, has become exceedingly popular in stock pricing and market valuation, (Hanley and Hoberg, 2012; Loughran and McDonald, 2013; Bajo and Raimondo, 2017; Sehrawat, 2019; Yan et al., 2019; Araci, 2019; Desola et al., 2019; Ly and Nguyen, 2020).

The sentiment construction described below is a strong middle-ground between percentage-of-words in a user defined list techniques common in the finance literature, (Loughran and McDonald, 2013; Hanley and Hoberg, 2012; Loughran and McDonald, 2020), and the effective techniques of word embedding (Araci, 2019; Sehrawat, 2019; Picasso et al., 2019; Peng and Jiang, 2016). In place of counting the number of times sentiment specific words occur in a document, the results utilize the cosine similarity between the embeddings of all words[1] and all words in each of the sentiment word lists of Loughran and McDonald (2011). Four pretrained embeddings are compared with standard percentage-of-words.

Therefore, the contribution of this work is threefold. Firstly, the analysis of initial prospectus sentiment in withdrawal prediction. The second con-

[1] All non-English words, symbols, and common stop words ('a','an','the', etc.) are purged from the document using the python package *ntlk.corpus* and are not included in the analysis.

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 154 - 162

tribution of this work is the expansion of existing sentiment scoring techniques to utilize a stronger, more modern tool: word embeddings. The third contribution of this work is the incorporation of all prospectus amendment document sentiments published before issuing, contributing to the growing literature studying the information revealed during the IPO filing process and how it relates to IPO valuation.

2 Literature Review

2.1 IPO Pricing

Many IPO intrinsic variables have been shown to be correlated with first-day returns; Ritter and Welch (2002) offer a review of IPO pricing factors. A recent investigation into the influence venture capitalist (VC) support has on IPOs shows that such firms are less susceptible to financial distress (Megginson et al., 2016) supporting the screening hypothesis, wherein VCs conduct their own screening analysis so only those firms that will perform well will receive VC backing, contrary to the treatment hypothesis in which VC backed firms do well because of the influence the VC has on the firm.

The litigation risk theory, as extended by Hanley and Hoberg (2010) and Hanley and Hoberg (2012), hypothesizes that underpricing exists to decrease the chance of a lawsuit for misleading investors on the positive quality of a firm (overpricing). In both papers, the authors use the number of root words (i.e if 'will' and 'willing' occurred in the section of text, the root word 'will' would only be counted once), as the total amount of information in each of section of the prospectus filing. Surprisingly, their findings suggest that firms whose filings contain more standard content experience higher first-day returns. Recently, McGuinness (2019) examines how IPO disclosures contained in the Risk Factors and Use of Proceeds sections affect returns and trading volume. Unsurprisingly, firms who deploy more of their proceeds to pay down debt (reduce risk factors) experience less IPO subscriptions and lower first-day and following returns; however, this has little affect on trading volume, compared to those firms that apportion more proceeds to internal investments.

2.2 IPO Withdrawals

Helbing (2019) offer a comprehensive review of the withdrawal literature, calling for more attention from NLP techniques. Busaba et al. (2001) hypothesize that the ability to withdraw from an IPO grants additional power to the issuing firm, since the underwriter can make no profits from underpricing if the offering never happens. Their study focuses on 113 firms that withdraw between 1990 and 1992 identified by the Securities Data Company (SDC) database and employs a probit model for the probability of withdrawal. They find a negative correlation between underpricing and probability of withdrawal, suggesting that underpricing is compensation for *information revelation* rather than *information production*, contradicting theories relating positive information to increased underpricing. Their findings also support the claim that IPO withdrawals are more common in periods of poor market performance. Importantly, the authors suggest that higher uncertainty about the firms value on the part of the underwriter creates a higher chance to receive negative news, increasing the possibility of withdrawal.

Benveniste et al. (2002) theorize that underwriters (investment banks) are responsible for the clustering of IPO timings by industry to overcome the coordination problem of pioneering firms taking the bulk of the cost, through underpricing, in newly developing industries; the authors take the example of Internet based firms following the highly successful Netscape IPO of 1995. Their research includes the 'option-to-abandon' by firms and withdraw their offering if a more favorable option, such as private funding, is available. Their study finds that despite the strong/poor performance of pioneer offerings, follower firms often withdraw/complete their IPO, contrary to original beliefs.

2.3 ML and NLP in Finance

Loughran and McDonald (2020) and Ke et al. (2019) offer recent reviews of natural language processing in finance.

Most famously, Loughran and McDonald (2013), examine the influence of initial prospectus sentiment, final prospectus sentiment, and the time between initial and final prospectus filings on IPO return and post IPO return volatility. Using a percent-of-words approach and regressing the percentage of words within each of a 6-sentiment corpus (Loughran and McDonald, 2011), their results are mixed, leaving more questions about how to quantify and evaluate the sentiment of these crucial documents, though they do find a strong correlation between prospectus uncertainty and high

first-day returns. This follows the theory that underpricing is a reward to the underwriter for their assumed risk. McGuinness (2019) amend Hanley and Hoberg (2012) and suggest that each section of the prospectus may have different sentiment for different reasons, such as having different audiences. Thng (2019) examine the differences in tone between firms with and without VC backing, using only the Management Discussion and Analysis section of the prospectus filing, and they conclude that VC-backed firms tend to be less optimistic. Similarly, González et al. (2019) use the Loughran McDonald approach to investigate the impact of tone in IPO prospectus filings in Latin America and find a significant positive relationship between board size and underpricing and a negative relationship between board independence and underpricing when controlling for uncertain tone.

Araci (2019) compare the performance of publicly available BERT, (Devlin et al., 2018), which is trained on a corpus of Wikipedia articles and books, to the performance of the BERT model trained solely on 10-k's from 1998-1999 and 2017-2019, which they call 'FinBERT,' on the task of sentiment classification for financial documents. There are several other published 'FinBERT' models including Desola et al. (2019), Yang et al. (2020), and (Liu et al., 2020)[2]. Their results show a clear improvement on language comprehension by the models on masked language model accuracy (MLM) and next sentence accuracy (NS) on new 10-Q data. Thus, as expected, domain knowledge and verbiage differ greatly from ordinary language, but for Earnings Calls, the models trained on financial documents are over-training, suggesting another language barrier. Araci (2019) compares the average in-list similarity of each Loughran McDonald sentiment using the publicly available pre-trained BERT embeddings and BERT embeddings trained on a corpus of financial documents; the results indicate that there is a significant difference between the resultant word embeddings, suggesting that the language of financial documents is unique to the field.

Tao et al. (2018) deploy deep learning techniques to extract 'forward looking statements' (FLS) from the final prospectus filing for successful IPOs between 2003 and 2013 to train a custom word2vec embedding. FLS are statements concerning the firms future projects, works-in-progress, and goals. Latent Dirichlet allocation (LDA) (Blei et al., 2003) is used to determine the common topics to which the FLS are addressing across all firms. The FLS features, including their Loughran McDonald sentiment and topics are combined with common IPO features (underwriter rank, industry, etc.) in several ML algorithms (Decision Tree, Bayes Classifier, Neural Network, etc.) and feature importance algorithms for the prediction tasks. With all of the machinery in place, the authors best report a 0.76 area under the curve (AUC) in predicting if the IPO will have a positive first-day return from an ensemble ML model and a 0.68 AUC if the IPO will have a positive up-revision. While this work is extensive and well-documented, the authors only review the final prospectus document and ignore the probability of withdrawal by only focusing on successful IPOs.

Recently, Ly and Nguyen (2020) apply several machine learning algorithms to prospectus sentiment factors as calculated using percent-of-words modeling and Loughran McDonald word lists to predict if the third, fifth, tenth, twentieth, and thirtieth day closing price is higher than the IPO offer price. Despite the expected strength of ML models, the logistic regression model performed the best and above 50% accuracy at all event horizons, without any market controls – the only data uses were text derivatives from the prospects filing such as total number of non-stop words, total characters in the document, and the Loughran McDonald word counts.

3 Data Collection

Following the method used by Lowry et al. (2017) and their published R code, IPO data is first collected from Thomson Financial Securities Data New Issues database (SDC) for firms who issue or withdraw between 2004 and 2020. Using the given SEC File Number, the correct CIK numbers are identified and all prospectus related forms, the initial prospectus (S-1), prospectus amendments (S-1/A, 424A, all 424B[3] variants), are collected using Loughran and McDonald (2013). Only those CIKs who filed an S-1 between 2004 and 2020 and issued or withdrew in that time are considered. Firms that have a non-missing *Withdrawn Date* field from SDC are considered to have withdrawn. While the

[2] Araci (2019) is used in the results as it was easily available at the start of this project Python FinBERT.

[3] Form 424B has variants 424B1-424B8, although Tao et al. (2018) only cite '424B' as the final prospectus.

withdrawn firms seldom publish after the S-1, the initial prospectus and final prospectus are fractions of the final picture for those that issue. The information revealed in the amendments must be taken into consideration when forecasting the final offer price and first-day return as it was likely disclosed strategically, (Hanley and Hoberg, 2012; Dambra et al., 2021); this is especially true considering that the final prospectus is often published **after** the issue date, (Loughran and McDonald, 2013). A total of 2201 unique CIK firms are found with sufficient, non-missing control variables following the method of (Lowry et al., 2017) with a total of 10,683 forms to evaluate, after the removal of common stop words (a, an, the, etc.) and all non-English characters[4].Of these, a remaining 1908 CIK firms have qualifying forms to be processed and analyzed in this model.

Index and first-day returns are collected from the Center for Research in Security Prices (CRSP) database accessed through the Wharton Research Data Services (WRDS). The collection of additional firm identifiers was attempted, but the best results were obtained by uniquely identifying all PERMNOs[5] on their first day of record, taking their CUSIP6[6] matches with the firms and taking the sample with the least missing data following Lowry et al. (2017). Carter and Manaster (1990) continue to publish a ranking on underwriters. While this ranking is standard in the underpricing literature, (Loughran and Ritter, 2004; Hanley and Hoberg, 2012; Loughran and McDonald, 2013), it only supplies a ranking in 1984, 1991, 2000, 2004, 2007, 2009, 2011, and 2015; therefore an underwriter ranked 8 in 2000 is still considered to be rank 8 in 2003, but if their rank is missing in 2004, it will also be missing in 2005. The most recent ranking is from 2015. The up-to-date 7-sentiment Loughran McDonald word lists are downloaded from their website.

[4]As Lowry et al. (2017) mention, there is nevertheless room for some errors of firm identification and form acquisition. Only those forms with more than 16 'clean words' are evaluated, as 16 was the 10th percentile of all documents and the 10.1th percentile was 35. This is to account for errors in form acquisition and noisy data.

[5]All publicly traded stocks are assigned a PERMNO (permanant number) by WRDS that follows them through Merger and Acquisition (M&A) activity, re-branding, corporate restructuring, etc. Some firms may have more than one PERMNO if they have multiple classes of stock traded.

[6]CUSIP is a 9-character identifier issued by CUSIP Global Services and uniquely identifies financial instruments and their issuers; CUSIP6 uses the first 6 characters of the CUSIP, which identify only the issuer.

4 Methodology

For these results, instead of using a percentage of words to represent each sentiment, the cosine similarity between every word in a document and every word in each Loughran McDonald sentiment list is calculated using the publicly available GloVe (Pennington et al., 2014) embeddings trained on Wikipedia, Sehrawat's GloVe embeddings (Sehrawat, 2019) trained on 10-K filings, BERT (Devlin et al., 2018) embeddings trained on BookCorpus and Wikipedia, and the FinBERT model from Araci (2019).

Algorithm 1 Sentiment Score Matrix: T

Inputs:
Embedding Matrix: M
Loughran McDonald Word List: LM
Vocabulary: V
Output:
Sentiment Score Matrix: T
for all $w \in V$ **do**
  $score = zeros(7)$
  if $w \in$ any $LM_{category}$ **then**
    $score_{category} = 1$
  else
    $v^w = M_w$
    $v^{lm} = M_{lm}$
    $score_i = max_i(cossim(v^w, v^{lm}))$
  end if
  $T_w = score$
end for

Algorithm 2 Document Scoring

Inputs:
Score Matrix: T
Document: D
Output:
Document Score: $score$
for all $w \in D$ **do**
  $score+ = T_w$
end for
$score = \frac{1}{||score||_2} score$

All seven Loughran McDonald categories are used (Positive, Negative, Constraining, Litigious, Uncertain, Strong Modal, and Weak Modal), giving every word a sentiment vector of dimension seven. Each word in the Loughran McDonald list receives a score of '1' for its category, and zero in all other categories; words not in an Loughran McDonald

list receive a score equal to its maximum similarity to any word in the Loughran McDonald lists for the category of that word, and zero else[7]. See Algorithm 1 for the construction of the word sentiment score matrix T. The formulas are the same for the standard GloVe, Sehrawat, BERT, and FinBERT embeddings[8]. For an entire document, the similarity vectors of all words are vector-summed into the document's total vector, which is normalized by dividing by the 2-norm to be the document's score vector; see Algorithm 2. This process is very similar to the one used by Araque et al. (2019). All documents are scored for each sentiment category by each embedding model; the percentage of words metric is also calculated for comparison. Additionally, for every document after the initial prospectus of each firm, the difference between its sentiment score over the previous document is recorded; these sentiment differences are then summarized by an expanding average leaving the final prospectus with an average difference in the changing published sentiment. This metric will capture large spikes in new sentiments that had not yet been revealed; a similar metric that takes the average in absolute differences between the documents was tested, but the results were insignificant. The following predictions are made: use the initial prospectus sentiment, present market average return, and underwriter ranking to predict if a firm will withdraw, for those firms that issue, use the sentiment of the final prospectus to predict the first-day return, for those firms that issue,use the sentiment of the final prospectus and the average sentiment update difference to predict the first-day return.

Supplementary materials for reproducibility are available upon request; however, the complete data set will take over a week of machine time due to the inclusion of four embedding models and the number of forms to process. Moreover, WRDS and SDC are proprietary, restricting the ability to publicize the entire data set. As noted in (Ritter and Welch, 2002), the years being evaluated often have measurable effects on the final results and thus a large volume of data is preferable.

[7]A few words appear in more than 1 list; these words are given a 1 in each category they appear.

[8]Both GloVe and Sehrawat embedding vocabularies were missing several words in each Loughran McDonald list but never more than 10%

5 Results

5.1 Probability of Withdrawal

Table 1 shows logistic regression coefficients and p-values of the left column regressors with withdrawal as the dependent variable[9]. While the pseudo R-squared is unimpressive, the p-values show significant relationships between the regressors and withdrawal. All of the sentiment scoring methods agree on the general form of the results, but the Sehrawat model achieves the highest pseudo R-squared with a significant positive relationship between negative, strong modal, and constraining language at a 0.05% level and a significant negative relationship between positive and weak modal language at a 0.05% level. The more positive language a firm includes in its filing, the less likely it will withdraw; this can be read as the firm has good intentions or good prospects within the offer, rather than needing to cover debts. Strong modal language is likely taking the role of commitments to future projects or ventures that firms are but eventually either drop or find cheaper capital.

While the percentage model only finds strong significance for the litigious and strong modal coefficients, the embedding methods capture a significantly positive coefficient on constraining language, suggesting that the embedding methods are better at disentangling the presence of constraining language from litigious. Additionally, the positive coefficient suggests that firms are more likely to withdraw given more obligations, and likely debt, acknowledged in their prospectus, all else equal. The change in significance of the litigious language factor between the percent and the proposed methods is likely due to the overlap between Loughran McDonald word lists and the concentration of legal-language words in the S-1 filing, being it is a registration statement; this confusion is better handled by the embedding methods as seen by the increased significance of negative, constraining, and weak modal language in the Sehrawat construction. Uncertain language in the context of new firms that are conducting their IPO is likely to be closely tied with projects whose possible outcomes upon are still under review, works-in-progress, and the FLS of Tao et al. (2018); thus, firms who have docu-

[9]Year fixed effects, the average market return at the time of S-1 filing, top tier, and log sales were included, but are not displayed for brevity. Additionally, the inclusion of the VC factor resulted in a singular matrix as did the separate inclusion of industry fixed effects.

	Percent Coef.	Percent p	GloVe Coef.	Glove p	Sehrawat Coef.	Sehrawat p	BERT Coef	BERT p	FinBERT Coef	FinBERT p
Positive	-51.572	0.0041	-62.7989	0.0004	-66.6374	0.0002	-49.3533	0.0011	-49.4866	0.001
Negative	7.1987	0.5731	32.8713	0.014	33.8713	0.0126	31.5083	0.0218	31.0429	0.0234
Uncertainty	-9.008	0.694	1.4227	0.9495	19.517	0.3971	10.4519	0.6568	11.1548	0.6349
Litigious	60.167	0	18.4975	0.1427	18.9262	0.1187	20.9076	0.0838	21.1957	0.0804
StrongModal	67.3125	0	89.6144	0	90.6276	0	93.2129	0	93.6087	0
WeakModal	-31.1342	0.2533	-45.7306	0.0843	-65.6305	0.017	-56.0963	0.0454	-56.2674	0.0441
Constraining	-21.3634	0.5155	63.4077	0.0127	51.4216	0.0374	68.447	0.0219	66.3964	0.0289
Pseudo R2		0.0961		0.1065		0.1073		0.1065		0.1067

Table 1: Probit regression coefficients and p-values for predicting the probability of withdrawal at time of S-1 filing.

mented an abundance of future projects are less likely to withdraw. While this insignificant result does not support existing theories that uncertainly should increase the probability of withdrawal, (Helbing, 2019), it opens the door for the potential of a more in-depth analysis as to why.

5.2 Amendment and Final Prospectus Sentiment on Underpricing

Since the initial and final prospectus sentiments are well studied (Hanley and Hoberg, 2010; Loughran and McDonald, 2013; Bajo and Raimondo, 2017; Tao et al., 2018), the tables are available upon request[10]. In Table 2, sentiment factors from the final filing and the average difference as described above bring all models significant coefficients on litigious and uncertainty at a 10% level. However, the Sehrawat, BERT, and FinBERT methods strong modal coefficient to significance at a 10% level and the strong modal average difference to be significant at nearly a 5% level. This result suggests that a sudden increase in the committal language during the filing process decreases first-day returns, all else equal. Given the significant coefficient on strong modal, this decrease is lessened if it is maintained in the final filing. The percentage based method appears to be unable to capture this in-process information spike. As stated previously, the current status of this technology is has a difficult time disentangling strong modal and uncertain language particularly with respect to opportunity, though they have different key words. Inclusion of more methods inspired by Tao et al. (2018), Bajo and Raimondo (2017), and Araque et al. (2019) may provide the answer.

5.3 Amendment and Final Prospectus Sentiment on IPO Price

As before, the analysis of the initial and final filings alone on IPO price is well documented, but the associated tables are available upon request. For all factors, the Sehrawat and FinBERT methods capture as many or more significant factors than their general counterparts. The Sehrawat embeddings capture a significant value for the probability of withdrawal derived from the initial prospectus, unlike the other metrics. Although it is debated whether or not the probability of withdrawal should increase the offer price, to entice the issuer to carry out the offer or be lower to hedge the underwriter against a bad investment, (Helbing, 2019), the results are unable to capture any significant relationship, all else equal.

Table 3 reinforces the conclusions on strong modal language from the underpricing regression; the joint conclusion is that it causes a belief between the underwriter and investors that firm is less valuable or a lower quality investment. All methods significantly suggest that average increases in the litigious language of filings over time increase the offer price over time while the degree to which it is present in the final filing decreases the offer price. This result is likely an escalation effect or a bi-product of the filing process itself, but a greater analysis could reveal more acute reasoning, such as appropriate compliance or inappropriate deviation from the law that causes improved or disproved evaluations. The GloVe method shows significant positive affects from the average difference in uncertainty with strong collaboration, save the Sehrawat model. While uncertain language appeared to have no relationship to underpricing from the market, it has a strong negative relationship to price of the IPO itself; however, if this language changed during the filing process, it was strongly positive. Coupled with the effects of constraining and modal language, this suggests that the uncertainty factor is better capturing growth opportunities rather than pitfalls for those that ultimately issue, being as those that do encounter unexpected hardship during the filing process have the option to withdraw.

Perhaps most interesting of all is that the change

[10]Controls for year, industry, market return, sales, positive earnings per share, number of shares, VC backing, mid-point price, top tier, share overhang, and a constant are employed in both Table 3 and Table 2 but not shown for brevity.

	Percent Coef.	Percent p	GloVe Coef.	Glove p	Sehrawat Coef.	Sehrawat p	BERT Coef	BERT p	FinBERT Coef	FinBERT p
Positive	1.9765	0.3259	1.257	0.5356	1.5567	0.4448	1.2468	0.4715	1.0749	0.5343
Negative	2.0639	0.16	2.6669	0.088	2.463	0.1192	2.65	0.0958	2.7093	0.0878
Uncertainty	-5.0807	0.0684	-5.0429	0.0616	-4.8004	0.0842	-5.1086	0.065	-5.1058	0.0653
Litigious	-1.3484	0.0665	-1.4648	0.0869	-1.4729	0.0851	-1.5364	0.0613	-1.5559	0.0584
StrongModal	3.1631	0.1288	3.6144	0.1284	3.899	0.0989	3.9006	0.0992	3.8934	0.0994
WeakModal	3.8294	0.2569	3.1572	0.3409	2.901	0.4047	3.107	0.3662	3.0628	0.3716
Constraining	-3.6886	0.3373	-4.0297	0.1962	-2.5583	0.3755	-3.1642	0.3745	-3.3668	0.3512
Positive_diff_av	-9.5925	0.3854	-6.0314	0.5721	-8.2718	0.4416	-7.4402	0.411	-6.3503	0.4833
Negative_diff_av	1.0585	0.912	-3.4888	0.7117	-2.7914	0.7717	-3.2401	0.7388	-3.7394	0.6995
Uncertainty_diff_av	16.8429	0.3593	17.1491	0.3427	19.3985	0.2707	18.2553	0.3189	19.0574	0.2971
Litigious_diff_av	5.5487	0.1983	8.5279	0.0771	9.0347	0.0585	8.1941	0.0758	8.4348	0.0679
StrongModal_diff_av	-19.3434	0.1745	-28.911	0.0721	-33.0815	0.0408	-32.9693	0.0429	-32.6582	0.0426
WeakModal_diff_av	-21.7586	0.312	-19.7252	0.3361	-22.2543	0.2909	-22.0189	0.3013	-22.3597	0.2903
Constraining_diff_av	35.2237	0.1005	18.9233	0.2967	9.9023	0.5515	16.8016	0.4076	18.7874	0.3625
Prob. Withdraw	-0.0132	0.8734	-0.0413	0.6314	-0.0443	0.6038	-0.0543	0.5231	-0.056	0.5103
Adj. R2		0.2347		0.2362		0.2363		0.2364		0.2364

Table 2: OLS regression coefficients and p-values for predicting the first-day return at time of last prospectus filing before the issue date and average sentiment difference factors.

	Percent Coef.	Percent p	GloVe Coef.	Glove p	Sehrawat Coef.	Sehrawat p	BERT Coef	BERT p	FinBERT Coef	FinBERT p
Positive	-6.1661	0.8453	5.2554	0.8694	19.521	0.5439	23.72	0.3851	18.9184	0.4878
Negative	-7.6335	0.741	-14.1084	0.5666	-21.1526	0.3965	-14.5755	0.5611	-12.723	0.6109
Uncertainty	-108.394	0.0132	-110.549	0.0091	-101.639	0.0202	-102.34	0.0188	-103.334	0.0177
Litigious	-20.5922	0.0745	-22.2152	0.099	-19.556	0.1472	-19.298	0.1356	-20.3935	0.1152
StrongModal	100.47	0.0022	100.85	0.0072	106.814	0.0042	106.684	0.0043	107.174	0.0041
WeakModal	140.97	0.0078	146.841	0.0048	136.551	0.0128	137.419	0.011	137.896	0.0105
Constraining	-209.425	0.0005	-144.483	0.0033	-99.3955	0.0293	-131.574	0.0192	-142.266	0.0125
Positive_diff_av	-35.6215	0.8375	-131.485	0.4345	-204.96	0.2272	-219.687	0.1239	-196.378	0.1693
Negative_diff_av	-75.4334	0.616	-11.5933	0.9378	29.436	0.8462	-15.7184	0.9182	-22.5583	0.8825
Uncertainty_diff_av	579.447	0.0446	662.025	0.0198	434.901	0.1173	546.091	0.0582	555.273	0.0536
Litigious_diff_av	162.524	0.0166	163.973	0.0309	133.752	0.0759	142.764	0.0497	146.61	0.044
StrongModal_diff_av	-429.83	0.0552	-517.888	0.041	-510.633	0.0455	-517.604	0.0438	-529.138	0.0371
WeakModal_diff_av	-424.648	0.2089	-592.593	0.0659	-368.377	0.2675	-458.019	0.172	-463.063	0.164
Constraining_diff_av	967.254	0.0041	592.201	0.0382	333.329	0.2042	506.593	0.1133	558.24	0.0861
Prob. Withdraw	-1.6947	0.1929	-1.4496	0.285	-1.1776	0.3823	-1.4537	0.2782	-1.4774	0.2703
Adj. R2		0.7338		0.7333		0.7324		0.7331		0.7332

Table 3: OLS regression coefficients and p-values for predicting the IPO price at time of last prospectus filing before the issue date and average sentiment difference factors.

in language is opposite in relationship to price to its level in the final filing. This implies that the the act of revealing this information during the IPO process has a measurable affect on the IPO price beyond the effect it has by being present in the final filing. This is especially true for the uncertain and litigious language that were themselves insignificant before the inclusion of their change factors in the final filing.

6 Contributions and Continuing Future Research

This work contributes to the ever growing literature on NLP in finance by first evaluating prospectus sentiment on the likelihood of withdrawal, second by expanding the sentiment evaluation to the use of word embedding methods, which does significantly better at disentangling uncertainty and constraining language from that of strong and weak modality, and thirdly by incorporating a measurement for the change in sentiment throughout the filing process, rather than just at the beginning and end. The BERT embedding method has additional strength beyond the embeddings themselves, which would imply that training a model on this data directly would likely improve the significance of the BERT factors by better capturing the context of prospectus filings. While the inclusion of a probability of withdrawal factor was statistically insignificant, its insignificance raises more questions. The method presented is able to disentangle the effects of uncertain and litigious language throughout the filing process, but more work needs to be done to better evaluate the factors behind the IPO price and first-day returns. Moreover, the ability of the embedding-based methods to first replicate and second out-perform that of the basic percentage-of-words method is a necessary bridge to advance the existing financial literature to more modern techniques.

Acknowledgements

Jared Sharpe's research is supported by the graduate fellowship from the Institute for Financial Services Analytics at University of Delaware.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Oscar Araque, Ganggao Zhu, and Carlos A. Iglesias. 2019. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346 – 359.

Emanuele Bajo and Carlo Raimondo. 2017. Media sentiment and ipo underpricing. *Journal of Corporate Finance*, 46:139 – 153.

Lawrence M. Benveniste, Walid Y. Busaba, and William J. Wilhelm. 2002. Information externalities and the role of underwriters in primary equity markets. *Journal of Financial Intermediation*, 11(1):61 – 86.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Walid Y. Busaba, Lawrence M. Benveniste, and Re-Jin Guo. 2001. The option to withdraw ipos during the premarket: empirical analysis. *Journal of Financial Economics*, 60(1):73 – 102.

Richard Carter and Steven Manaster. 1990. Initial public offerings and underwriter reputation. *The Journal of Finance*, 45(4):1045–1067.

Michael Dambra, Bryce Schonberger, and Charles E Wasley. 2021. Creating visibility: Voluntary disclosure by private firms pursuing an initial public offering. *Available at SSRN 3213482*.

Vinicio Desola, Kevin Hanna, and Pri Nonis. 2019. Finbert: pre-trained model on sec filings for financial natural language tasks.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Maximiliano González, Diego Téllez, María Andrea Trujillo, et al. 2019. Governance, sentiment analysis, and initial public offering underpricing. *Corporate Governance: An International Review*, 27(3):226–244.

Kathleen Weiss Hanley and Gerard Hoberg. 2010. The information content of ipo prospectuses. *The Review of Financial Studies*, 23(7):2821–2864.

Kathleen Weiss Hanley and Gerard Hoberg. 2012. Litigation risk, strategic disclosure and the underpricing of initial public offerings. *Journal of Financial Economics*, 103(2):235 – 254.

Pia Helbing. 2019. A review on ipo withdrawal. *International Review of Financial Analysis*, 62(C):200–208.

Fred E. Huibers. 2020. Towards an optimal ipo mechanism. *Journal of Risk and Financial Management*, 13(6):115.

Shikun Ke, José Luis Montiel Olea, and James Nesbit. 2019. A robust machine learning algorithm for text analysis. Technical report, Working paper.

Emmet King and Luca Banderet. 2014. Ipo stock performance and the financial crisis. *Econometric Modeling: Capital Markets - Asset Pricing eJournal*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, pages 5–10.

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

Tim Loughran and Bill McDonald. 2013. Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics*, 109(2):307–326.

Tim Loughran and Bill McDonald. 2020. Textual analysis in finance. *Annual Review of Financial Economics*, 12:357–375.

Tim Loughran and Jay Ritter. 2004. Why has ipo underpricing changed over time? *Financial Management*, 33(3):5–37.

Michelle Lowry, Roni Michaely, and Ekaterina Volkova. 2017. Initial public offerings: A synthesis of the literature and directions for future research. *Forthcoming Foundations and Trends in Finance*.

Michelle Lowry and G. Schwert. 2004. Is the ipo pricing process efficient? *Journal of Financial Economics*, 71(1):3–26.

T. H. Ly and K. Nguyen. 2020. Do words matter: Predicting ipo performance from prospectus sentiment. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 307–310.

Paul B. McGuinness. 2019. Risk factor and use of proceeds declarations and their effects on ipo subscription, price 'fixings', liquidity and after-market returns. *The European Journal of Finance*, 25(12):1122–1146.

William Megginson, Antonio Meles, Gabriele Sampagnaro, and Vincenzo Verdoliva. 2016. Financial distress risk in initial public offerings: How much do venture capitalists matter?*. *Journal of Corporate Finance*.

Pablo Moran and J. Ari Pandes. 2019. Elite law firms in the ipo market. *Journal of Banking & Finance*, 107:105612.

Yangtuo Peng and Hui Jiang. 2016. Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 374–379, San Diego, California. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Andrea Picasso, Simone Merello, Yukun Ma, Luca Oneto, and Erik Cambria. 2019. Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, 135:60 – 70.

David Quintana, Cristóbal Arco-Calderón, and Pedro Isasi. 2005. Evolutionary rule-based system for ipo underpricing prediction. pages 983–989.

Jay R Ritter and Ivo Welch. 2002. A review of ipo activity, pricing, and allocations. *The journal of Finance*, 57(4):1795–1828.

Saurabh Sehrawat. 2019. Learning word embeddings from 10-k filings using pytorch. *Available at SSRN 3480902*.

Rama Seth, S. R. Vishwanatha, and Durga Prasad. 2019. Allocation to anchor investors, underpricing, and the after-market performance of ipos. *Financial Management*, 48(1):159–186.

Ann E. Sherman and S. Titman. 2000. Building the ipo order book: Underpricing and participation limits with costly information. *Capital Markets: Market Efficiency*.

Jie Tao, Amit V Deokar, and Ashutosh Deshmukh. 2018. Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. *Journal of Business Analytics*, 1(1):54–70.

Tiffany Thng. 2019. Do vc-backed ipos manage tone? *The European Journal of Finance*, 25(17):1655–1682.

Yumeng Yan, Xiong Xiong, J Ginger Meng, and Gaofeng Zou. 2019. Uncertainty and ipo initial returns: evidence from the tone analysis of china's ipo prospectuses. *Pacific-Basin Finance Journal*, 57:101075.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.