THE UNIVERSITY OF CHICAGO

Why are IPOs Always Underpriced?
An Empirical Study of Behavioral Economics

By

Yanwei Pan

June 2021

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: Philip Waggoner

Preceptor: Shilin Jia

Abstract

IPO underpricing has become one of the most famous market anomalies in the modern financial market. This phenomenon, called the "New Issue Puzzle," stimulated a hot debate among economists. Many of them focused on explaining the cause of this anomaly using the classical economics models. It is relatively novel to explain this anomaly based on behavioral economics theories. This paper took a comprehensive perspective, considered two possible stages that the underpricing derived from: the pricing stage and the trading stage. For each stage, the paper proposed two hypotheses based on the behavioral economics theory and examined the hypotheses with the OLS regression model and the most up-to-date machine learning techniques (KNN, Decision Tree, Random Forests, and Gradient Boosting). The results demonstrate that the underpricing level of IPO is positively correlated to managerial confidence. This furtherly implies that the underpricing of IPOs is partially derived from the behaviors in the pricing stage. Besides, the result of four classification models validates that investor sentiments would contribute to the underpricing. Using company features and investor sentiments extracted from the online discussion could classify whether the IPO will be underpriced or not at a precision level of about 74%.

Keywords: IPO; Underpricing; Behavioral Economics; Prospectus; Investor Sentiments; GBDT

1. Introduction

IPO underpricing ("New issues Puzzle") has become one of the most famous market anomalies since Ibbotson (1975) documented in their research. Ibbotson used the data of new issue companies in the US from 1960 to 1969 and built up a time-series model to test the dependency of new issue premia, which is the first empirical analysis that justifies the IPO underpricing phenomenon in the US capital market. Ibbotson's research shows that the average underpricing rate is 11.4% among the sample data, which triggered researchers to explore the factors that cause this "puzzle."

In the initial public offering process in the US financial market, there are three stages: project establishment, management, and subsequent support. In the first stage, the issuer has to negotiate with an investment bank (underwriter) in the United States and jointly study issues such as the issuance market, issuance, and promotion methods, market conditions, stock value, etc. At this stage, both parties should reach an agreement on market conditions, corporate goals, and potential investment targets. Then in the management stage, the investment bank should take responsibility for the issuance of stocks. It should have the ability to construct a vast sales network so that the stocks can be reached quickly by the investors and at a low cost. The latter stage of support is a crucial link in determining the vitality of stock issuance. The investment bank's role is to stimulate and maintain the enthusiasm of investors through its research and advertisements.

Much work has been done in this exploration. As there are two determinants of IPO underpricing-offer price and the 1st-day closing price, the underpricing can only originate from the relatively low offer price or the relatively high 1st-day closing price. The offer price is determined through the pricing process in the pre-IPO period by the underwriters (usually are the investment banks) and issuers. In contrast, the 1st-day closing price is determined by the investors (including

institutional investors and individual investors). Most of the research related to IPO underpricing focuses on these two perspectives.

1.1. The Incentives of Underwriter and Issuers in the Pricing Process

The most common one focused on the underwriters' and issuer's incentives. The majority of researchers regard low pricing as the cause of IPO underpricing. Logue firstly took this perspective. As he considered the investment bankers would underprice a new issue due to the consideration of minimizing cost, risks and gaining favors from issuers, he built a linear regression model to explore the relationship between new issue performance (underpricing level relative to the market index) and competing issues variable, market ebullience variable, etc., Then, Baron utilized principal-agent theory to explain the IPO underpricing in 1982, which built a more generalized theory model in this perspective. His model indicated that, as the underwriters had more advantages in gathering information and resources in the stock markets and had more experience in issuing shares, they would underprice the new issues to ensure issuance success. This is because whether the issuance is a success depends on the 1st-day performance of the stock. If the 1st day closed price falls below the offer price, the new issues would be marked as “unsuccessful” and then influence the reputation and performance of underwriters. As a result, the underpricing rate would positively be related to the risks of the new issue. In 1986, Beatty and Ritter used a data set including firms going public from 1977 through 1982 to prove two prepositions they proposed – whether there is a positive relationship between the underpricing level of initial public offerings and the investors' uncertainties, as well as whether the investment banks conduct the underpricing due to their consideration of reputation. Johnson and Miller (1988) specifically took Logue's consideration of investment banker's prestige to explore the impact of investment banker prestige on the IPO underpricing. They used two methodologies to measure the prestige level of each

investment bank in their sample with a size of 196, including binary measurement and a four-point ranking scale. The empirical result shows that the underpricing level is positively related to the prestige of investment bankers using the OLS regression. Another famous theory in this perspective is known as the "Winner's Curse," which was proposed by Rock in 1986. He built up an adverse selection model and assumed that there were two kinds of investors in the stock market- one is informed investors who have more information about stocks; The other is the uninformed investors (e.g., individual investors) lack of information. If the stocks were overpriced, the uninformed investors, who played essential roles in the new issue process, wouldn't actively invest in the stocks. This is due to their risk aversions. As long as those uninformed investors don't have enough information that indicates the potential growth of the stocks, they wouldn't take risks to invest. Therefore, to ensure the success of IPO, the underwriters would underprice the new issues. In 1992, Welch established a herd behavior model to explain the underpricing phenomenon in the IPO market. In the aspect of social psychology, individual behavior would be significantly affected by collective behavior. Welch's theory model demonstrates that the potential investors would pay attention to other investors' behavior during the new issue process, which would then affect their investment behaviors. In order to attract those potential investors to get involved and trigger the herd effect, the underwriters would underprice the new issues.

In 1989, Welch proposed using a signaling model to explain the underpricing, one of the most well-known theories in explaining the "New Issue Puzzle." He assumed that there are two kinds of firms in the stock market, including low-quality firms and high-quality firms. As the low-quality firms tended to imitate the high-quality firms in their issuing process, they would invest in imitation expenses. However, the high-quality firms could discover their imitations between offerings. In order to give sufficient signals to let the investors identify the high-quality firms, the

high-quality firms would underprice at the IPO since they had much confidence in getting profits in the future and could wait for the long-term gains. But the low-quality firms aimed at achieving profits as soon as possible. Thus, low-quality firms wouldn't underprice their stocks. Welch also used data to support his assumption. He used 1028 IPO firms in the 1977-1982 period as his sample, comparing the IPO data with the SO data and found that the empirical result was consistent with the implications of the signaling model. In the same year, Allen and Faulhaber (1989) built up a pooling equilibrium model and assumed investment bankers play no active role in the pricing process. Their models explain the incentives of good companies to underprice their issuing. Hanley (1993) also used a signaling model to explore the IPO underpricing phenomenon. She considered the rate of the final offer price to the range of anticipated offer prices that the firms disclosed in the preliminary prospectus as an efficient signal to reveal the quality of the firms, which could affect the initial returns. She used the IPO data of 1,430 firms from January 1983 to September 1987, which were compiled from Investment Dealer's Digest Corporate Database, and built a regression model to examine the relationship. The result shows that the prices only partially adjust to new information. The underwriters would like to increase the underpricing rate to compensate the ones who reveal the actual information. Cliff and Denis (2004) hypothesized that the IPO underpricing was positively related to analyst coverage according to that new-issued companies were willing to pay for the research coverage and thus allocate part of the underpriced profits on them. They conducted an OLS regression on the completed IPO data set between 1993 and 2000, including 1050 firms and data of analyst recommendation, offer price, and the 1st-day closing price. The results were consistent with the hypothesis that the underpricing is the compensation of analyst coverage, which demonstrates the issuers' incentives to underprice the new issues.

1.2. The Result of Investor Sentiments in the Trading Process

Compared with the research considering that the underpricing is generated in the pricing process, some scholars take the investors into account as behavioral finance came into our sight. They supposed that the underpricing derives from the trading process on the first trading day. Jaggia and Thosar (2004) referred to the DHS theory (Daniel et al., 1998) to build up an ordered logit regression model and analyzed the high-tech IPO underpricing. Their sample included all IPOs from January 1, 1998, through October 30, 1999, in the specific sectors. Their result shows that investors' overconfidence and biased self-attribution contributed to the underpricing in high-tech companies. However, they didn't provide much evidence that this model could apply to other sectors and generalize it. In 2003, Ljungqvist, Nanda, and Singh did a simulation, which shows how the sentiment investors' behavior influenced the IPO's first-day return. They modeled a new listing in a "hot" IPO market and two types of investors: sentiment investors and rational investors, which justified the underpricing phenomenon triggered by those sentiment investors.

1.3. Quantitative Explanation of IPO Underpricing

Besides focusing on the possible theoretical explanation of IPO underpricing, some researchers intend to find the quantitative determinants of underpricing degrees of IPO. Tian (2011) used some explicit variables to find the determinants of Chinese extreme IPO returns. She utilized the data of Chinese IPO to examine the factors that are significant in the regression model and found that although the asymmetric information about the quality of firms would cause the IPO underpricing, the effects of financial regulations played an essential role in this phenomenon. Tian, Butler, Keefe, and Kieschnick (2014) researched robust determinants of IPO underpricing in the US IPO market in 2014. They examined the variables (more than 40) from previous studies and found that half of

these variables were significant in the regression model. They gathered all the IPO data and the stock price data from 1981 to 2007. Their results demonstrated that the total share volume of the specific month would affect IPO issuing.

In recent years, many researchers begin to pay more attention to the impact of the IPO prospectus. They regard it as a significant factor that could influence investors' sentiment, which could indirectly cause the underpricing. Loughran and McDonald (2013) studied the relationship between the first day returns of the IPO and the sentiments of the prospectus. They used the word lists they established in their previous research in 2011 to classify each prospectus in his sample (1,887 IPOs in the US with an offer price higher than $5 per share during 1997-2010) into uncertain, weak modal, negative, etc. And the result shows that the IPOs with uncertain text in their prospectus have higher first-day returns, which means as the uncertainty of the text increased, the underpricing is more severe. Their research gives me strong evidence that the prospectus could be a factor that causes the underpricing phenomenon, and there could be a significant correlation between these two aspects. Then, Ly and Nguyen furtherly studied this relationship in 2020. They used different models, including the OLS regression model, random forest, decision tree, naïve Bayes, etc., to examine the impact of prospectus sentiments on IPO performance. They used similar sample data as Loughran and McDonald, both from the Electronic Data Gathering, Analysis, and Retrieval system. However, they used a more comprehensive sample, including all the IPOs in the US from 1975 – 2019. In addition, they considered the sentiments of the prospectus as the independent variables and took the complexity of the text, count of characters, and other factors into account. Their result shows that the models they trained can predict the first-day price with an accuracy of up to 9.6% higher than chance.

1.4. About This Paper

As the cause of the "New Issue Puzzle” is complex, researchers haven't come to an agreement through decades of debates. It is reasonable to conduct a more comprehensive study that takes all possible perspectives into account. This paper takes such a broad perspective, considering both the underwriters/issuers' incentives and investor sentiments' influences, which is different from the previous studies that only focused on individual theory or possible cause. Besides, as the result of underpricing is a complicated interaction between the behaviors (or the phycological behaviors) of underwriters, issuers, and investors, this paper utilizes the theories in behavioral economics as the basic theory to explain the phenomenon. It also uses the novel techniques of machine learning algorithms to build up empirical models, which is novel in studying the causes of IPO underpricing. On the perspective that underpricing originates from the pricing process (by underwriters and issuers), this paper focuses on managerial confidence. Since the issuers and underwriters are the dominators on pricing an IPO, their confidence will significantly impact the final offer prices. The sentiment of IPO prospectus can measure the managerial confidence as it is composed by the issuer and underwriters, and their opinions are embedded in the words. And on the perspective that the underpricing originates from the investor sentiments through the trading process, we utilize the online discussions as the measurement of the investor sentiments and the company features that investors would analyze before they make the investment decisions.

2. Theories

Behavioral economics is a branch of applied economics. It combines behavioral analysis theory with economic operation laws, psychology, and economic sciences to explain the anomalies that violets the basic assumptions in current economic models, like self-interest, complete information, utility maximization, and consistent preferences. It breaks the fundamental assumption of rationality in the traditional economics theory and first introduced the irrational market participants to economics, which improves the explanation ability of economic theories towards the real-world phenomenon, especially the anomalies. Behavioral economics emphasizes 'sentiments' and 'bias,' which is aligned with the possible causes of 'Hot Issue Puzzle.' Based on the basic behavioral economics theories, we could propose the following hypothesis towards the two stages where the underpricing occurs- the pricing process (the underpricing is original from the sentiments/bias of issuers and underwriters) and the trading process (the underpricing is original from the sentiments/bias of investors).

There are four fundamental theories in behavioral economics: prospect theory, regret theory, overreaction theory, and overconfidence theory. The four hypotheses proposed in this paper are derived from these four theories in behavioral economics.

Prospect Theory: Many scholars have studied decision-making under risk and uncertainty conditions, and they have proposed many models. The most commonly accepted rational choice model is the expected utility theory of wealth, developed by Von Neumann and Morgenstern (1953). This theory indicates that individual investors would make investment decisions based on the expected wealth and its probability under uncertain circumstances, which provides a standardized model (solves the problem of how people should act when facing risk choices) and

is more convenient to apply. However, in recent decades, the theory has encountered many difficulties. It cannot explain anomalies in the real world. Experimental data violate several fundamental axioms of its. These problems have also stimulated other attempts to explain risks or failures and individual behavior under certain conditions. Prospect theory is one of the best ones.

Prospect theory believes that people usually consider issues not from the perspective of wealth but from winning or losing. They are more careful about gains and losses. While facing risky decisions, will people choose to avoid them? Or to go forward? There is no absolute and straightforward answer since we need to consider the environment of the decision-maker, the state of the company, etc.

There are two laws in the prospect theory. Firstly, while facing gains, people are often cautious and unwilling to take risks; however, while facing losses, everyone becomes an adventurer. The second law is that people have different sensitivity towards loss and gain, and the pain of loss is far greater than the joy of gain.

Regret Theory: Regret theory holds that an individual would evaluate his expected reaction to future events or situations. Bell (1982) described regret as the emotion produced by comparing the outcome or state of a given event with the condition to be chosen. For example, when selecting between familiar and unfamiliar brands, consumers may regret choosing alien brands to cause poor results rather than regret selecting familiar brands and then generate a poor outcome. Therefore, consumers seldom choose unfamiliar brands.

Regret theory can be applied to the field of investor psychology in the stock market. Regardless of whether investors intend to buy stocks or funds that are falling or rising, buying the securities they prefer will produce an emotional response. Investors may avoid selling stocks whose prices have fallen. This is to avoid regrets about making wrong decisions and the embarrassment of reporting

losses. When the choice fails to achieve the expected result or the result is inferior to other alternatives, the regret of making the wrong decision will appear. Therefore, even if the results of decisions are the same, if a particular decision-making method can reduce regret, this decision-making method is still superior to other decision-making methods.

In essence, investors have a herd mentality. To avoid regrets caused by making wrong decisions, investors may refuse to sell stocks whose prices have fallen. When investors consider that many investors have also suffered losses on the same investment, investors may reduce their emotional reactions or feelings. Therefore, investors find it easy to follow the psychology of the crowd and buy stocks that are hot this week or are chased by everyone, leading to the "herding effect" in the stock market.

Overreaction Theory: The overreaction theory is one of the critical theories in western investment psychology. This theory explains that the market always has overreaction. Due to psychological factors such as emotions and cognition, people will show intensification during the investment process.

Classical economics and financial theories believe that individuals are rational in investment activities. They will make logical analyses before making investment decisions. When the stock price is lower than the intrinsic value of the listed company, investors begin to buy the stock; and when the stock price is higher than the inherent value of the listed company, they begin to sell the stock. The securities market has thus formed an atmosphere of value investment, but this is not the case. In the investment field, there is a situation where prices deviate from their intrinsic value for a relatively long time. The main reason is that the future value of listed companies has many uncertainties. It is the uncertainty that triggers irrational factors in investors' psychology. The

common irrational speculation of investors has formed the phenomenon of skyrocketing and collapse of the market.

Professor Robert Shiller called the rising stock market "an irrational, self-driven, and self-inflating bubble" in March 2000. One month later, the Nasdaq stock index, which represents the so-called new American economy, fell from the highest peak of more than 5,000 points to 3,000 points, and after nearly two years of decline, the lowest fell to more than 1,100 points. The Internet bubble is not uncommon in the investment field. Why do people always make the same mistake? Professor Robert Shiller believes that irrational human factors play a significant role, and historical lessons are not enough to rationalize people. Irrationality is a deep-rooted limitation of human beings. Professor Shiller once found in a study that when the Japanese stock market peaked, only 14% of people believed that the stock market would plummet, but when the stock market plummeted, 32% of investors believed that the stock market would plummet. Investors usually over-analyze the recent experience and derive current trends from it but rarely consider the degree of deviation from the long-term average.

Overconfidence Theory: A large amount of cognitive psychology literature believes that people are overconfident, especially overconfidence in the accuracy of their knowledge. People systematically underestimate certain types of information and overestimate other information. Gervaris, Heaton, and Odean (2002) define overconfidence as a belief that the accuracy of one's knowledge is higher than that of the facts. That is, the weight given to one's information is greater than the actual weight. Research on subjective probability measurement has also found that there is indeed a situation of overestimating the accuracy of one's knowledge.

Overconfident people will overestimate prominent and noticeable information when making decisions, especially overestimating information that is consistent with their existing beliefs.

Besides, they tend to collect information that supports their beliefs while ignoring those that do not support their beliefs. When specific points of view are supported by vivid information, essential cases, and obvious scenarios, people will be more confident and overreact to this information. When certain opinions are supported by relevant, concise, statistical, and basic probability information, people usually underestimate the information and respond insufficiently.

Humans tend to see the law from disorder, especially from a large amount of random economic data, to deduce the so-called law. The inherent regularity produces biases of cognition and judgment. Investors' attribution preferences have also aggravated this cognitive bias. That is, accidental success is attributed to their operating skills, and failed investment operations are attributed to factors beyond the control of the outside world, resulting in the psychological phenomenon of so-called overconfidence. Overconfidence means that people are overconfident in their judgment. Investors tend to think that other people's investment decisions are irrational, while their own choices are rational, operating based on superior information, but this is not the case. Daniel Kadmeman believes that overconfidence comes from investors' wrong estimation of probabilistic events. People have overestimated the possibility of small-probability events and believe that it is always possible. This is also the psychological basis for various gambling behaviors; For the moderately high probability events, it is easy to produce too low estimates; but for the probability events of more than 90%, it is considered that it will happen. This is a major cause of overconfidence. In addition, participating in investment activities will give investors an illusion of contro1, which is also an important reason for overconfidence.

Investors and securities analysts are particularly overconfident in areas where they have specific knowledge. However, the level of self-confidence is not related to successful investment. Fund managers and investors always believe that they can outperform the market, but this is not the case.

Brad Barber and Terrance Odean (2001) have done a lot of research in this area. Men always overestimate themselves in many fields (sports skills, leadership skills, ability to get along with others). From 1991 to mid-1997, they studied the investment behavior of 38,000 investors and used annual trading volume as an indicator of overconfidence. They found that the yearly trading volume of male investors was 20% higher than that of female investors overall. Above all, the investment income is slightly lower than that of female investors. The data shows that overconfident investors frequently trade in the market, and the overall performance is an increase in annual trading volume. However, frequent trading due to overconfidence does not allow investors to obtain higher returns. In another study, they sampled 78,000 investors from 1991 to 1996 and found that the higher the annual trading volume, the lower the actual investment income of investors. In a series of studies, they also found that overconfident investors are more likely to take risks and ignore transaction costs. These are also the two main reasons why its investment income is lower than average.

3. Hypothesis

Based on the two stages that underpricing might originate from, we proposed two hypotheses for each stage. The hypotheses focus on the pricing stage are competing, while the other two hypotheses focus on the trading stage are complimentary.

3.1. Underpricing Occurs in the Pricing Stage

Our first hypothesis draws on the market timing theory proposed by Stein (1996). He analyzed how a rational manager, who represents the interests of shareholders and aims to maximize the company's market value, should act towards irrational investors. He believes that when the stock market is irrational, rational managers should use market opportunities and implement different strategies. When the market overvalues the stock price of a company, managers should issue new stocks to take advantage of the high investment sentiment of investors; on the contrary, when the stock price is undervalued, the company should buy back stocks. A confident manager would overvalue the firm, which could generate disagreement between the manager and investors and make equity issuance less likely. Therefore, a confident manager would wait to issue the stock when the stock prices are high, and the disagreement is low. We could propose the following hypothesis:

H1a. *The level of IPO underpricing is negatively correlated with managerial confidence.*

Another possibility is based on the theory of the signaling model (Welch, 1989; Hanley, 1993). High-quality firms would underprice their IPO to signal their quality to the market to differentiate themselves from low-quality firms. Therefore, the manager with higher confidence intends to

underprice more to convey their optimistic expectations of the company quality. If this is the case, we could predict the following:

H1b. *The level of IPO underpricing is positively correlated with managerial confidence.*

To examine which of the hypothesis above is valid, we built up an OLS regression model to find the relationship between the underpriced rate (1st Day Percentage Change) and the proportions of sentimental words in the prospectus.

3.2. Underpricing Occurs in the Trading Stage

The first hypothesis that could explain the IPO underpricing that occurs in the trading stage is the overconfidence theory. Gervaris, Heaton, and Odean (2002) define overconfidence as a belief that the accuracy of one's knowledge is higher than that of the facts. Humans have a characteristic of representative heuristic. That is, humans tend to infer from the surface characteristics of some data, which produces biases of cognition and judgment. Those biases are common and would then lead to the deviations of price to the inherent value. People believe the company's features would affect the company's value, and they would overvalue the IPO with good quality and undervalue the IPO with worse quality. However, as the underwriters and issuers tended to disclose good data to ensure the success of IPO, the underpricing of IPO became an average phenomenon. Therefore, we could propose the following:

H2a. *Company features could predict whether the IPO would be underpriced or not.*

The other hypothesis draws on the theory of herd behavior. Bernheim (1994) argued that humans' behavior would finally conform because humans don't want to look different from others, and our behavior will have a breakpoint in the crowd. Therefore, the sentiments of online comments would

significantly affect the overall sentiments of individual investors. And the behaviors of investors would align with the online views. If this is the case, we could have the following hypothesis:

H2b. *Investor sentiments could predict whether the IPO would be underpriced or not.*

We used the classification algorithms to examine these two hypotheses. We selected the classification model: KNN, Decision Tree, Random Forests, and Gradient Boosting Decision Tree Model.

4. Data and Sample Descriptions

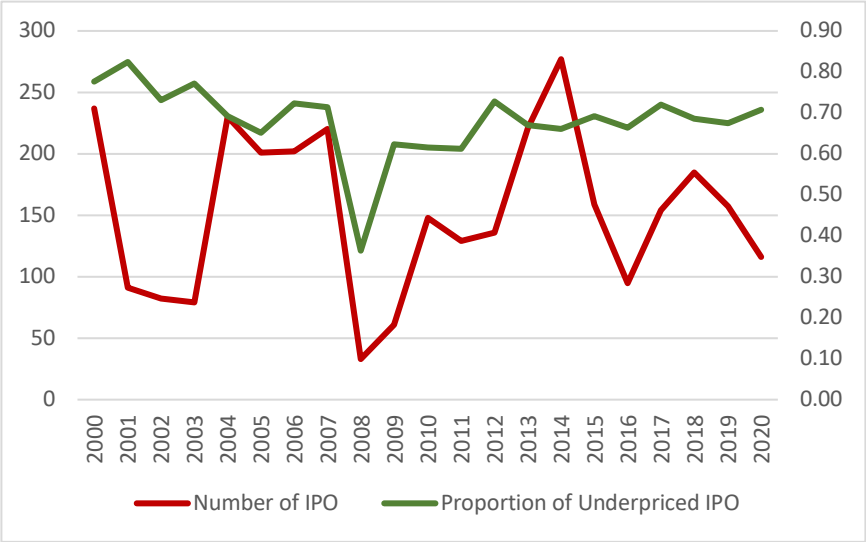
4.1.Completed IPO List

The initial completed IPO list was gathered from IPOScoop.com[1]. Considering the structures of S-1 filings were not stable before 1996, we used the sample IPOs between January 2000 and September 2020, including 3633 IPOs, to construct our datasets. We excluded all of the SPAC (Special Purpose Acquisition Company, which have stationary offer price of $10 and tickers ending with "U") and the IPOs whose offer price is under $5. Our final IPO list contains 3214 IPOs, including their issue dates, issuer names, symbols (tickers), leading underwriters, offer prices, and their first-day closed prices.

Figure 1 shows the change in the number of IPOs and the proportion of underpriced IPOs from 2000 to 2020 (the data in 2020 only includes the IPOs issued before September 11, 2020) in our data set. We can identify a sharp decrease in 2008 when the number of IPOs and the proportion of underpriced IPOs reached a shallow. As in 2008, there was a severe global financial crisis, which stroked the financial markets worldwide, especially the US stock market. As the participants in the market, including issuers, underwriters, and investors, became more conservative in investment during this period, we could see there were fewer IPO underpriced. The overall valuation of IPOs was at a low level. However, the proportion of underpriced IPO is about 60% to 70% in most of the years. Therefore, the underpriced phenomenon is the mainstream.

[1] https://www.iposcoop.com/scoop-track-record-from-2000-to-present/

Figure 1. The No. of IPOs and Proportion of Underpriced IPOs in 2000-2020



4.2.Company Features

The industry/sector data for each company and the Nasdaq monthly trading volume data were scraped down from Yahoo Finance. And other features, including the fiscal data and company information in the pre-IPO period, were collected from the dataset that Prose uploaded on Kaggle.com[2]. The reputation ranking of underwriter was retrieved from Jay R. Ritter's website of IPO data[3]. The ranking contains 1193 investment banks that have issued IPOs in US financial market, ranging from 0 to 9 (-9 means there was no activities). For each IPO, we matched with the reputation ranking of the lead manager in the year of issuing. We dropped the IPO with NaN values and got 1345 IPOs with completed company features.

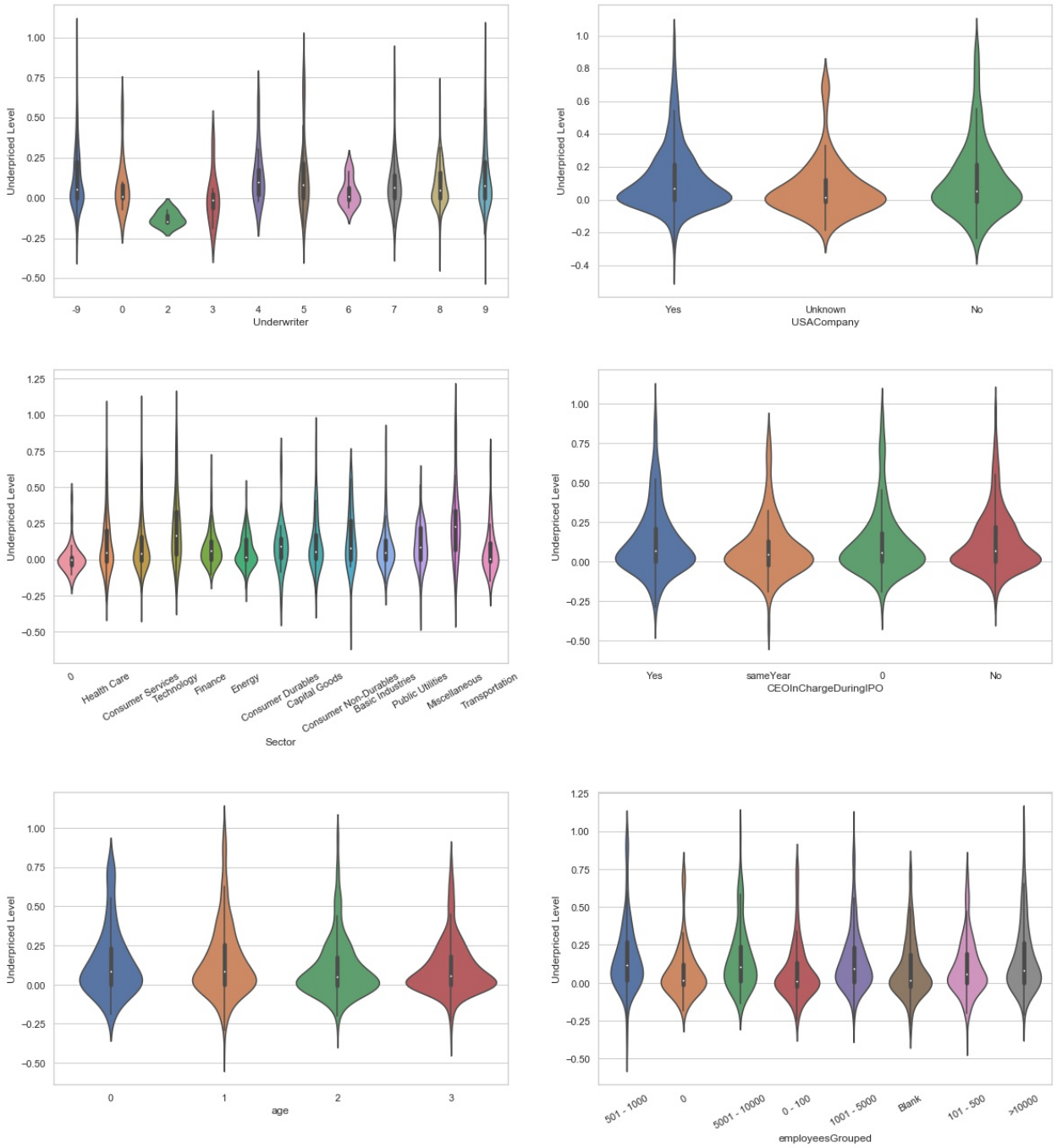
[2] https://www.kaggle.com/proselotis/financial-ipo-data?select=IPODataFull.csv
[3] https://site.warrington.ufl.edu/ritter/ipo-data/

Table 1. Features List in the Data Sample

Variable	Obs	Mean	Std	Min	25%	50%	75%	Max
Underwriter Reputation	1345	2.10	8.24	-9.00	-9.00	8.00	9.00	9.00
Offer Price	1345	15.90	6.49	5.00	12.00	15.00	19.00	91.00
Opening Price	1345	18.22	9.09	4.64	12.41	16.79	21.80	100.01
1st Day Closing Price	1345	18.51	9.70	4.28	12.53	16.75	22.00	122.54
1st Day Percentage Change	1345	0.15	0.29	-0.41	0.00	0.07	0.22	3.54
Avg SPY Volume in Prv 30 Days	1345	396.70	52.03	291.23	358.18	392.69	426.88	572.03
Last Sale	1345	33.48	51.39	0.17	8.51	18.87	39.75	1037.98
Market Cap	1345	5.27E+09	2.97E+10	0.00E+00	3.00E+08	1.02E+09	2.96E+09	7.22E+11
CEO Age	1345	48.38	17.56	0.00	46.00	53.00	58.00	83.00
No. of Employees	1345	5265.09	24118.89	0.00	43.00	495.00	2600.00	446225.00
Year Difference btw Founding and IPO	1345	12.85	22.47	-14.00	2.00	7.00	13.00	186.00
Sector Dummy	1345	6.33	2.96	0.00	4.00	7.00	9.00	12.00
Industry Dummy	1345	6.33	2.96	0.00	4.00	7.00	9.00	12.00
Whether CEO in Charge Dummy	1345	0.91	1.07	0.00	0.00	1.00	1.00	3.00
Whether is US Firm Dummy	1345	1.60	0.75	0.00	2.00	2.00	2.00	2.00

Figure 2 shows the different distribution between the categorical variables of IPO underpriced level. The IPOs issued by the underwriter with a higher reputation are more likely to be significantly underpriced than with a lower reputation since such distributions have lower kurtosis. Besides, the majority of companies would choose the underwriter with a high reputation in their IPO progress. Non-US companies, companies in the technology sector, and companies with CEO in charge during the IPO process are more likely to be significantly underpriced. The companies whose CEO's age is under 50 and the number of employees under 500 are more likely to be greatly underpriced. The figure shows that the underpriced level of IPOs is associated with those categorical variables.

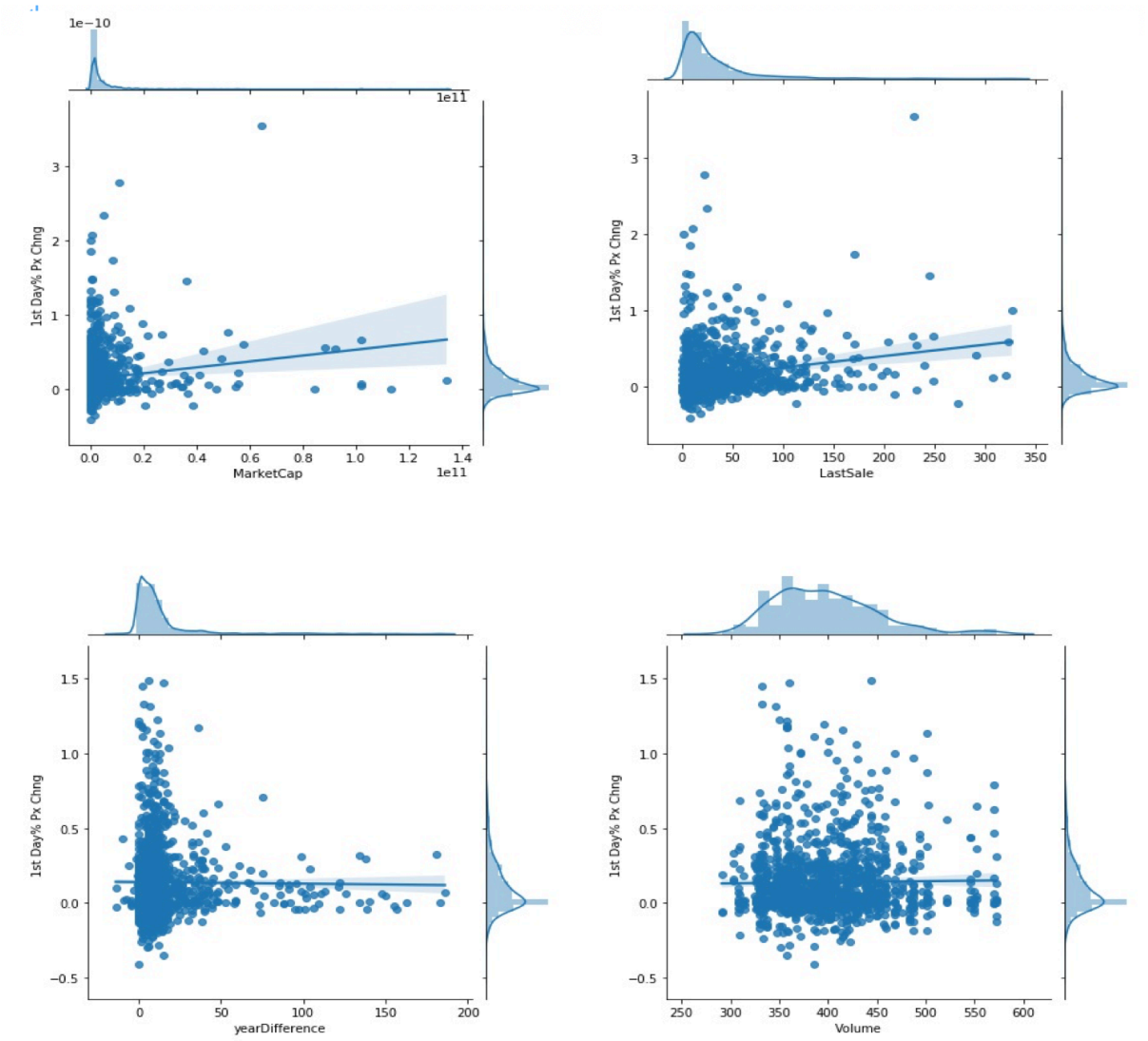
Figure 2. The Distribution of IPO Underpriced Level in Categorical Features



* The above plots excluded the outliers (underpriced level > 1.0).

** The half of the shape in each category is the distribution.

Figure 3. The Distribution of IPO Underpriced Level VS Continuous Variables



* The lines in the plot are the linear regression model fits, and the shadow areas are the confidence interval at 95% confidence level.

** Except for volume, the other three continuous variables showed above are highly concentrated. The data is sparse as the value increase. As the plots aim to demonstrate the possible relationship between the continuous variables and the underpriced level, we used linear regression fits instead of nonparametric LOESS smoothers.

Figure 3 shows the distributions of IPO underpriced level versus the continuous variables, including market cap, last sale, year difference between the founding year and the IPO year, and the average SPY trading volume in the 30 days before issuing. We can see positive correlations

between the underpriced level and market cap and the company's last sale. Since these two variables represent the companies' quality, the companies with greater market cap and sales could have a higher probability of making profits in the future. In addition, we can see a slightly downward fitted line between the underpriced level and the year difference. It demonstrates that the more extended period between the company founded and goes public, the less probability that the issue be underpriced. The average SPY volume represents the activity in the pre-IPO market, and there may be a positive correlation between the underpriced level and the volume.

4.3. IPO Prospectus

We used S-1 filings as the IPO prospectus. We didn't use the final prospectus files like Form 424B1 to 424B8 since they are not the required files for IPO. Only a small proportion of companies that goes public file Form 424. For example, in the sample of Loughran's (2012) research, there were only about 24% filed Form 424. Tim also proved that the changes in expressions between S-1 filings and Form 424 are unrelated to the offer price revisions. So here, it is appropriate to use S-1 filings to measure the managerial confidence), the texts of online discussion about the IPOs.

The text data of IPO prospectus (S-1 filings) were retrieved from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) based on the Python script developed by Professor Tim Loughran and Bill McDonald[4]. We collected 34386 S-1 filings from EDGAR. Then, after removing all the HTML, XBRL segments, and stop words using the stop words list (Loughran and McDonald, 2011), we matched each text with the tickers in our IPO lists and then got 1935 pieces of text. To measure the sentiments in the prospectus, we used the frequencies of sentiment words. The classifications of words are derived from the Loughran and McDonald (2018) word lists,

[4] https://sraf.nd.edu/textual-analysis/resources

including the wordlist of uncertain, positive, negative, weak modal, and strong modal. Since the word lists were summarized from the financial documents (financial news, articles, etc.), it is applicable in measuring the sentiments in S-1 filings.

Table 2. Statistical Summary of Prospectus Data

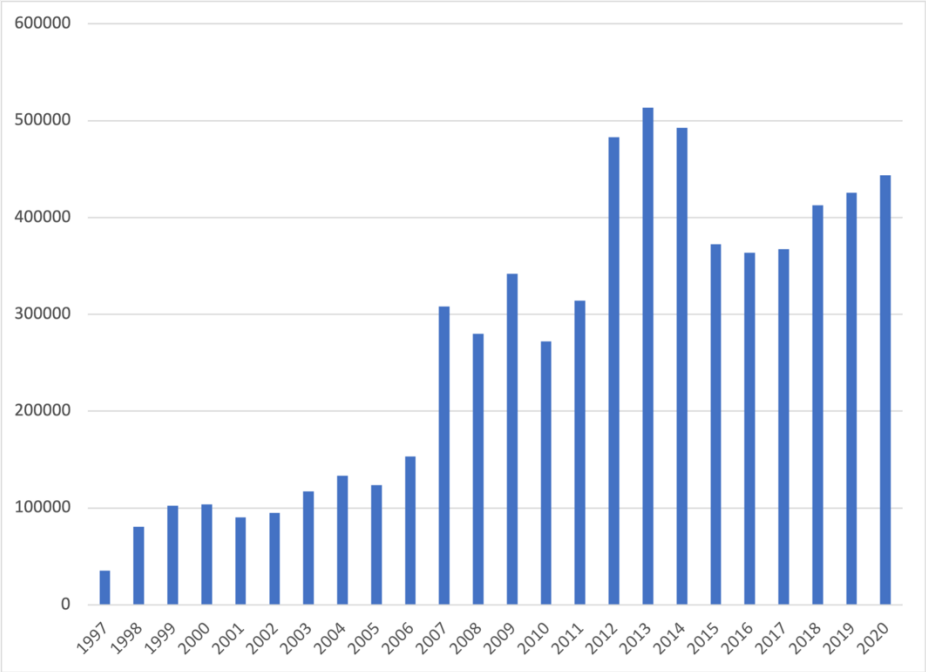
Variable	Obs	Mean	Std	Min	25%	50%	75%	Max
file size,	1935	7.02E+06	7.86E+06	2.14E+05	1.95E+06	5.08E+06	9.36E+06	9.95E+07
number of words,	1935	4.91E+05	4.39E+05	1.96E+04	1.66E+05	4.00E+05	6.60E+05	4.42E+06
% positive,	1935	0.41	0.28	0.06	0.25	0.33	0.43	1.85
% negative,	1935	1.18	0.58	0.14	0.72	1.07	1.60	6.09
% uncertainty,	1935	0.58	0.30	0.12	0.40	0.51	0.67	2.64
% litigious,	1935	1.05	0.66	0.13	0.50	0.93	1.49	3.77
% modal-weak,	1935	0.30	0.14	0.04	0.20	0.28	0.37	0.99
% modal-moderate,	1935	0.11	0.05	0.02	0.08	0.10	0.13	0.36
% modal-strong,	1935	0.28	0.12	0.03	0.20	0.28	0.34	0.96
% constraining,	1935	0.37	0.17	0.06	0.24	0.35	0.46	1.21
# of alphabetic,	1935	4.12E+06	4.30E+06	1.46E+05	1.20E+06	3.05E+06	5.45E+06	4.90E+07
# of digits,	1935	5.45E+05	8.46E+05	3.17E+03	8.88E+04	2.93E+05	7.18E+05	1.35E+07
# of numbers,	1935	1.21E+05	1.75E+05	1.04E+03	2.85E+04	7.43E+04	1.59E+05	2.89E+06
avg # of syllables per word,	1935	1.60	0.08	1.35	1.54	1.60	1.66	1.90
average word length,	1935	5.04	0.14	4.05	4.94	5.03	5.12	5.63
vocabulary	1935	5010.49	1651.47	1281.00	3869.00	4965.00	6265.00	11737.00
1st Day% Px Chng	1935	0.18	0.31	-0.41	0.00	0.08	0.29	2.78

The Sentiments in S-1 filings would reflect the managerial confidence since the managers (underwriters and issuers) would convey their attitudes towards their firms through S-1 filings while composing. They would share their confidence through the texts as well. The more positive, modal-strong of the sentiments of S-1 filings, the more confident the managers are towards their

firms; on the contrary, the more negative, uncertain, and modal-weak of the sentiments, the less confident they are.

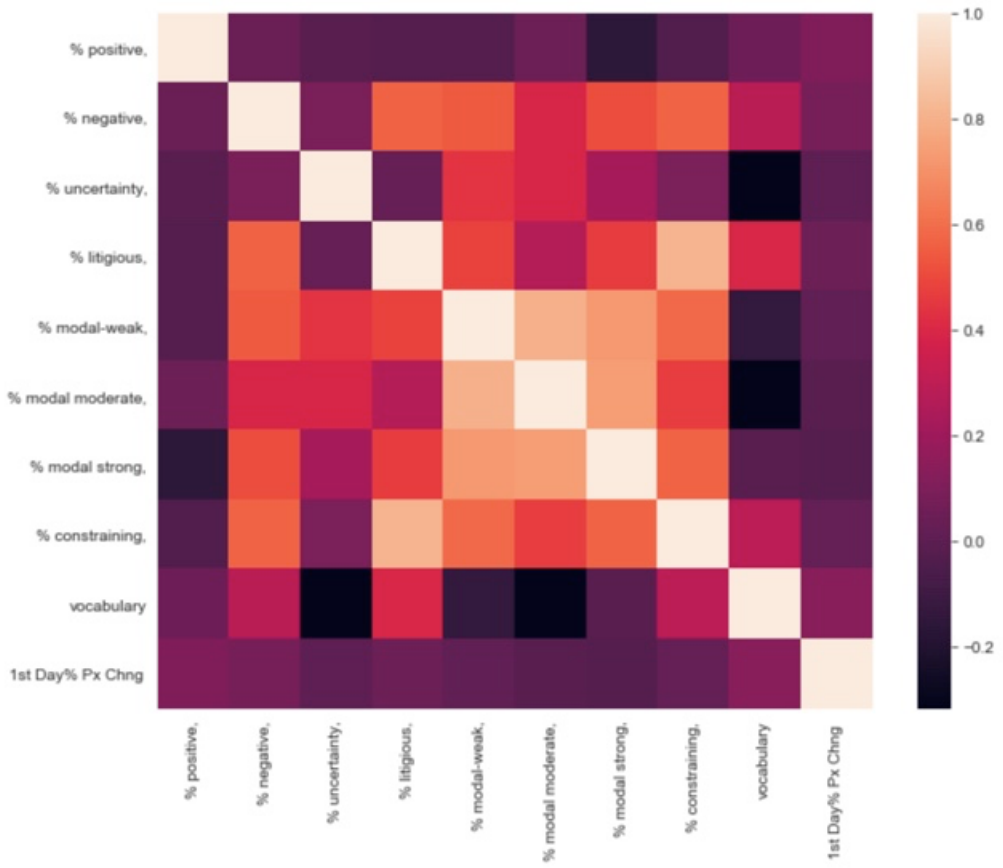
Figure 4 shows the average number of words in S-1 filings through 1997-2020. There is an explicit growth in the number of words throughout the years.

Figure 4. The Average No. of Words in S-1 Filings



The heatmap (Figure 5) demonstrates that the underpricing level of IPO is correlated with the percentage of sentiment words. Among all the types, the percentages of positive and litigious words are highly correlated with underpricing level.

Figure 5. Heatmap of Correlations in Prospectus Variables



4.4. Online Discussions

Internet message board enables investors to interact each other, discuss their opinions and gather information about other investors' opinions and behaviors (Tsukioka et al., 2018). Therefore, the internet message board is an ideal place to analyze the investor sentiments towards an IPO. Among all the alternatives, like Twitter, Reddit, the conversation section on Yahoo! Finance is much preferable since each listing has its threads. However, unlike Yahoo! Japan Finance (YJF) message boards with specific boards for new listings, there are no separate sections on US Yahoo! Finance.

However, we could scrape the message posted in the pre-IPO period based on comparing the posting date and the issuing date.

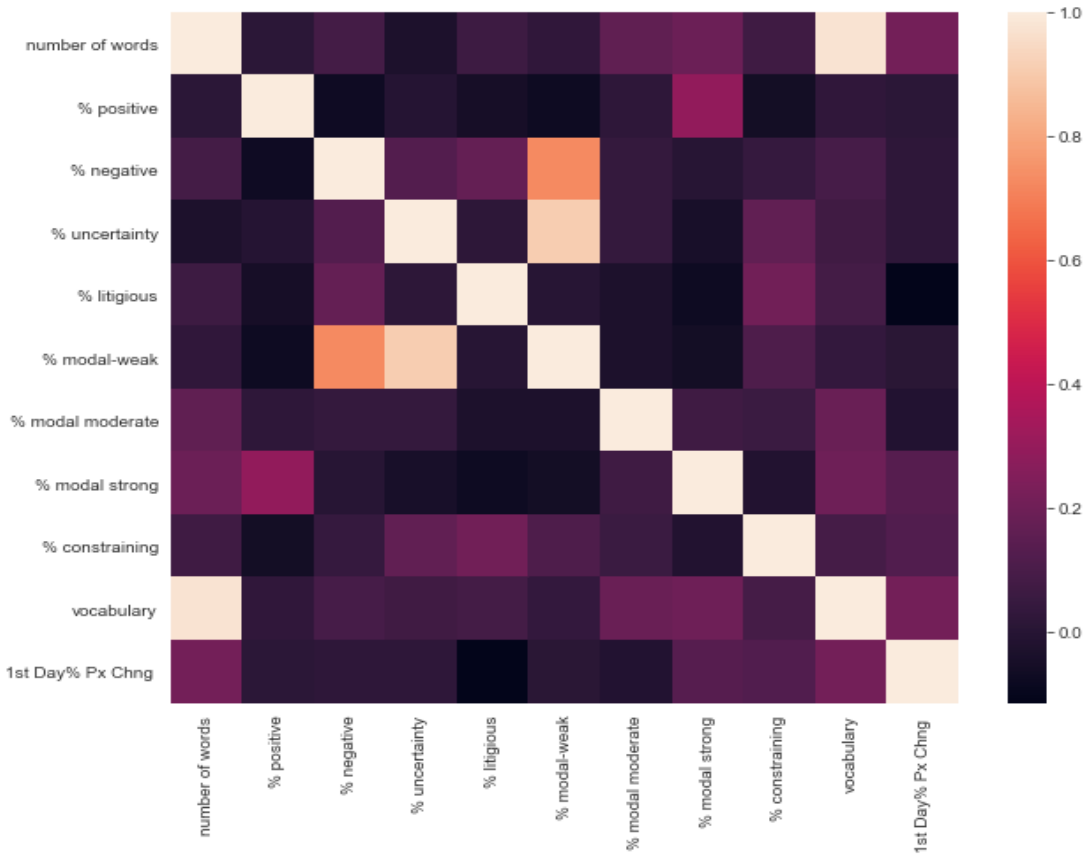
As the time stamp of the posts on Yahoo! Finance is ambiguous (the time of posts are showed as the period from the current time, like three minutes ago, last month, two years ago, etc.), it is hard to identify whether the comment was posted before the 1st trading date of the IPO for the listings that went public several years ago. We chose to focus on the IPO from Jan 2018 to Sep 2020. We collected 17445 pieces of comments associated with 174 IPOs in our list and grouped the comments by the tickers to construct the corpus of investors' opinions towards each IPO. The basic statistics of the corpus data are represented in Table 3.

Table 3. Statistical Summary of Investors Comment Data

	Obs	Mean	Std	Min	25%	50%	75%	Max
file size	174	10703.13	22147.42	19.00	452.75	1167.00	5316.25	99761.00
number of words	174	1587.08	3317.83	3.00	68.00	173.00	708.50	15226.00
% positive	174	1.42	2.24	0.00	0.00	1.25	1.83	25.00
% negative	174	1.25	2.00	0.00	0.00	1.08	1.90	22.22
% uncertainty	174	0.74	1.06	0.00	0.00	0.62	1.03	11.11
% litigious	174	0.12	0.45	0.00	0.00	0.00	0.07	5.26
% modal-weak	174	0.43	0.94	0.00	0.00	0.29	0.55	11.11
% modal moderate	174	0.63	0.73	0.00	0.00	0.57	0.97	3.85
% modal strong	174	0.95	1.09	0.00	0.00	0.85	1.47	8.33
% constraining	174	0.06	0.22	0.00	0.00	0.00	0.03	2.26
# of alphabetic	174	7882.78	16506.70	10.00	306.50	810.00	4091.00	76698.00
# of digits	174	411.13	2282.71	0.00	23.25	71.50	178.00	29797.00
# of numbers	174	117.89	478.05	0.00	8.00	20.50	56.75	6043.00
avg # of syllables per word	174	1.36	0.17	1.00	1.28	1.37	1.45	1.93
average word length	174	4.24	0.51	2.00	3.96	4.30	4.51	5.76
vocabulary	174	371.55	590.94	3.00	40.25	103.00	315.75	2636.00
1st Day% Px Chng	174	0.30	0.46	-0.37	0.00	0.21	0.49	2.49

Figure 6 demonstrates the correlations between the variables in investor comments. The underpriced level is strongly correlated with the number of words in comments, % of modal-strong, % of constraining words in the comments and vocabulary.

Figure 6. Heatmap of Correlations in Investors Comment Variables



5. Models

5.1.Model for H1a and H2b- Regression Model

OLS regression model is a good model to find the link between the variables. As H1a and H1b indicate the opposite relationship between the underpricing level and the managerial confidence (which is characterized by the positive word use in the text of prospectus), OLS regression could answer. The dependent variable is the 1st-day percentage change of price, which is the measurement of underpriced level and could be defined as:

$$\Delta \mathrm{P} = \frac{P_{1^{\mathrm{st}}\ \mathrm{day\ closed}} - P_{offer}}{P_{offer}} \times 100\%$$

The independent variables are the sentiments (% of sentiment words), including the number of words in S-1 filings, the percentage of positive, negative, uncertain, weak-modal, strong-modal, litigious words in the text, the average number of syllables per word, average word length and vocabulary. The control variables are the company features, including the underwriters' reputation, last sale of the firm, and their market cap.

Table 4. VIF Test Result

Variables	VIF
number of words	4.123840467
% positive	1.10182252
% negative	1.86667333
% uncertainty	1.585524289
% litigious	1.891028709
# of alphabetic	3.273488567
vocabulary	2.165728593

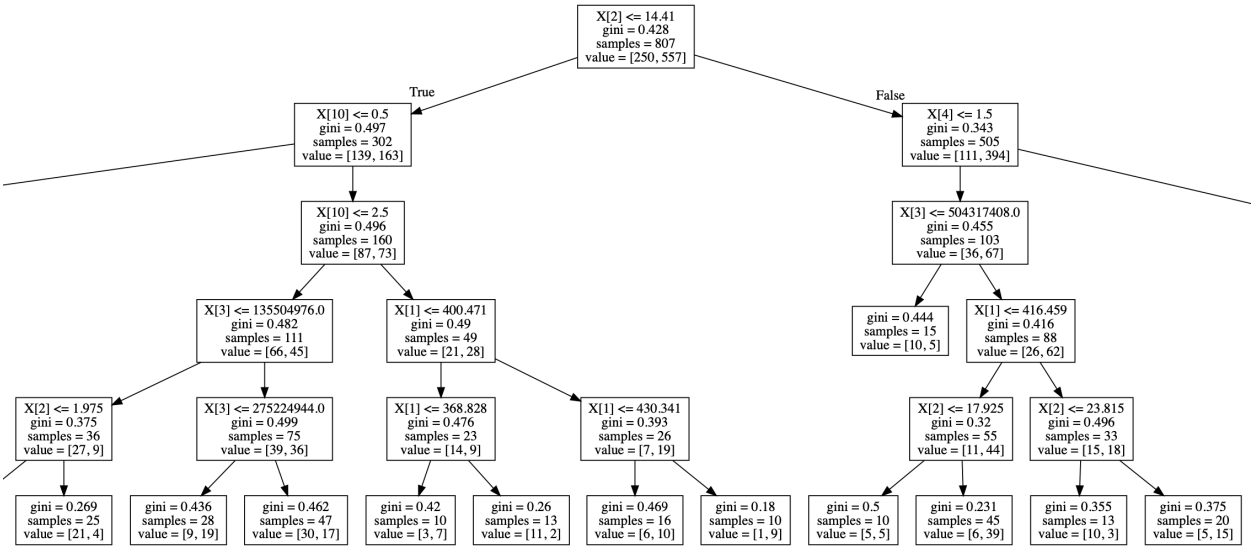
The heatmap of correlations shown in Figure 5 demonstrates that there might be multicollinearity in the independent variables. Therefore, we firstly conducted a VIF test on those variables. The result is showed in Table 4

The VIF of all independent variables is under the threshold of 5. Thus, we could build a linear regression model on the variables.

5.2.Model for H2a and H2b- Classification Models

Since H2a and H2b are hypotheses related to the classification issues (which variables could predict if the IPO is underpriced), we choose four classification models to make the prediction and comparison.

Figure 7. Decision Tree



* To ensure the readability of the model, here only shows the part of the Decision Tree graph.

The Decision Tree is a good baseline model for classification problems. It is relatively simple, using entropy to decide the classification parameters and automatically adjust the data. In order to tune the hyperparameters, I did a random search for every model and used the optimal parameters to build the model and fit the data. Using the company features, I trained the following decision tree model (Figure 7):

K-nearest Neighbors (KNN) classifier is a nonparametric model based on the value of the parameter K that predicts the category of a given observation. It is often referred to as an unsupervised learning method because it, in some sense, requires no model. The KNN method simply uses the data in the neighborhood of each data point to predict type.

As a bagging approach, the random forest contains multiple trees, each of which uses a random subset of features to build a tree. It allows us to explore the complete set of possible predictors better. For the sample with a large scale of features, random forests could have better prediction results. I also did a random search for every random forest model to tune the hyperparameters.

Except for bagging models, boosting models are important ensemble learning models as well. We chose the Gradient Boosting Decision Tree (GBDT) model to fit our data. It uses the weak classifier (like decision tree) to iteratively train the data and obtain the optimal model, which has a good training effect and significantly avoids overfitting. Besides, considering the number of classes in the sectors/industries in the company features, using a one-hot label approach could generate the sparse matrix. Light GBM optimizes the support for category features, and we can directly input categorical features to the model, which is suitable for our data.

Since about 70% of new listings would be underpriced in US stock markets, the data we obtained is unbalanced. The number of underpriced IPOs greatly exceeds the number of non-underpriced IPO, which would lead to a biased prediction result. In order to balance the data (underpriced and non-underpriced), we used SMOTE algorithm to process the data. In 2002, Chawla proposed the SMOTE algorithm, a synthesis of oversampling techniques. It improved the scheme based on the random oversampling algorithm. This technology becomes a common method for processing unbalanced data nowadays.

The basic idea of SMOTE algorithm is to analyze and simulate the minority of samples and add new simulated samples to the data set. Therefore, the categories in the original data are no longer seriously unbalanced. The simulation process of the algorithm uses KNN, and the steps to simulate and generate new samples are as follows:

i) Finding the K nearest neighbors of each minority sample using KNN;

ii) Randomly select N samples from K nearest neighbors for random linear interpolation;

iii) Construct new minority samples;

iv) Synthesize the new sample with the original data to generate a new training set;

6. Results

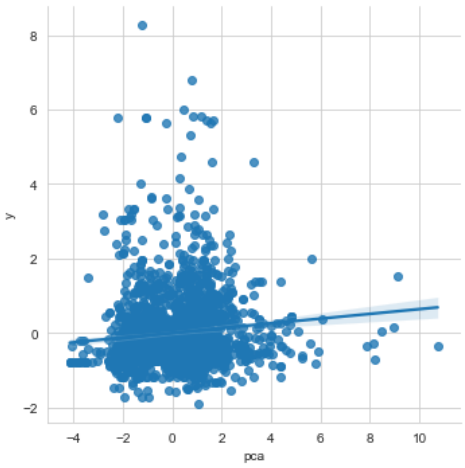
6.1. Relationship Between Managerial Confidence and Underpricing Level

After using the 1st-day percentage change of price as the dependent variable and the sentiment measures extracted from the prospectus text as the independent variables to fit a regression model, we could find the coefficients of sentiments modal-weak, modal-moderate, modal-strong, and constraining, are insignificant. Therefore, we excluded those sentiment variables and got the following empirical result:

Table 5. OLS Result of 1st Day Percentage Change of Price ~ Managerial Confidence

	coef	std err	t	P>|t|	[0.025	0.975]
number of words	-0.1990	0.073	-2.740	0.006	-0.341	-0.057
% positive	0.1080	0.028	2.818	0.000	0.053	0.164
% negative	0.0076	0.036	0.213	0.831	-0.062	0.077
% uncertainty	0.0379	0.030	1.272	0.204	-0.021	0.096
% litigious	-0.0235	0.036	-0.661	0.509	-0.093	0.046
number of alphabetic	0.1292	0.069	1.871	0.062	-0.006	0.265
vocabulary	0.2077	0.046	4.509	0.000	0.117	0.298

Figure 8. 1st Day Percentage Change of Price ~ Managerial Confidence (Primary Component)

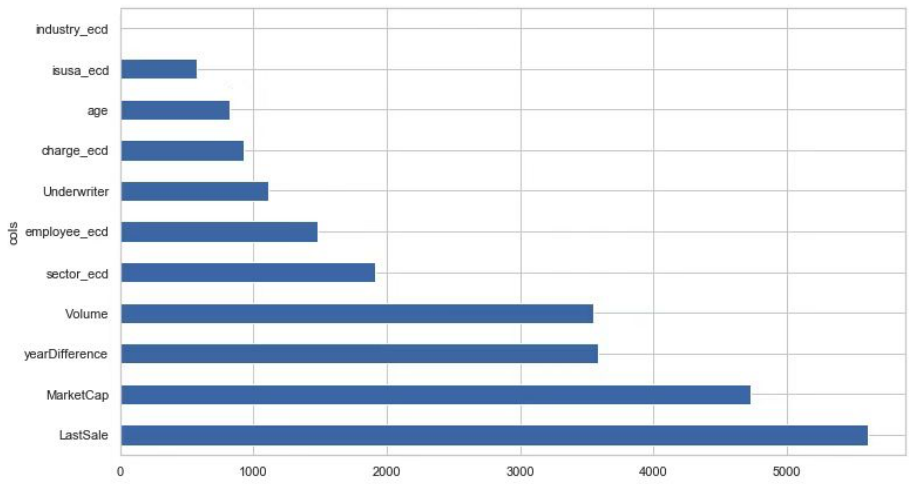


We used PCA (Principal Component Analysis) to reduce the dimension of independent variables and plot a scatter plot (Figure 8). As we could see an upward fitting line between the underpriced and primary components, we could conclude that the underpriced level is positively correlated with managerial confidence.

6.2. Classification Evaluation

Figure 9. Feature Importance of Company Features in Light GBM

	cols	feat_imp
2	LastSale	5609
3	MarketCap	4724
4	yearDifference	3586
1	Volume	3544
5	sector_ecd	1909
7	employee_ecd	1483
0	Underwriter	1113
10	charge_ecd	925
8	age	820
9	isusa_ecd	573
6	industry_ecd	0



We used KNN, Decision Tree, Random Forest, and GBDT classifiers to fit company feature data and investor sentiment data (after transforming by SMOTE). In training the Light GBM classifier on the company features, we found that the feature importance of the industry is 0 (Figure 9), so we considered it a useless variable and excluded it in fitting the Light GBM classifier based on the company features.

Table 6. Classification Result

		Company Features					
		1	2	3	4	5	Average
KNN	p	0.66	0.58	0.72	0.71	0.67	0.67
	r	0.58	0.60	0.62	0.56	0.62	0.60
	f	0.62	0.59	0.67	0.63	0.65	0.63
Decision Tree	p	0.54	0.51	0.71	0.68	0.66	0.62
	r	0.72	0.63	0.68	0.59	0.49	0.62
	f	0.62	0.56	0.69	0.63	0.56	0.61
Random Forests	p	0.56	0.51	0.70	0.64	0.72	0.63
	r	0.72	0.59	0.60	0.62	0.62	0.63
	f	0.52	0.52	0.61	0.68	0.69	0.60
GBDT	p	0.54	0.45	0.82	0.85	0.87	0.71
	r	0.78	0.60	0.65	0.60	0.55	0.63
	f	0.64	0.51	0.73	0.71	0.67	0.65

		Investor Sentiments (Measured by Discussion Texts)					
		1	2	3	4	5	Average
KNN	p	0.57	0.59	0.68	0.52	0.81	0.64
	r	0.64	0.54	0.54	0.54	0.54	0.56
	f	0.60	0.57	0.60	0.53	0.65	0.59
Decision Tree	p	0.58	0.56	0.68	0.55	0.77	0.63
	r	0.36	0.63	0.79	0.67	0.42	0.57
	f	0.42	0.59	0.65	0.60	0.54	0.56
Random Forests	p	0.58	0.55	0.70	0.74	0.71	0.65
	r	0.60	0.67	0.58	0.67	0.63	0.63
	f	0.58	0.68	0.61	0.55	0.51	0.59
GBDT	p	0.61	0.62	0.68	0.94	0.94	0.76
	r	0.68	0.67	0.54	0.71	0.63	0.64
	f	0.64	0.64	0.60	0.81	0.75	0.69

The precision, recall, and F1 score for each model and examined hypothesis are shown in Table 6. We took the cross-validation of 5 folds to train the models. The result shows that when using company features to classify whether the IPO would be underpriced or not could achieve the highest average precision of 71% by using the GBDT model. The other three models could only classify the underpriced IPO with an accuracy of about 60%. However, while using investor sentiments to do the classification, we could get the highest average precision of 76% by using the GBDT classifier. Besides, we could get higher precision through the other three classifiers when we used investor sentiments to make the prediction. However, both the recall rate and F1 score are relatively lower than using the company features.

7. Conclusion

OLS regression models validate H1b. That is, the underpriced level is positively correlated to managerial confidence. The coefficient of the proportion of positive words in the prospectus represents the confidence of the issuer and underwriter, and it is significantly higher than zero at the significant level of 1%. It demonstrates that the more positive (or more confident) they are, the more probabilities of underpricing their new issues. This empirical result aligns with the signaling model proposed by Welch and gives us a new approach to predicting the underpricing level of IPOs and making profits. The text-based approach could extract the sentiments and affect the real-world price.

The classification results demonstrate that both company features and investor sentiments could predict whether the IPO would be underpriced or not, which validates the two hypotheses that originated from the trading process. However, as the precisions of classification are not significantly different between the two data sets, we couldn't take any one of the hypotheses as to the major explanation of underpricing occurs in the trading process.

Overall, we validated that the underpricing of IPO came from two stages. In the pricing process, the issuers and underwriters would underprice the new listings in order to convey their confidence towards the company qualities (the signaling model). Their confidence could be measured in many aspects. We took the sentiments of the prospectus as the measurement to explore the application of the text-based approach (Natural Language Processing), which is an indirect measure of managers' sentiments contains lots of psychological information.

In the trading stage, the investors would actively participate in the trade if they regard the company as profitable and has many potentials through their analysis of company features and following

others' sentiments and behaviors if most people are optimistic towards the IPO. Therefore, using company features and the investor sentiment extracted from the online discussion can predict whether the IPO would be underpriced or not.

8. Future Work

There are some limitations and possible issues that would affect the result and conclusion. First of all, the data is insufficient. Butler, Keefe, and Kieschnick (2014) use 48 variables like the company features and IPO characteristics based on their previous research study. Due to the complexity of databases and usage constraints, we only chose a small number of features to study. Besides, the size of online discussions between investors is relatively small, leading to biases in our result.

Secondly, in the progress of extracting investor sentiments from their online discussions, we utilized the same word list we used in parsing S-1 filings. Since the wordlist is summarized from the finance corpus, it may be inappropriate to represent the sentiments in the online discussions.

In order to improve the robustness and effectiveness of the models, I need to gather more data, balanced the sample, and try different models, like LSTM, in extracting the sentiments from the text.

References:

- Ibbotson, R. G., & Jaffe, J. F. (2012, April 30). "HOT ISSUE" MARKETS. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1975.tb01019.x
- Logue, D. E. (1973). On the Pricing of Unseasoned Equity Issues: 1965-1969. The Journal of Financial and Quantitative Analysis, 8(1), 91. doi: 10.2307/2329751
- Baron, D. P. (1982). A Model of the Demand for Investment Banking Advising and Distribution Services for New Issues. The Journal of Finance, 37(4), 955–976. doi: 10.1111/j.1540-6261.1982.tb03591.x
- Welch, I. (1989). Seasoned Offerings, Imitation Costs, and the Underpricing of Initial Public Offerings. The Journal of Finance, 44(2), 421–449. doi: 10.1111/j.1540-6261.1989.tb05064.x
- Hanley, K. W. (1993). The underpricing of initial public offerings and the partial adjustment phenomenon. Journal of Financial Economics, 34(2), 231–250. doi: 10.1016/0304-405x(93)90019-8
- Johnson, J. M., & Miller, R. E. (1988). Investment Banker Prestige and the Underpricing of Initial Public Offerings. Financial Management, 17(2), 19. doi: 10.2307/3665523
- Rock, K. (1986). Why new issues are underpriced. Journal of Financial Economics, 15(1-2), 187–212. doi: 10.1016/0304-405x(86)90054-1
- Welch, I. (1992). Sequential Sales, Learning, and Cascades. The Journal of Finance, 47(2), 695–732. doi: 10.1111/j.1540-6261.1992.tb04406.x
- Jaggia, S., & Thosar, S. (2004). The medium-term aftermarket in high-tech IPOs: Patterns and implications. Journal of Banking & Finance, 28(5), 931–950. doi: 10.1016/s0378-4266(03)00040-2
- Ljungqvist, A., Nanda, V. K., & Singh, R. (2003). Hot Markets, Investor Sentiment, and IPO Pricing. SSRN Electronic Journal. doi: 10.2139/ssrn.282293
- Tian, L. (2011). Regulatory underpricing: Determinants of Chinese extreme IPO returns. Journal of Empirical Finance, 18(1), 78–90. doi: 10.1016/j.jempfin.2010.10.004
- Butler, A. W., Keefe, M. O., & Kieschnick, R. (2014). Robust determinants of IPO underpricing and their implications for IPO research. Journal of Corporate Finance, 27, 367–383. doi: 10.1016/j.jcorpfin.2014.06.002

- 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. doi: 10.1016/j.jfineco.2013.02.017
- Ly, T. H., & Nguyen, K. (2020). Do Words Matter: Predicting IPO Performance from Prospectus Sentiment. 2020 IEEE 14th International Conference on Semantic Computing (ICSC). doi: 10.1109/icsc.2020.00061
- Liu, Y., & Liu, L. (2009). Sales Forecasting through Fuzzy Neural Networks. 2009 International Conference on Electronic Computer Technology. doi: 10.1109/icect.2009.65