ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Economics and Business Economics

The impact of sentiment displayed in social- and traditional media on IPO performance.

Name student: Mathijs Goethals

Student ID number: 455954

Supervisor: Dr. R. Quaedvlieg
Second assessor: Dr. J. Chalabi

Date final version: 23-1-2021

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Acknowledgments

I extend my sincere gratitude towards each person that has helped me in this effort. Without their assistance this paper would not have been written. I would like to extend my gratitude especially to Dr. R. Quaedvlieg (ESE), for his continuous guidance during this endeavour.

I also express my sincere gratitude to my parents: Marc Goethals and Mary Goethals-van der Kuip, for their continuous support during the pursuit of my academic career. They have supported me both morally and economically throughout this bachelor’s degree for which I am forever in their debt.

Abstract

This paper aims to predict the first-day returns of various American IPO’s between 2015 and 2018 by looking at sentiment expressed in tweets and newspaper articles. The sentiments that have been evaluated are positive, negative and uncertain sentiments as defined in Loughran and McDonald’s financial dictionary. There are indications that for both tweets and newspaper articles positive sentiment has a positive impact on first-day returns of IPO’s. In contrast, negative sentiment in tweets and newspaper articles indicates lower first-day returns. There seems to be a difference between the impact of uncertain sentiments in tweets and newspaper articles. Where for newspaper articles increasing uncertain sentiment indicates declining first-day returns, the opposite is true for tweets. For tweets are indications that increased uncertain sentiment produces higher first-returns.

Table of Contents

1. Introduction

In previous years an increased interest from retail investors has been witnessed regarding initial public offerings (IPO’s). Proceeds generated by IPO’s have reached record heights. For example the IPO of Saudi Aramco on December 11, 2019 amassed a total amount of $25.6 billion. These large sums are in part due to big anticipation for certain stock among retail investors. Retail investors or non-institutional investors can be described as non-professional investors who invest smaller amounts of money per person. Institutional investors on the other hand are large organisations that trade with much larger sums of money. The recent IPO’s of companies such as AirBnB and Doordash serve as others examples. These two companies raised $3.5 billion and $3.4 billion through their IPO respectively. At the first day of trading, when retail investors entered the market the share prices of the two stocks shot up. This means that if the underwriter would have made a better estimation of the popularity of the shares, and raised the offer price of both IPO’s, the companies could have raised a greater amount of capital. This paper aims to estimate the sentiment among the public and its influence on the first-day returns of an IPO. If it is possible to accurately estimate the public’s sentiment then this could mean that IPO’s can be priced much more efficiently.

Institutional investors have greater access to sources of private information. Whereas most retail investors only have public sources of information, which leads to information asymmetry. Retail investors can be expected to rely more on public information when it comes to estimating the value of a company going public through an IPO. Although the prospectus, a document containing all financial details and risks for the company going public, is published for every IPO, this is not the best forecaster of retail investor’s behaviour. Due to the size of this document many retail investors do not use it, instead relying on more accessible sources of information such as traditional media. This could explain scenario’s where institutional investors tend to be more conservative in their valuations, whereas retail investors are willing to pay more for shares of the company going public.

Traditional media publish articles on IPO’s which could influence retail investors’ opinions to a great extent. The sentiment in these media outlets could forecast a reaction by the public. If this sentiment is positive (negative), the returns generated by the IPO would therefore be expected to increase (decrease). The same could be said for publicly available online information. Websites such as Yahoo Finance or Investing.com could offer an alternative to newspapers when it comes to influencing the opinion of the retail investors.

IPO’s are influenced by both private and public information. Since the inception of IPO’s, private information has been available to institutional investors to decide on what should be the price for a company going public. In previous years, the role of non-institutional investors has seen an increase in impact on trading prices of IPO’s. This is partly due to the fact that the number of retail investors has increased over the years. As interest rates remain low, investing in equities is relatively more attractive than saving. Websites such as robinhood.com deliver easy access for retail investors to the financial markets. Their impact on IPO’s can be expected to steadily increase over the coming years. However, how could the public appetite for a certain IPO best be estimated? On the one hand retail investors might base their opinions regarding IPO’s on information provided to them by newspapers. On the other hand, the internet and social media might have overtaken traditional media as a more popular source of information regarding IPO’s.

Traditional media sources could give more legitimacy to its message then would be the case for a random online source. On the other hand social media could be a better estimation of what is actually on the minds of retail investors. Perhaps in some cases traditional media can be too conservative in their valuations. This could result in a surge on the first-day of trading as has been witnessed for the IPO’s AirBnB and Doordash. The impact of social media sentiment is an area that requires further exploration, as it could be the key to understanding how an IPO can be priced more efficiently. Furthermore, social media distinguish themselves from traditional media when it comes to feedback. Depending on the platform, social media gives its users the possibility to “like” or “retweet” a message. A message that receives many of these likes or retweets might have greater influence on public opinion or be a better representative of public opinion.

To assess the impact of these different forms of media on the first-day returns of IPO’s, this paper aims to answer the following question:

“Do positive, negative and uncertain sentiments measured in traditional- and social media have a significant impact on first-day returns of American IPO’s?”

When answering this main question, several sub-questions come to mind. These questions regard the impact of both social- and traditional media sentiment on IPO’s. Furthermore these questions aim to determine the extent of the interaction between sentiment of social media users and the sentiment of publications by traditional media outlets. To answer these questions, several control variables are added to the regression. These sub-questions are the following:

- “Is the correlation between sentiment and first-day returns in tweets significantly different from that in newspaper articles?”
- “Is there a significant difference between the impact of different sentiments on the first-day returns of an IPO?”
- “Do first-day returns of an IPO increase with size of such IPO measured in total proceeds?”
- “Do first-day returns significantly differ across various industries?”

- “Does a tweet’s popularity, measured in retweets, increase the IPO’s first-day returns?”
- “Does the timing of a tweet in relation to the IPO date influence the impact of its sentiment?”

Analysis of social media provides insight into the public’s sentiment concerning certain IPO’s. Sentiment on social media could therefore be a better indicator for public sentiment than merely the sentiment derived from newspapers. On the other hand it is possible that Twitter users simply follow traditional media and copy the sentiment expressed in newspaper articles.

The social media platform of choice for this paper is Twitter, an online platform where the users can express their thoughts in so called “tweets”. The Twitter user is bound by a character limit of 280 characters per tweet, forcing users to express themselves concisely. If a certain tweet triggers a response from another, then that user can decide to retweet such tweet. A tweet that receives many retweets might have a larger impact on first-day returns as it receives more attention than a tweet that receives no retweets at all, possibly because many people agree with the content of that tweet and therefore choose to retweet it. Furthermore, the number of retweets gives an indication of the general popularity of a tweet. A tweet with no retweets might be seen by no one, therefore it offers not much insight into the public opinion as it can be seen as only one opinion among many.

In this paper both tweets and newspaper articles published by some of the most influential financial newspapers, are analysed. The tweets have been obtained through the program “Twint” on Python. Python is a programming language that has recently known a rise to popularity and is nowadays one of the most used programming languages worldwide. The newspaper articles have been accessed through a database called “Factiva”, a global news monitoring service which provides a database with over 33,000 sources. The IPO’s of interest have be derived from the Thomson One Database, a database with extensive information on public companies which is very suitable for research into IPO’s.

After collecting the data, Python is used to extract the sentiment from both tweets and newspaper articles. The sentiment words are based on Loughran & McDonald’s financial dictionary, published in 2011. When the sentiment is extracted from the media sources, a regression is conducted where the returns of the selected IPO’s are predicted through the sentiment scores in both media outlets. Sentiment scores are calculated by dividing the total number of words regarding a certain sentiment, through the total number of words in a text. Furthermore, multiple control variables have been added to the regression based upon previous research and new variables which are of interest in a social media setting.

2. Theoretical Framework

The underwriter has a strong role for the marketing of IPO's. He creates the prospectus and visits potential institutional investors to gage interest. It is on the basis of these visits that the underwriters sets an offer price for the IPO. The retail investor therefore has little influence on the offer price of an IPO. Benveniste and Wilhelm state that the pricing of an IPO can reveal private information towards retail investors (Benveniste & Wilhelm, 1990). Private information determines the valuation made by institutional investors regarding the company going public. These valuations are reflected in the offer price set by the underwriter on behalf of the company going public. Therefore the offer price can be viewed as a representation of private information.

Benveniste and Wilhelm identify two groups of investors. One group consists of "regular investors", this group is described in this paper as the institutional investor. The other group is described by Benveniste and Wilhelm as retail investors. It is the second group that this paper focuses on, as they are likely to be more influenced by sentiments in media. This is because institutional investors rely more on private information to decide whether they invest in an IPO. Underpricing might occur according to Benveniste and Wilhelm because the underwriter also suffers from information asymmetry. For institutional investors the willingness to pay is easier to estimate, given the private information they know and the information the underwriter has gained through gaging interest among institutional investors. For retail investors this estimation is difficult as the underwriter does not know the willingness to pay of all retail investors. This can lead to IPO underpricing. If an underwriter can gage public opinion through measuring displayed sentiment, then information asymmetry will decrease, as will levels of under- or overpricing.

Another possible explanation for IPO underpricing is given by Rock. His main explanation for IPO underpricing is found in the winner's curse. The winner's curse can be found in any scenario where people bid against each other. Several subjective elements can cause the price of the item up for bidding to exceed its intrinsic value (Rock, 1986). Aggressive bidding by overconfident investors can be seen as a main cause of overvaluation (Neupane & Poshakwane, 2012). Even in a setting where there is no competition with institutional investors there seem to be cognitive biases amongst retail investors bidding on IPO shares. This is associated with prices to no longer correspond correctly to demand, as the winner's curse might inflate demand. Given that this inflated price will revert back to normal levels, an opportunity arises for investors that recognize the inflated demand. This opportunity can according to Rock be seen as a reward for the cost of investigation. Referring back to the AirBnB example, say that an investor put time and effort into evaluating the company and finds out the first-day results are inflated due to a winner's curse. He could profit from this in the shape of selling short or buying put options, profiting of the expected correction towards the natural, non-inflated price. The profit he would gain through this is his reward for the cost of researching the company.

When looking at impact of retail sentiment on IPO returns previous research by Gao, Meng and Chan has shown that positive sentiment can lead to overpriced IPO's with high first-day returns and long term price reversals (Gao, Meng & Chan, 2016). The price reversal is associated with the winner's curse deflating back to normal levels. Gao, Meng and Chan measured sentiment in the number of retail trades for an IPO. When number of trades increase, the IPO returns increase as well. Therefore the conclusion that positive sentiments in this case is correlated with high first-day returns makes sense. In this paper, the approach to retail sentiment is less financial. High sentiment in either type of media does not directly translate into purchases of a certain IPO, meaning results in this paper might differ significantly from the findings of Gao, Meng and Chan. As sentiment in tweets and newspapers decreases for the sample used in this paper, first-day returns increase. The paper by Gao, Meng and Chen also describes relevance of this topic. As retail investors gain greater impact on first-day returns, their sentiment will become increasingly important in future pricing of IPO's. Social media however is a different method of measuring sentiment displayed by retail investors, and one that could provide a novel contribution to current works.

The signaling theory provides a possible explanation for sentiments displayed in traditional media. This theory has been used in previous research by Bajo and Raimondo to explain the importance of traditional media sentiment in impacting prospective investors (Bajo & Raimondo, 2017). Putting this theory into practice for this paper, positive sentiment in newspaper articles would be expected to signal high IPO quality towards retail investors. However, newspaper articles possibly do not influence retail investors that use other sources of information for their investment decisions.

Textual analysis is done to assess the impact of the sentiment displayed in the tweets and newspaper articles on first-day returns of the selected IPO's. A great amount of literature has been published on textual analysis of sentiment and its impact on various topics. Tim Loughran and Bill McDonald could be seen as the so to speak "godfathers" of this type of analysis. At Notre Dame University they have published various works in recent years on how to conduct textual analysis. In these works they dive deep into the methodology of textual analysis as they provide the reader with an understanding of how to extract sentiment from a variety of texts. These methods and their financial dictionary are used.

The texts that are analyzed by Loughran and McDonald tend to be traditional. For instance in their papers of 2011 and 2013, 10-K lists and S-1 forms respectively, are textually analyzed. In contrast to the texts analyzed by Loughran and McDonald, this paper seeks to put the methods of Loughran and McDonald to use in a more untraditional setting, in the environment of social media. For the sentiment in S-1 forms Loughran and McDonald find that high levels of uncertain sentiment in the text correlate with higher first-day returns.

This paper aims to add to existing literature by accurately describing the relationship between traditional- and social media sentiments. A variety of existing literature can be found on the effect of the Twitter sentiment on IPO's. Relevant research on this topic has been published by Wang and Liew in 2015. However, they do not include any information on sentiment on traditional media. By including sentiment displayed in traditional media, a relationship could be established between the two types of media. Besides this, the research by Wang and Liew focusses on the sentiment up to three days prior to the IPO date. This paper also includes research into the impact of a tweet's timing on first-day returns. Taking one month instead of three days provides a better opportunity to fairly estimate this variable.

In previous research by Loughran and McDonald, the impact of sentiment due to positive words has been difficult to examine because this research would look at company documents, such as a prospectus. In these type of documents, the publishing company has incentive to present themselves in a positive daylight. Therefore negative sentences tend to be expressed in positive words in these documents. Retail investors or journalists have little need for concealing negative judgements, therefore this paper could provide a better impact analysis for words with a positive tone in finance. The fact that this paper does indeed find positive sentiment in tweets to have an impact on first-day results shows that there might be added value in analysing social media in addition to traditional media when it comes to impact on first-day results. The language used in social media is less complicated which makes for easier estimations of sentiment.

3. Data

3.1 Tweets

To gage the sentiment of retail investors, the database from Twitter's database is used. This database is a collection of all tweets, including those relevant to the IPO's in the sample. Twitter is selected to represent the impact of social media. This platform is specifically relevant to share messages containing sentiment regarding a certain topic such as a company's IPO. Other social media platforms such as Facebook or LinkedIn are not used in the same manner.

Using the Twint command through Python a total of 70.197 tweets were retrieved concerning the IPO's in the sample. The code that is used for Twint can be found in the attached appendix under 1.1. More than half of these tweets, 40.442 to be precise concern the IPO of Snap Inc, a social media platform. The fact that this IPO is the most tweeted about from our sample could have been expected, since it is also the IPO that raised the biggest amount of capital. Snap Inc is also a relevant company to social media users, increasing chances of people tweeting about it. Besides the text of the tweets, the number of retweets for all of the tweets are collected. Retweets could act as an indicator of potential first-day returns.

An overview for the total number of tweets for the IPO's, where the IPO's have been ranked by proceeds, can be found in the appendix under 1.2. Twitter users do not seem to pay a great amount of attention to what potential proceeds for an IPO will be before deciding whether they find the IPO worth tweeting about. From the nine IPO's that have reached over 1000 tweets, six IPO's concern companies in the high technology sector. It seems as if Twitter users are more interested in certain industries than they are in IPO proceeds. High Technology IPO's make for 27,33% of the total IPO's however in the list of the nine biggest tweets they represent 66,67%. This is also due to the fact that of the top ten IPO's with the highest proceeds, four companies are active in the high technology sector. However, AXA Equitable and US Foods Holding Corp, the second and third biggest IPO's are severely disregarded by the Twitter users. Sectors outside of high technology seem not to gain as much attention. Especially the finance industry is being disregarded in terms of tweets.

Twitter data indicates that social media does not blindly follow traditional media. Twitter users seem to have their own fields of interest, mainly the high technology sector when it comes to IPO's. For certain industries we might also expect to see an increased impact of sentiment since there is so much more sentiment to be found surrounding these industries. This certainly is the case for the high technology sector, which produces significantly higher impact on first-day returns compared to other industries.

3.2 Traditional media

Traditional financial newspapers possibly have a great impact on the sentiment of retail investors on certain IPO's. An interesting issue at hand is to look at how the sentiment expressed on Twitter relates to the sentiment expressed in the most influential newspapers. Bajo and Raimondo (2017) express that the more reputable and geographically dispersed outlets should produce the greatest impact on IPO pricing. From looking at daily circulation of various influential financial newspapers this research looks at the sentiment that is being produced by the three biggest newspapers: The Wall Street Journal, The Financial Times and The New York Times (Bajo & Raimondo, 2013). The articles from these newspapers on IPOs from 2015 till 2018 with at least a price of $5 per share have been collected, to examine the expressed sentiment. After removing irrelevant companies and articles written at or after the IPO date, 43 articles are included in the sample. A total of 22 companies from the IPO database are covered in these articles.

The distribution of newspaper articles published for the IPO's increasing in proceeds have been published in the appendix under 2.1. Similar to the results for tweets, Snap Inc has gained the most attention from the newspapers in the sample. However, the distribution is not nearly as skewed towards Snap Inc as it was for the tweets. Overall the number of newspaper articles published seems to increase with total proceeds per IPO. Similar to the number of tweets, the number of newspaper articles published on IPO's is higher for the IPO's in the high technology sector. However, the relative difference between the different IPO's in terms of newspaper articles published is much lower than it was for the number of tweets.

Overall there seems to be a bias for both newspaper articles and tweets towards the high technology sector. No matter the sentiment displayed, high technology is significantly correlated with first-day returns. The high technology could be described as a more exciting industry then other industries in the sample, such as financials or healthcare. With the reader in mind, this might lead traditional media to publish more articles on IPO's in this sector in the hope of capturing a larger audience. However, where Twitter users relatively seem to ignore IPO's in the financial sector, newspapers devote a higher amount of their total newspapers published to such sectors which could be described as less glamorous.

3.3 IPO’s

The selection of the IPO data can be visually summarized as follows:

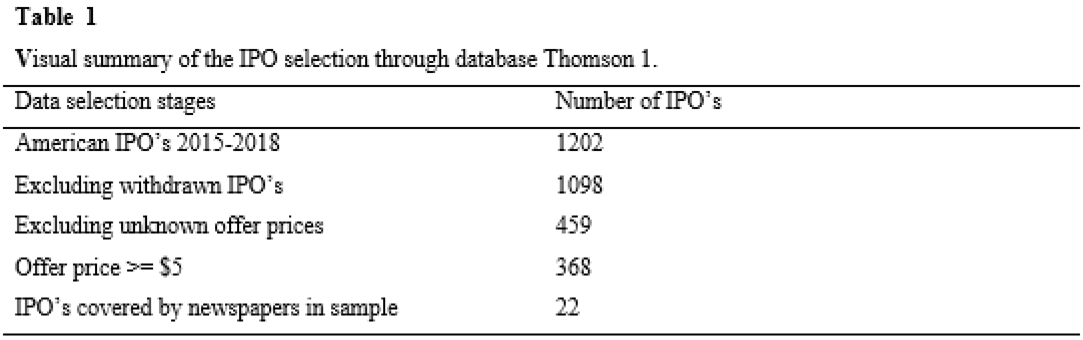


Table 1

Visual summary of the IPO selection through database Thomson 1.

Data selection stages	Number of IPO’s
American IPO’s 2015-2018	1202
Excluding withdrawn IPO’s	1098
Excluding unknown offer prices	459
Offer price >= $5	368
IPO’s covered by newspapers in sample	22

Although the specifications of the IPO, such as offer price, amount raised or business segment vary greatly among these 22 IPO’s, the size of the sample proves to be a limitation. The sample size of 22 unfortunately prevents this study of finding any significant results for the impact of sentiment on first-day returns. Therefore the results in this paper can, except for some control variables, only be considered indicative. A larger sample of IPO’s should make it easier to produce significant results. For further research I recommend to increase this sample size. For instance, more newspapers could be included. Other options to increase sample size are increasing time span or lowering minimum offer price. A summary of the IPO’s can be found in the appendix under 2.2. These remedies are given the fact that traditional media sentiment is of interest. If this is not the case then a larger sample can be obtained by simply abandoning traditional media, as Twitter discusses a wider variety of IPO’s.

If positive sentiment were to lead to positive returns, then the articles or tweets at the day of the IPO and after will also display positive sentiment. However, in the case of no link between sentiment and returns there will still be positive sentiment on the days of the IPO and after. Therefore, it is best not to include the sentiment at the day of the IPO.

3.4 Sentiment

To measure the sentiment regarding an IPO in a tweet or newspaper article, textual analysis is used to derive the sentiment from a text. The use of a dictionary is crucial in defining what texts contain sentiment. The Harvard Psychology Dictionary for instance publishes a list of words and their corresponding sentiment. The sentiment described in this dictionary does not necessarily translate well into what sentiment a certain word has in a financial context. This can be problematic as applying dictionaries outside the domain for which they were developed can lead to serious errors (Grimmer & Stewart, 2013). For example, 73.8% of the negative word counts according to the Harvard list are attributable to words that are not necessarily negative in a financial sense (Loughran

& McDonald, 2011). In this paper a dictionary that has been developed by Loughran and McDonald with regards to a financial context, is used.

In previous research by Loughran and McDonald in 2013, the emphasis for tone has been put on negative tone. This paper established that from all the sentiment words listed in their dictionary, the negative and uncertain words were related to IPO underpricing. Because of the establishment of the relation between these word lists and IPO performance, both word lists have been included in this paper to measure sentiment. However other research describes the fact that retail investors are more attentive to potentially positive signals (Bajo & Raimondo, 2017). Loughran and McDonalds research did not evaluate sentiment as expressed by the public or retail investors, for that reason this paper includes the positive word list from Loughran and McDonald's sentiment dictionary. The inclusion of the positive words list allows the examination of sentiment effects of both positive and negative sentiment on IPO returns. The difference between the two could establish whether there is a certain degree of attention bias among retail investors. The results for newspaper articles indicate that there indeed is some degree of attention bias towards positive sentiment, as positive sentiment triggers a stronger reaction for first-day results than other sentiment.

Furthermore, the timeline to examine American IPO's is from 2015 up to and including 2018. By looking at Twitter's number of active users we see that from 2015 on this number stabilizes around 300 million active users. American IPO's on the American markets are best suited for this research as the subject of most tweets and influential newspaper articles concerning IPO's are about IPO's in the American market. By selecting the American market the assumption can also be made that the amount of sentiment that is being expressed in any language other than English is kept to a minimal amount. Since the dictionary that is used to derive sentiment is written in English, this research disregards any sentiment expressed in other foreign languages.

4. Methodology

4.1 Textual analysis

The method of textual analysis is a process that aims to accurately convert qualitative information from texts into quantitative measures. There are various methods of transforming text into data. The method that is used in this paper, is the dictionary method as described by both Loughran and McDonald in 2011 and Grimmer and Stewart in 2013. This method does not only lend itself well for use on large texts such as newspaper articles, but could be applied to any text, such as a tweet in this case (Grimmer & Stewart, 2013). This method matches the text in the database to the words of a dictionary, in this case the financial sentiment dictionary published by Loughran and McDonald. The code on Python, found in the appendix under 3.1, analyses the tweets and newspaper articles to find all words containing sentiment according to the chosen sentiment dictionary of Loughran and McDonald.

A term weighting scheme is used. Sentiment is expressed as a percentage of the total text. For instance, a tweet that consists of ten words and includes one positive word according to Loughrand and McDonald's dictionary, has a positive score of 10%. This normalizes document length, which is important when the sentiment of a tweet is compared to the sentiment in a newspaper article as the tweet is substantially shorter.

This method can be described as the “bag of words” method, as it looks at texts as a big collection of words. The order of these words is not taken into consideration. This could be problematic if negative statements are rephrased with positive words, or with positive statements the other way around. The bag of words method could prove more problematic for newspaper articles than tweets. However there is no indication from the data that this way of phrasing results in a severely smaller impact of sentiment for newspaper articles, compared to tweets. A feature of tweets is that tweets are very concise and would not use any style forms such as the example referenced above. For instance, a great amount of the tweets in our sample are designed as follows:

“*More on @Fitbit filing for IPO as sales of its #wearables rocket - MedCity News http://medcitynews.com/2015/05/fitbit-files-ipo-sales-wearables-rocket/Â â€¦ #mhealth #digitalhealth*”

These sentences are very suitable for analysis using the bag of words method. In the regression a variable is added which contains the number of retweets. This way it is possible to measure which tweets could have more influence on the overall sentiment and the performance of the IPO in question. A well-recognized Twitter account that has many followers, resulting in many retweets per tweet might have greater influence on the sentiment regarding a certain IPO then might be the case for an average retail investor voicing his opinion.

The Factiva report is published in Microsoft Word. Should there have been hundreds of articles published then it would be necessary to parse the articles in a similar manner to what Loughran & McDonald have done. In contrast to the tweets, these articles need to be parsed because they contain components that are not relevant for the measurement of sentiment. Just as was the case for Loughran & McDonald, the header is removed together with visuals such as tables. By running the code on Python, which can be found in the appendix under 3.1, it is possible to extract the sentiment of both text types.

The Fin-Neg, Fin-Pos and Fin-Unc word lists from Loughran and McDonald are used. It can be assumed that the definitions are still applicable in the current context since the lists are updated every other year (Loughran & McDonald, 2016). These are lists of words that have a respectively negative, positive and uncertain connotation in finance. (Hereafter referred to as “Fin-Neg”, “Fin-Pos” and “Fin-Unc”.) Earlier research shows that Fin-Neg and Fin-Unc are in fact related to IPO underpricing (Loughran & McDonald, 2013). However, due to potential attention bias among retail investors positive sentiment has also been included (Bajo & Raimondo, 2017).

Performing textual analysis for both the tweets and the newspaper articles results in an output where the percentage of either negative, positive, or uncertain words in the text explains the first-day returns of the IPO.

4.2 Twitter Analysis

Microblogging is a way of conveying thoughts online through posting updates, ideas or quick notifications (McFedries, 2007).The selected tweets are different from regular texts as they have certain features that apply to microblogging. These features include hashtags, emoticons, and the necessity of normalizing the text. The latter applies to texts such as: ”Todayyyyy I am HAPPY!!!!”. Where multiple letters are replaced by one and the higher case must be changed into lower case letters (Koulompis, Wilson & Moore, 2011). In Python the text of the tweets are changed into lower case letters. Furthermore, an important feature of microblogging is the fact that there is a limited amount of characters for people to express their sentiment. For this reason, the beforementioned sentiment expressed in a percentage of the total text is important.

In recent studies it has been found that the accuracy of an n-gram classification model can exceed that of a vector model, given that the necessary text transformations have taken place (Dickinson & Hu, 2015). Compared to these options, the bag of words method that is used for this research is just as valid. The bag of words method has been used as a financial standard for many years because of its simplicity and ease of use (Schumaker & Chen, 2006).

4.3 Sentiment IPO performance

To measure the relation between sentiment and IPO performance three different regressions are conducted. One regression concerning the Twitter sentiment, another concerning newspaper sentiment and a third where both sentiments are combined. Taking three different regressions allows for comparisons between the R^2 of the regressions. In the regressions with only tweets or articles their respective explanatory power can be compared. The regression that combines tweets and newspaper articles indicates which of the two has more impact on first-day IPO returns.

For the three regressions, two different specifications are performed. One specification contains all control variables as well as the main explanatory variable. The other specification contains only the explanatory variable. The specification without any control variables is added because it has the biggest chance of reporting significant results. By omitting control variables the estimation of the remaining parameter, namely that of sentiment impact, is the most precise. Therefore general indications of sentiment's impact on first-day returns are based on the specification without control variables

After the sentiment is extracted from the tweets and articles in our sample, several regression have been produced. The y-variable of the regression, named "FIRST_DAY_RETURN" is created by taking the first-day returns. The first-day returns have been calculated as the percental change from the offer price to the price after the first-day of trading. Both offer price and the price after the first trading day have been provided in the Thompson 1 database and can be found in the appendix under 2.2.

Several control variables have been added to the regression. The selection of these control variables is based on previous research conducted by Loughran and McDonald in 2016 and the relevant environment of this study. Based on previous research the dummy variables for calendar years and positive earnings per share (EPS) have been added to the regression. These variables are dummy variables that are either one, for a specific year or if the EPS of a company is positive at time of the IPO. Or zero, if the IPO is not in a specific year and the EPS at time of IPO was negative. The retweets variable and the rt_sentiment variable have been developed specifically for a Twitter environment.

The variable "retweets" counts the number of retweets per tweet and generates the impact of this number on the first-day returns. This means that all sentiments are taken into account. It is possible that retweeted messages, containing relatively high percentages of a sentiment, have a significantly higher impact on first-day returns. To account for the differences between sentiments in retweeted messages a variable is created. This variable, rt_unc, rt_pos or rt_neg, takes the number of retweets from a specific tweet and multiplies it by the number of words from the specific sentiment of interest. To illustrate, for negative sentiments the calculation is as follows: $rt_{neg} = retweets * negativetwt.$ Where the variable "negativetwt" is the number of negative words according to the sentiment list as a percentage of the total number of words in the text.

The data on the timing of the tweets is used to see whether timing impacts first-day returns. As all tweets of a month before an IPO have been selected, every tweet is assigned a value between 1 and 31 for how many days before the IPO the message was sent.

Besides the retweet-related variables other control variables have been added to address several sub questions. These are the control variables for industry and proceeds collected by the IPO. The data for both industry and proceeds have been provided by the Tomson 1 database. The industry variable is a dummy variable for the seven industries included in the sample. By looking at the coefficients for these variables an understanding could be gained into whether first-day returns differ significantly across industries. To answer the question whether first-day returns are higher for larger IPO's collecting more proceeds the log_proceeds variable is added. In the sample proceeds vary greatly across the different companies, ranging from $50 million to $3.91 billion. This variation leads to the distribution of proceeds to be highly skewed. To address this problem and to improve the fit of the model, the logarithm of proceeds are used in the regression. This means that the coefficient for this variable is a percental change.

The final control variable is the returns of the S&P 500 index. This control variable is added to the regression to check for market developments around the time of the IPO. If the return index is very high at the time of an IPO then it might lead to wrong correlations between sentiment and returns when left out of the regression. The S&P 500 specifically is very suitable as it consists of 500 American companies. If the return index did not provide information on a certain IPO date then the information closest before the IPO is taken. The Chicago Board Options Exchange has a freely available file containing a large amount of data on the S&P 500 return index from 1988 up to and including 2019.

To see if the regressions using standard errors are homoscedastic, a white test is conducted. The white test is executed without cross products since this is not possible for the regression regarding newspaper articles. This is because for this regression there would be more variables than there are observations, 42 in this case. The number of variables exceeds the number of observations due to the high amount of cross products generated by the control variables. This results in linear dependency, therefore the white test without cross products is conducted for all regressions. The null hypothesis of this test, homoskedasticity, is rejected for all regressions as can be found in the appendix under 3.2.

After checking for heteroskedasticity it is evident that standard errors need adjustment. To solve this issue, the regressions are performed using clustered standard errors. These standard errors account for heteroskedasticity across several clusters. The clusters in this sample are the 22 different IPO’s. Seen as the cross-sectional regression contains over 70,000 observations for Twitter, the 22 different observations for first-day returns is low. The cluster-robust standard error solves this by clustering the data to company level. By using this standard-error heteroskedasticity within clusters is accounted for.

5. Results

5.1 Unit of Observation and General Remarks

The regressions in this chapter estimate the impact of various variables on the first-day returns of the selected IPO's. For every IPO in the sample there is one first-day return, meaning there is a total of 22 first-day returns in the sample. For these 22 observations many observations of the explanatory variables have been found through the web scraping tool. All 70,035 tweets have been used to extract sentiment that impacts the 22 first-day returns of the IPO's from the sample, as follows from the next paragraphs. As can be deducted from the several regressions that follow, no explaining variables have produced significant correlations with first-day returns. Therefore this paper can merely provide indications. The cause for the insignificant correlation between sentiment and first-day returns lies in the low number of observations for first-day returns in the sample compared to the number of texts. The reduction of the sample to 22 IPO's has proven troublesome. In future research this could be avoided in several ways. For example, the number of newspapers in the sample could be increased. Another solution could be to increase the timespan. Looking at articles over the span of 10 years for example, provides a greater number of IPO's that are covered by newspapers than 22. Without these two limitations, future research should be able to provide significant correlations. As a result of insignificant explaining variables, it is not possible to reject the null hypothesis that social- and traditional media sentiment has no impact on first-day returns.

5.2 Newspaper Regressions

In table 2 the results for the regression including newspaper articles can be found. For newspaper articles all sentiments appear to have a positive impact on first-day returns, whereas the impact of uncertain sentiment is the least positive. This could be explained as retail investors in general do not appreciate uncertainty. The positive coefficient for negative sentiment is surprising. The reason for this could be that negative attention in popular newspapers coincides with underwriters lowering the offer price. This provides a possibility for higher first-day returns, given that retail investors are not as deterred by negative news the newspaper articles because they are affected by the winner's curse.

The year in which the IPO was offered to the public does not have a significant impact on first-day returns in this regression. Perhaps there is little difference between the impact of newspaper articles on first-day returns over the years as the number of subscribers vary slightly over time. This in contrast to tweets, where the audience can be expected to rise with time.

For mostly the bigger industries in the sample there is a significant difference in first-day returns. For the industries of high technology, retail and energy and power first-day returns rise significantly. Apart from energy and power, these are also the industries that represent most companies in the sample. In total these three industries account for 54.55 percent points of the total sample. That all the coefficients for these industries are positive could have been expected as there are only two companies in the sample that have a negative first-day return, both companies are not in the industries mentioned above. As has been discussed before, the large amount of coverage that high technology firms receive might attract many investors, causing companies in this industry to have the highest significant increase in first-day returns.

Proceeds have no significant impact on first-day returns for this regression. This could be explained by the fact that newspaper attention for IPO's was not as concentrated on size of IPO in terms of proceeds, when compared to tweets. Positive earnings per share at the time of IPO do on the other hand have a significantly negative impact on first-day returns. This result is unexpected as high earnings per share would expect to see a positive response for the retail investors during the first day of trading. The cause of this is an interesting topic for further research.

Table 2

Regressions with first-day returns as dependant variable. First-day returns are the change from the offer price to the closing price after the first-day of trading. The Loughran & McDonald dictionaries have been used to classify newspaper articles into positive, negative and uncertain sentiment categories. All regressions include an intercept. The t-statistics are in parentheses. The Cluster-Robust standard error is used, dividing the sample into 22 different clusters, one for every IPO. Regressions include 42 observations concerning 22 companies during 2015-2018. For every sentiment there are two specifications, one with control variables and one without control variables.

Variables	(1)	(2)	(3)	(4)	(5)	(6)
Constant	40.15	28.56***	164.74	27.11**	65.50	29.72**
	(0.13)	(3.58)	(0.60)	(2.48)	(0.19)	(3.08)
% Positive	10.18	14.07				
	(0.98)	(1.17)				
% Negative			8.83	7.33		
			(1.42)	(0.87)		
% Uncertain					-2.00	6.41
					(-0.24)	(1.44)
Control Variables						
Calendar year	-11.59		-23.56		-12.41	
2016	(-0.55)		(-0.94)		(-0.52)	
Calendar year	-25.99		-11.97		-23.81	
2017	(-0.64)		(0.75)		(-0.51)	
Calendar year	16.43		47.59		20.66	
2018	(0.20)		(0.66)		(0.22)	

Energy and	51.54**	44.15	57.14**
Power	(2.30)	(1.64)	(2.24)
Financials	20.01	15.94	20.72
	(1.06)	(0.66)	(0.86)
Healthcare	38.11	8.85	32.95
	(0.58)	(0.14)	(0.47)
High Technology	54.63**	51.73*	57.37**
	(2.31)	(1.86)	(2.17)
Materials	15.21	16.65	18.67
	(0.63)	(0.57)	(0.67)
Retail	49.44**	43.13	50.97*
	(2.47)	(1.61)	(2.05)
Log(proceeds)	0.64	-3.42	1.62
	(0.04)	(-0.20)	(0.08)
S&P 500 Return	-0.01	-0.03	-0.02
Index	(-0.17)	(-0.62)	(-0.27)
Positive EPS	-29.34*	-34.37**	-29.55*
dummy	(-1.94)	(-2.14)	(-1.79)
R^2 *value*	53.78%	54.33%	51.79%

** $p < 0.10$*

*** $p < 0.05$*

**** $p < 0.01$*

5.3 Twitter Regressions

Table 3 shows the regression results for the regressions conducted using the Twitter data. The main difference between newspaper articles and tweets concerns negative sentiment. Where negative sentiment produces a positive impact on first-day returns for newspaper articles, the opposite is the case for tweets. A possible explanation for this difference could be that newspaper articles are a better representation of the institutional investor and bookmaker, whereas Twitter and social media are a better indicator of retail investors. After all, if retail investors are negative about a certain IPO then it should be expected to fall on the first day of trading, given that the underwriter has not priced this negativity into his offer price. Examining what forces are behind the differences of sentiment impact on first-day returns makes for an interesting topic of further research. It is surprising that uncertain sentiment has a more positive impact on first-day returns than positive sentiment. Benveniste & Wilhelm’s information asymmetry could explain this. Given that there is a great amount of positivity, this should give the underwriter easier access to this information, which leads to lower levels of underpricing. Higher levels of underpricing in a setting where there is a great amount of uncertainty could offer larger possible first-day returns as offer prices are likely to be set lower.

Unlike the newspaper regressions, the calendar year 2017 does have a significant impact for tweets. This is likely due to the fact that Snap Inc's IPO was in 2017, generating a great number of tweets for this year. It does significantly have negative impact however, signifying that 2017 was not as good a year for first-day returns as were the other years. The calendar year 2015 is included in the constant. For this regression, the industries of high technology, retail and energy and power provide a significant increase in first-day returns. Similar to the previous regression, high technology is the industry producing the highest first-day returns. This is in accordance with the theory that both tweets and newspaper articles display an attention bias towards this industry, resulting in higher returns.

Although tweets in the sample are more focused on the IPO's that generated a large amount of proceeds, proceeds have no significant impact on first-day returns. The indication is according to expectations however, as it signifies that a percent point increase in proceeds increases first-day returns by around 0.4 percent points. As was the case for newspaper articles, for tweets positive earnings per share at the time of IPO produce a significantly negative impact on first-day results. The impact for Twitter is stronger than is the case for newspaper articles.

To measure impact of popularity the Retweets and Retweets * Sentiment variables have been added to the regression. However, no evidence have been found to support the theory that popularity of a tweet has any consequences for the first-day returns or for the weight that a certain tweet might have when deciding its influence on first-day returns.

Although the impact is insignificant, the "Days until IPO" variable indicates that the more tweets are published over a long period of time, the higher first-day returns are. How this impacts differs over various periods of time could serve as a good subject for further research. It might very well be that if an IPO gains a great amount of attention a year before the IPO, but as time further attention drops that this impact decreases with it.

Table 3

Regressions with first-day returns as dependant variable. First-day returns are the change from the offer price to the closing price after the first-day of trading. The Loughran & McDonald dictionaries have been used to classify tweets into positive, negative and uncertain sentiment categories. All regressions include an intercept. The t-statistics are in parentheses. The Cluster-Robust standard error is used, dividing the sample into 22 different clusters, one for every IPO. Regressions include 70,035 observations concerning 22 companies during 2015-2018. For every sentiment there are two specifications, one regression with control variables and one without control variables.

Variables	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Constant	-34.32	40.24***	-33.95	40.76***	-34.27	40.30***
	(-1.42)	(8.21)	(-1.38)	(8.14)	(-1.42)	(7.92)
% Positive	-0.13	0.28				
	(-0.97)	(1.01)				
% Negative			0.07	-0.41		
			(0.32)	(-0.77)		
% Uncertain					-0.08	0.40
					(-0.47)	(0.20)
<u>Control Variables</u>						
Calendar year	5.00		4.89		4.99	
2016	(0.35)		(0.35)		(0.35)	
Calendar year	-54.88***		-54.98***		-54.83***	
2017	(-3.60)		(-3.64)		(-3.60)	
Calendar year	7.72		7.87		7.80	
2018	(0.25)		(0.25)		(0.25)	
Energy and	54.93**		54.49**		54.92**	
Power	(2.17)		(2.12)		(2.17)	
Financials	37.40		37.16		37.37	
	(1.51)		(1.48)		(1.51)	
Healthcare	56.97		56.89		57.00	
	(1.20)		(1.20)		(1.20)	
High Technology	72.98***		72.93***		72.96***	
	(3.77)		(3.76)		(3.77)	
Materials	39.35		39.04		39.32	
	(1.55)		(1.51)		(1.54)	
Retail	54.07**		53.93**		54.07**	
	(2.69)		(2.69)		(2.69)	
Log(proceeds)	9.59		9.51		9.60	
	(1.21)		(1.21)		(1.21)	
S&P 500 Return	-0.01		-0.01		-0.01	
Index	(-0.45)		(-0.45)		(-0.45)	

Positive EPS	-54.34***	-54.39***	-54.32***
dummy	(-3.96)	(-3.96)	(-3.96)
Retweets	-0.00	-0.00	-0.00
	(-0.41)	(-0.30)	(-0.65)
*Retweets **	-0.01	-0.01	0.00
Sentiment	(-0.15)	(-0.22)	(1.04)
Days Until IPO	0.07	0.06	0.07
	(0.83)	(0.84)	(0.84)
R^2 *value*	77.49%	77.49%	77.49%

* $p < 0.10$

** $p < 0.05$

*** $p < 0.01$

5.4 Newspaper & Twitter Regressions

In table 4 the final regressions are stated. Here a combination can be found of both the tweets and newspaper articles, therefore containing all texts of the total corpus. As is found in the previous separate regressions for tweets, negative sentiment has a negative impact on first-day returns whereas positive and uncertain sentiments produce positive results for first-day returns. For newspapers the results are only different for uncertain sentiments, as in this regression they produce negative impact on first-day returns for those articles. The correlation coefficients for sentiment have decreased slightly for newspaper articles if compared to previous regressions. Taking all regressions into consideration, there are indications that correlation between sentiment and first-day returns are different for tweets and newspaper articles. This can be deducted from the correlation coefficients, as they are much larger for newspaper articles. An explanation for this could be the amount of exposure that the average newspaper article receives compared to a tweet. Every newspaper article is guaranteed to be read by a number of subscribers of the newspaper, there is no such guarantee for a tweet's exposure. Furthermore, from the different coefficients it can be deduced that there is a difference between the impact of the different sentiments on the first-day returns of an IPO. For both tweets and newspaper articles positive sentiment seems to have the largest positive impact on first-day returns. However, for negative and uncertain sentiment the results are different for tweets and newspaper articles. The source of this difference is an interesting topic for further research.

For every calendar year the first-day returns increase. This increase over time, especially rising by a great amount for 2018, is surprising. In the appendix under 4.1 the average first-day return for the IPO's per year can be found. The jump in the regression for 2018 is surprising since in the sample first-day returns decline steadily over the course of time.

From this regression it can also be deduced that first-day returns differ significantly across certain different industries. Just as was the case for earlier regressions, the high technology sector produces the highest significant coefficient. An overview of the average first-day returns can be found in the appendix under 4.2. Looking at this table it comes as no surprise that high technology sector produces the highest significant correlation with first-day returns. This average consists out of 6 different IPO's. For the industry of Energy and Power the average might be highest of the sample. However this average has been derived from just one IPO. Because of the small representation in the sample the dummy variable for Energy and Power produced insignificant results.

Total proceeds of an IPO increase the first-day returns. Higher proceeds could indicate higher levels of demand for a certain IPO. This demand would then increase the gap between offer price and stock price after the first-day of trading. The S&P 500 return index indicates that when returns on the S&P 500 are high that first-day returns decrease. Perhaps that at such times of optimism ruling the markets, underwriters increase their offering prices, leading to lower possibilities for first-day returns. Also for this regression a positive EPS has a significantly negative impact on the first-day results of an IPO. Researching what causes the market to react in such a strange manner to positive earnings is a good topic for further research.

Table 4

Regressions with first-day returns. First-day returns are the change from the offer price to the closing price after the first-day of trading. The Loughran & McDonald dictionaries have been used to classify tweets into positive, negative and uncertain sentiment categories. All regressions include an intercept. The t-statistics are in parentheses. The Cluster-Robust standard error is used, dividing the sample into 22 different clusters, one for every IPO. Regressions include 70,077 observations concerning 22 companies during 2015-2018.

Variables	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Constant	335.27	40.24***	335.94	40.76***	334.35	40.30***
	(0.81)	(8.20)	(0.81)	(8.14)	(0.81)	(7.92)
% Positive Twitter	-0.17	0.28				
	(-1.03)	(0.32)				
% Negative Twitter			0.18	-0.41		
			(0.79)	(-0.77)		
% Uncertain Twitter					-0.06	0.10
					(-0.36)	(0.20)
% Positive News	3.03	3.06				
	(1.14)	(0.38)				
% Negative News			-0.54	-0.28		
			(-0.20)	(-0.01)		

% Uncertain News			1.82	-0.75
			(0.20)	(-0.22)
<u>*Control Variables*</u>				
Calendar year 2016	12.84	12.69	12.78	
	(0.57)	(0.57)	(0.57)	
Calendar year 2017	17.04	16.85	16.87	
	(0.19)	(0.19)	(0.19)	
Calendar year 2018	143.10	143.45	142.76	
	(0.93)	(0.93)	(0.92)	
Energy and Power	72.69	71.93	72.58	
	(1.65)	(1.62)	(1.64)	
Financials	27.14	26.64	27.12	
	(1.16)	(1.13)	(1.15)	
Healthcare	37.23	37.07	37.28	
	(0.63)	(0.62)	(0.62)	
High Technology	65.15***	65.07***	65.12***	
	(2.92)	(2.91)	(2.92)	
Materials	50.80	50.20	50.72	
	(1.48)	(1.46)	(1.48)	
Retail	53.88**	53.65**	53.82**	
	(2.67)	(2.65)	(2.67)	
Log(proceeds)	9.37	9.21	9.37	
	(1.53)	(1.51)	(1.53)	
S&P 500 Return	-0.10	-0.10	-0.10	
Index	(-0.90)	(-0.90)	(-0.89)	
Positive EPS	-50.25***	-50.36***	-50.24***	
dummy	(-3.61)	(-3.61)	(-3.60)	
R^2 *value*	77.47%	77.50%	77.46%	

*: $p < 0.10$

**: $p < 0.05$

***: $p < 0.01$

6. Conclusion

By analysing the corpus, consisting of texts from Twitter and newspapers, this paper tries to gain an understanding of the influence of sentiments displayed in traditional- and social media's on the first-day results of various IPO's. The main question this paper aims to answer is: Do positive, negative and uncertain sentiments measured in traditional- and social media have a significant impact on first-day returns of American IPO's? The answer to this question is that for all sentiments no evidence is found in support of a significant impact on the first-day returns. This is true for both newspaper articles and tweets.

Insignificant results for the explanatory variables are due to a limitation for this research in the size of the sample. The cause of the limitation is that the sample contains merely 22 different IPO's, while thousands of texts are analysed. In future research significant impact for sentiment could be measured by using a regression containing more IPO's than this one. As is described in paragraph 3.3, the main constraint on the sample is the selection of IPO's covered by the three most prestigious newspapers. For future research, the sample can be increased whilst still maintaining the analysis of traditional media. Ways of doing this are increasing the timespan of the sample, or including more newspapers. However, comparing social media with online media sources such as Yahoo Finance could also serve as an interesting topic for further research.

For tweets and newspaper articles there are indications that positive sentiment has a positive impact on first-day returns. This is according to expectation just like the indications that for both newspaper articles and tweets, negative sentiment has a negative impact on the first-day returns. However, for uncertain sentiment the reaction for first-day results concerning tweets and newspaper articles differ. For tweets higher levels of uncertainty indicate higher first-day returns. In contrast for newspaper articles, high levels of uncertainty are correlated with lower first-day returns. What the reason is for these inverse relationships between sentiments and first-day returns is an interesting topic for further research.

Furthermore this paper answers several sub-questions regarding sentiment analysis and the specific sample used. Firstly, the null hypothesis that there is no significant difference between the correlation of sentiment in tweets and newspaper articles and first-day results cannot be rejected based on the p-scores of the conducted regressions. Moreover, this study indicates that there is a difference between the impact of different sentiments on the first-day returns of an IPO. However due to statistical limitations in the sample this result is not significant. Therefore the null hypothesis that there is no difference between the impact of different sentiments on the first-day returns is not rejected.

For tweets it seems that proceeds have larger impact on first-day returns than would be the case for newspaper articles. However the p-score for the variable *log_proceeds* is higher than 0.1 for all regressions. This means that the null hypothesis that an increase in proceeds does not increase first-day returns is not rejected.

Industry does play a significant role in deciding first-day returns. For the eight industry dummy variables, the industries of high technology and retail produced significantly positive coefficients. The significantly best performing industry in terms of first-day returns is High Technology. This raises the question whether this industry performs better due to increased attention from the public, which serves as an interesting topic for further research. The null hypothesis that first-day returns do not differ across various industries is rejected.

For tweets, surprisingly the popularity, measured in retweets, has no significant impact on its relationship with first-day returns. In all regressions the "Retweets" and "Retweets * Sentiment" variables are insignificant and close to zero. The timing of a tweet, expressed in days prior to the IPO neither produces a significant coefficient for its impact on first-day returns.

Based on these conclusions, practitioners should consider using the sentiment expressed by both journalists and the general public when estimating the first-day returns of future American IPO's. By collecting tweets through Python and using a database for newspaper articles a good estimation can be made of the first-day returns.

In terms of other further research several topics relating to this thesis could provide interesting additional information. As this paper analyses retail investor's sentiment, it would also be interesting to gain insight into what institutional sentiment means for first-day returns. Another suggestion is to analyse the impact of sentiment, retail or institutional, on the offer price. The offer price has a great amount of influence on the potential for first-day returns. Therefore this is relevant to this paper. Furthermore, a limitation of this paper is that it has little information of the changes made in offer prices. Future research that provides a possible link between sentiment and offer price revisions made by underwriters could help establish the source of some surprising results my analysis produces.

7. Literature

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, *27*(2), 1–19.

Bajo, E., & Raimondo, C. (2017). Media sentiment and IPO underpricing. *Journal of Corporate Finance*, *46*, 139–153.

Gao, S., Meng, Q., & Chan, K. C. (2016). IPO pricing: Do institutional and retail investor sentiments differ? *Economics Letters*, *148*, 115–117.

Bollen, J., Mao, H., & Pepe, A. (2009). Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. *Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena*, 450–453.

Rock, K. (1986). Why new issues are underpriced. *Journal of Financial Economics*, *15*(1–2), 187–212.

Liew, J. K.-S., & Wang, G. Z. (2015). Twitter Sentiment and IPO Performance: A Cross-Sectional Examination. *SSRN Electronic Journal*, 1–2.

Das, S. R. (2014). Text and Context: Language Analytics in Finance. *Foundations and Trends® in Finance*, *8*(3), 145–261.

Neupane, S., & Poshakwale, S. S. (2012). Transparency in IPO Mechanism: Retail investors' participation, IPO pricing and Returns. *SSRN Electronic Journal*, 2065–2075.

Dickinson, B., & Hu, W. (2015). Sentiment Analysis of Investor Opinions on Twitter. *Social Networking*, *04*(03), 62–71.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, *40*(16), 6266–6282.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297.

Koulompis, E., Moore, J., & Wilson, T. (2009). *Twitter Sentiment Analysis: The Good the Bad and the OMG!* Fifth International AAAI Conference on Weblogs and Social Media.

Benveniste, L. M., & Wilhelm, W. J. (1990). A comparative analysis of IPO proceeds under alternative regulatory environments. *Journal of Financial Economics*, *28*(1–2), 173–207.

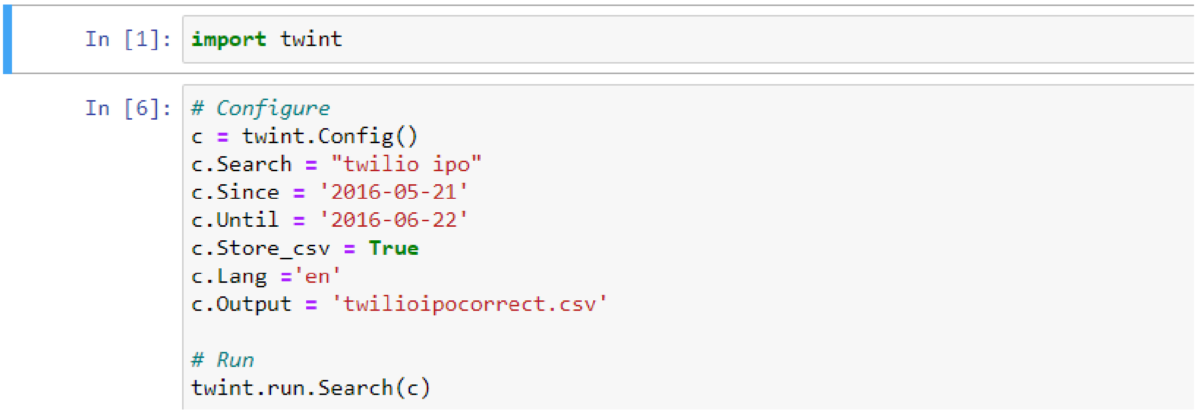
Loughran, T.,& McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Loughran, T., & McDonald, B. (2012). IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language. *SSRN Electronic Journal*, 307–326.

McFedries, P. (2007). *Technically Speaking: All A-Twitter* (Nr. 10). IEEE.

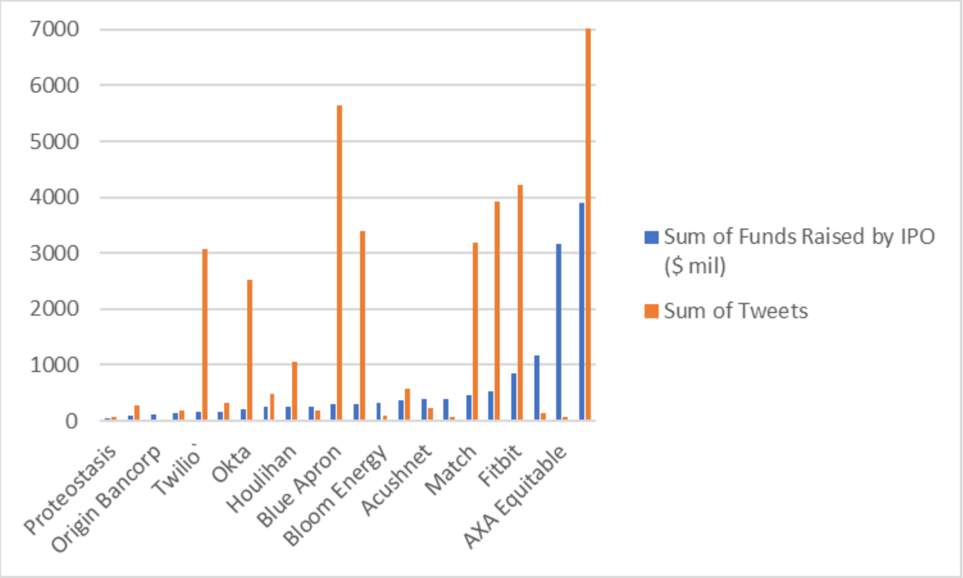
8. Appendix

1.1: Twint code for obtaining the tweets regarding the IPO of the company Twilio.

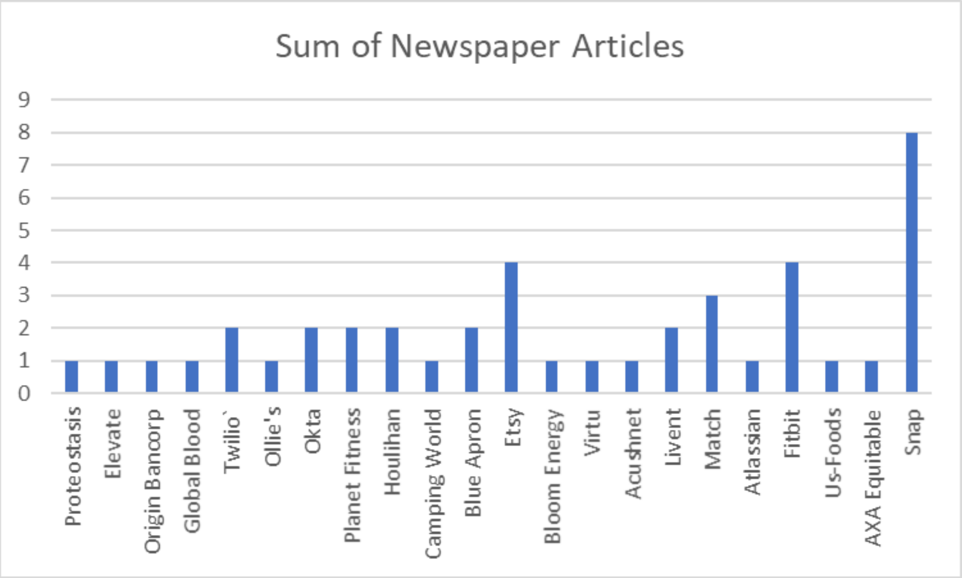


```
In [1]: import twint

In [6]: # Configure
        c = twint.Config()
        c.Search = "twilio ipo"
        c.Since = '2016-05-21'
        c.Until = '2016-06-22'
        c.Store_csv = True
        c.Lang ='en'
        c.Output = 'twilioipocorrect.csv'

        # Run
        twint.run.Search(c)
```

1.2: Total number of tweets for the IPO's in the sample, where IPO's have been ranked by proceeds.



2.1: Distribution of newspaper articles published per IPO, ranked by proceeds.



2.2: Summary of IPO's

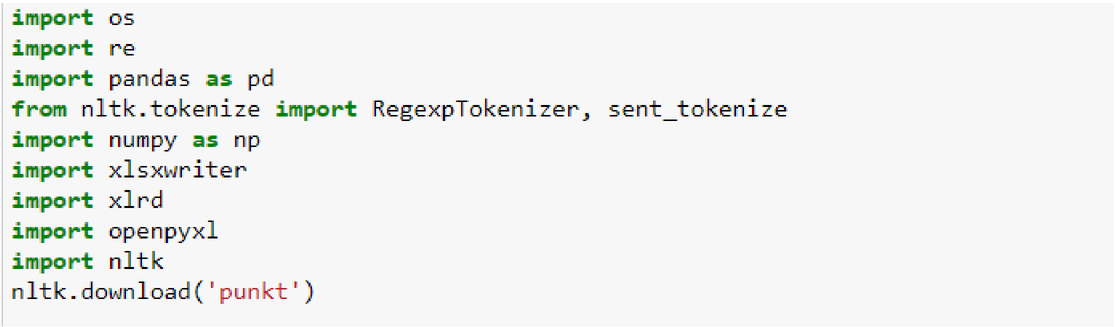
Issuer	TF Macro Description	Issue Date	Market Place	Proceeds Amt inc. over Sold – in this Mkt ($ mil)	Offer Price ($)	Stock Price at Close of Offer/ 1st Trade
Acushnet Holding Corp	Consumer Products and Services	27/10/2016	U.S. Public	377.97	17	18
Bloom Energy Corp	Energy and Power	24/07/2018	U.S. Public	310.50	15	25
Virtu Financial Inc	Financials	15/04/2015	U.S. Public	361.23	19	22
Houlihan Lokey Inc	Financials	12/08/2015	U.S. Public	253.58	21	22
Camping World Holdings Inc	Financials	06/10/2016	U.S. Public	261.19	22	23

Elevate Credit Inc	Financials	05/04/2017	U.S. Public	92.69	7	8
Origin Bancorp Inc	Financials	08/05/2018	U.S. Public	123.63	34	38
AXA Equitable Holdings Inc	Financials	09/05/2018	U.S. Public	3,156.75	20	20
Global Blood Therapeutics Inc	Healthcare	11/08/2015	U.S. Public	138.00	20	43
Proteostasis Therapeutics Inc	Healthcare	11/02/2016	U.S. Public	50.00	8	6
Fitbit Inc	High Technology	17/06/2015	U.S. Public	841.23	20	30
Match Group Inc	High Technology	18/11/2015	U.S. Public	460.00	12	15
Atlassian Corp Plc	High Technology	09/12/2015	U.S. Public	531.30	21	28
Twilio Inc	High Technology	22/06/2016	U.S. Public	150.00	15	29
Snap Inc	High Technology	01/03/2017	U.S. Public	3,910.00	17	24
Okta Inc	High Technology	06/04/2017	U.S. Public	215.05	17	24
Livent Corp	Materials	10/10/2018	U.S. Public	391.00	17	17

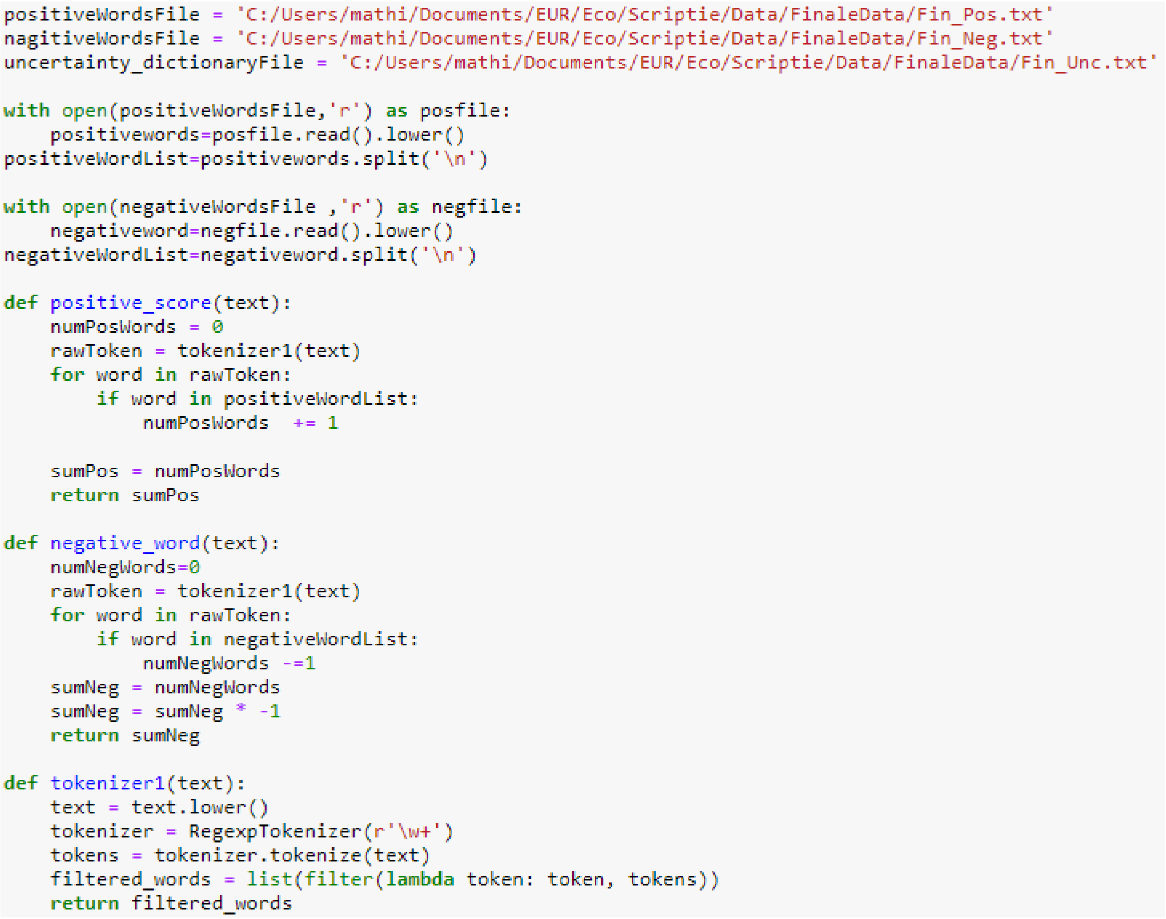
Planet Fitness Inc	Retail	05/08/2015	U.S. Public	248.40	16	16
Etsy Inc	Retail	15/04/2015	U.S. Public	306.67	16	30
Ollie’s Bargain Outlet Hldg	Retail	15/07/2015	U.S. Public	164.22	16	21
US Foods Holding Corp	Retail	25/05/2016	U.S. Public	1,175.56	23	25
Blue Apron Holdings Inc	Retail	28/06/2017	U.S. Public	300.00	10	10

3.1: Python Code for the textual analysis of newspaper articles and tweets

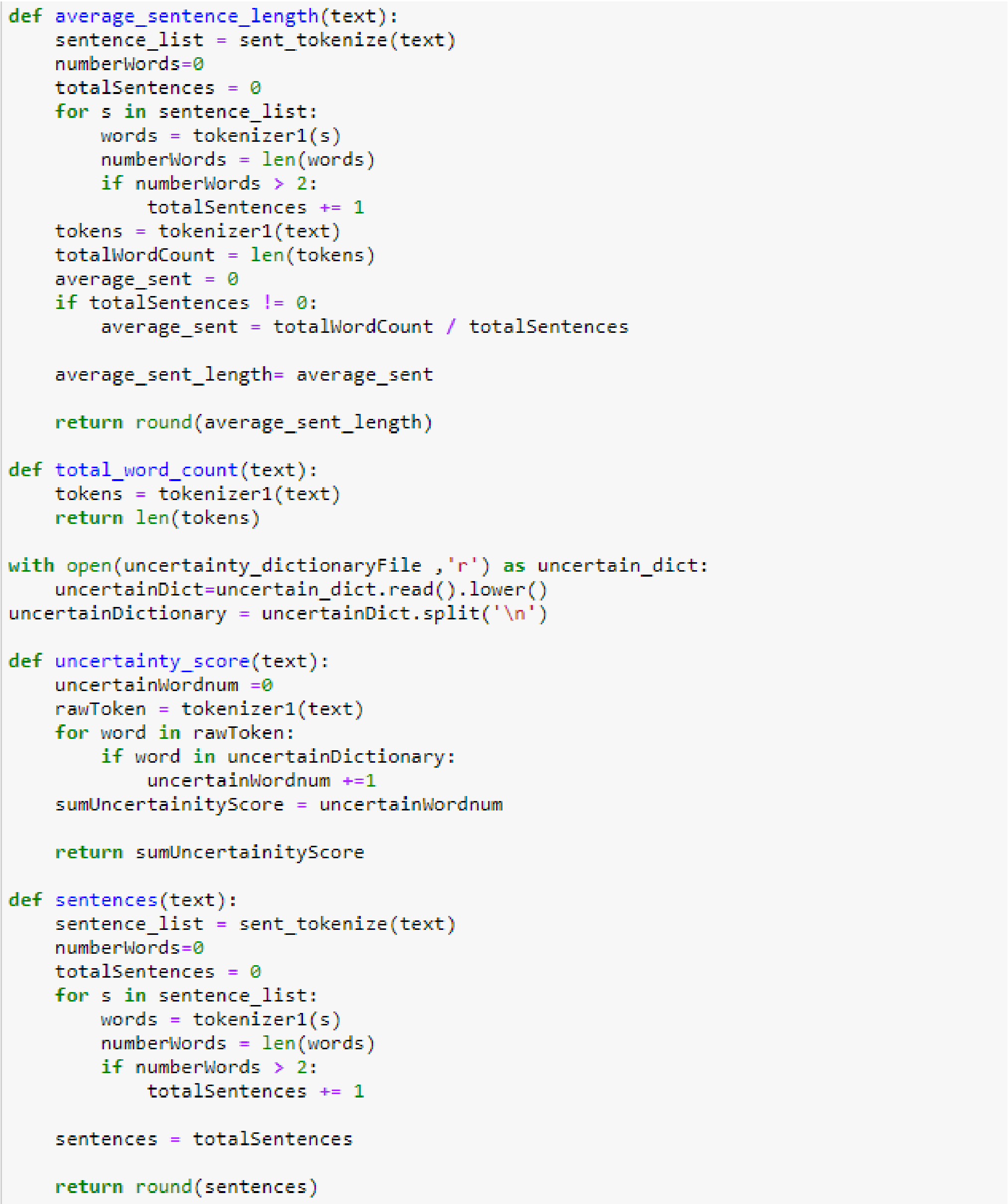
```
Algorithm 1
Input
    -   Word lists from Loughran & McDonald’s sentiment dicitionary.
    -   List of tweets and newspaper articles.
Output
    -   Sentiment words in each text.
For each text in list do
    -   Count number of sentiment words
```



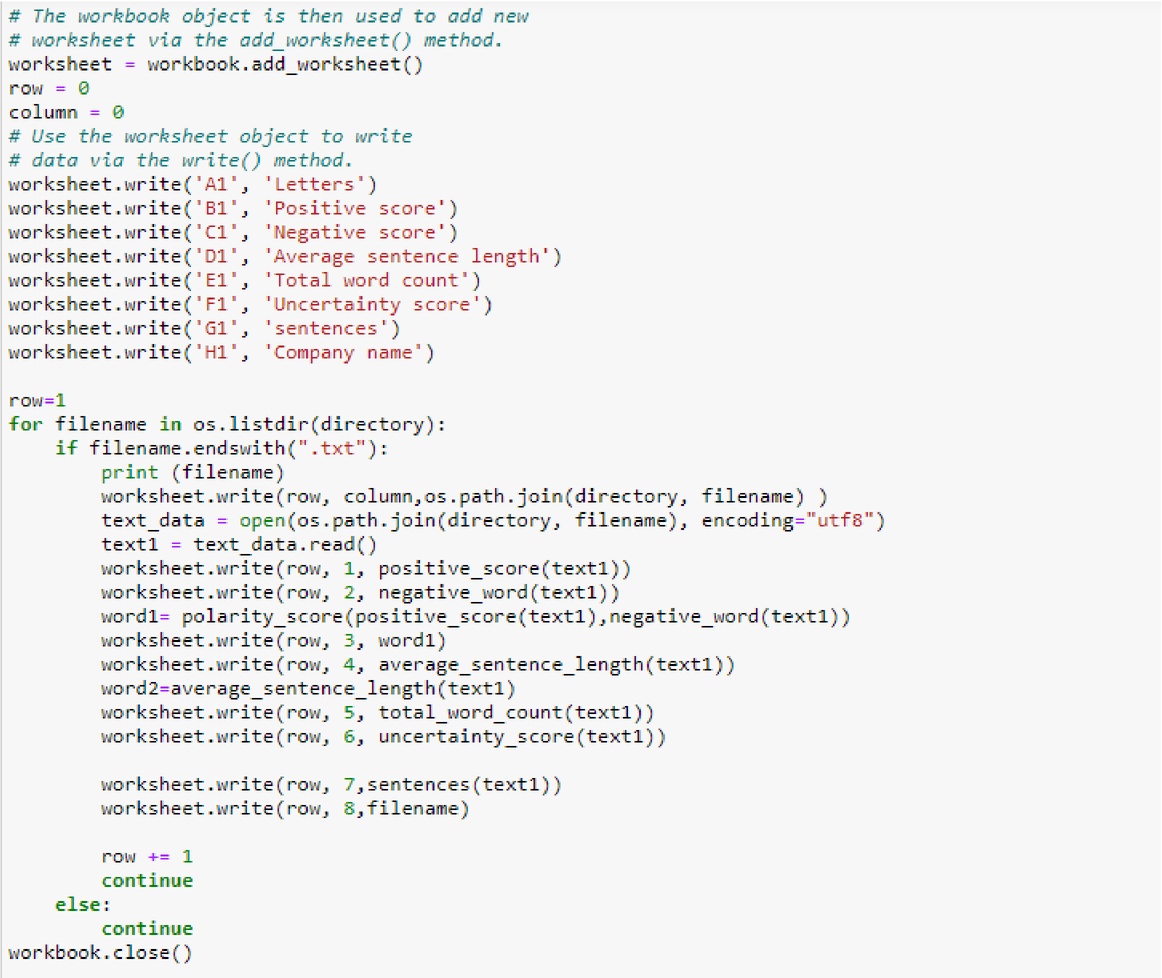
```
import os
import re
import pandas as pd
from nltk.tokenize import RegexpTokenizer, sent_tokenize
import numpy as np
import xlsxwriter
import xlrd
import openpyxl
import nltk
nltk.download('punkt')
```



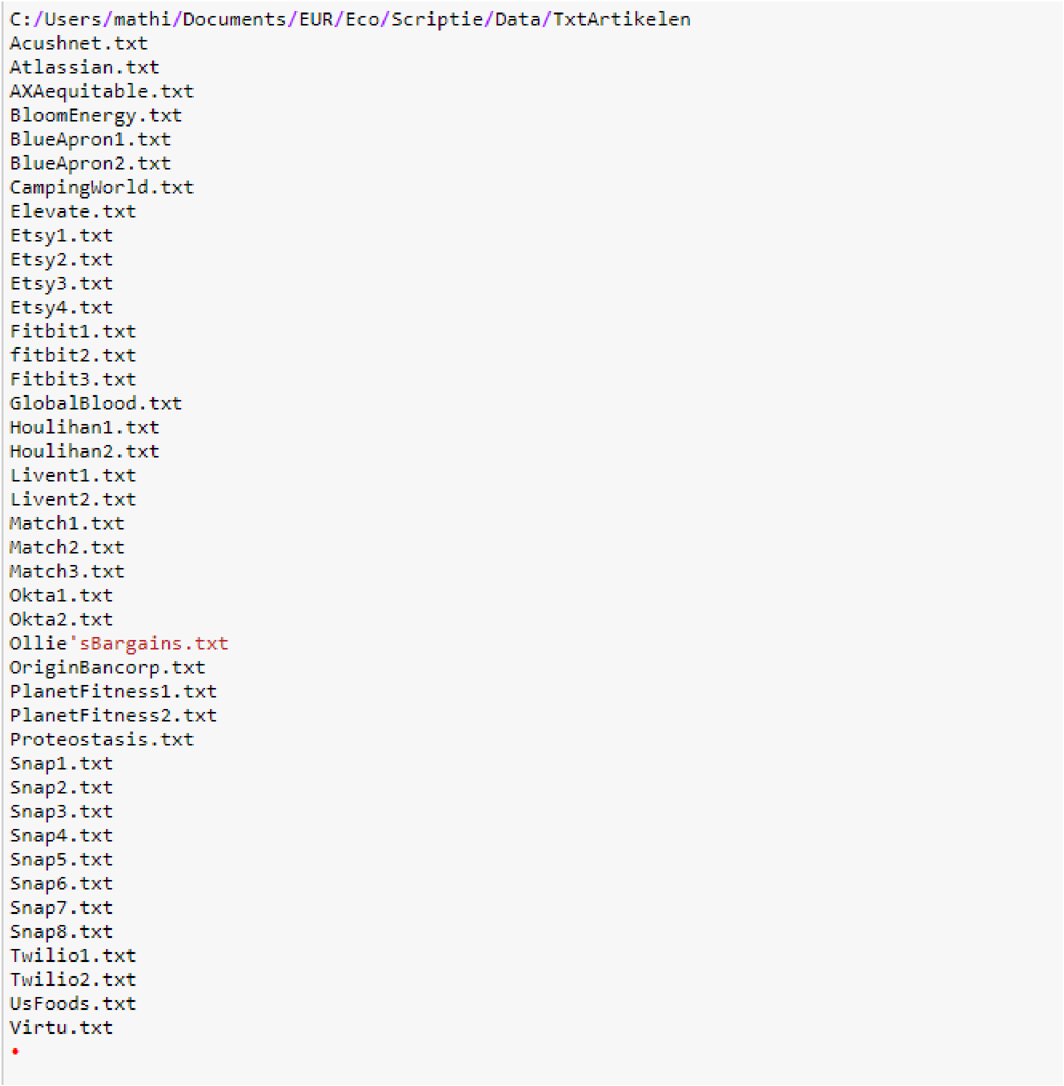
```
positiveWordsFile = 'C:/Users/mathi/Documents/EUR/Eco/Scriptie/Data/FinaleData/Fin_Pos.txt'
nagitiveWordsFile = 'C:/Users/mathi/Documents/EUR/Eco/Scriptie/Data/FinaleData/Fin_Neg.txt'
uncertainty_dictionaryFile = 'C:/Users/mathi/Documents/EUR/Eco/Scriptie/Data/FinaleData/Fin_Unc.txt'

with open(positiveWordsFile,'r') as posfile:
    positivewords=posfile.read().lower()
positiveWordList=positivewords.split('\n')

with open(negativeWordsFile ,'r') as negfile:
    negativeword=negfile.read().lower()
negativeWordList=negativeword.split('\n')

def positive_score(text):
    numPosWords = 0
    rawToken = tokenizer1(text)
    for word in rawToken:
        if word in positiveWordList:
            numPosWords  += 1

    sumPos = numPosWords
    return sumPos

def negative_word(text):
    numNegWords=0
    rawToken = tokenizer1(text)
    for word in rawToken:
        if word in negativeWordList:
            numNegWords -=1
    sumNeg = numNegWords
    sumNeg = sumNeg * -1
    return sumNeg

def tokenizer1(text):
    text = text.lower()
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(text)
    filtered_words = list(filter(lambda token: token, tokens))
    return filtered_words
```



```python
def average_sentence_length(text):
    sentence_list = sent_tokenize(text)
    numberWords=0
    totalSentences = 0
    for s in sentence_list:
        words = tokenizer1(s)
        numberWords = len(words)
        if numberWords > 2:
            totalSentences += 1
    tokens = tokenizer1(text)
    totalWordCount = len(tokens)
    average_sent = 0
    if totalSentences != 0:
        average_sent = totalWordCount / totalSentences

    average_sent_length= average_sent

    return round(average_sent_length)

def total_word_count(text):
    tokens = tokenizer1(text)
    return len(tokens)

with open(uncertainty_dictionaryFile ,'r') as uncertain_dict:
    uncertainDict=uncertain_dict.read().lower()
uncertainDictionary = uncertainDict.split('\n')

def uncertainty_score(text):
    uncertainWordnum =0
    rawToken = tokenizer1(text)
    for word in rawToken:
        if word in uncertainDictionary:
            uncertainWordnum +=1
    sumUncertainityScore = uncertainWordnum

    return sumUncertainityScore

def sentences(text):
    sentence_list = sent_tokenize(text)
    numberWords=0
    totalSentences = 0
    for s in sentence_list:
        words = tokenizer1(s)
        numberWords = len(words)
        if numberWords > 2:
            totalSentences += 1

    sentences = totalSentences

    return round(sentences)
```



```
# The workbook object is then used to add new
# worksheet via the add_worksheet() method.
worksheet = workbook.add_worksheet()
row = 0
column = 0
# Use the worksheet object to write
# data via the write() method.
worksheet.write('A1', 'Letters')
worksheet.write('B1', 'Positive score')
worksheet.write('C1', 'Negative score')
worksheet.write('D1', 'Average sentence length')
worksheet.write('E1', 'Total word count')
worksheet.write('F1', 'Uncertainty score')
worksheet.write('G1', 'sentences')
worksheet.write('H1', 'Company name')

row=1
for filename in os.listdir(directory):
    if filename.endswith(".txt"):
        print (filename)
        worksheet.write(row, column,os.path.join(directory, filename) )
        text_data = open(os.path.join(directory, filename), encoding="utf8")
        text1 = text_data.read()
        worksheet.write(row, 1, positive_score(text1))
        worksheet.write(row, 2, negative_word(text1))
        word1= polarity_score(positive_score(text1),negative_word(text1))
        worksheet.write(row, 3, word1)
        worksheet.write(row, 4, average_sentence_length(text1))
        word2=average_sentence_length(text1)
        worksheet.write(row, 5, total_word_count(text1))
        worksheet.write(row, 6, uncertainty_score(text1))

        worksheet.write(row, 7,sentences(text1))
        worksheet.write(row, 8,filename)

        row += 1
        continue
    else:
        continue
workbook.close()
```



```
C:/Users/mathi/Documents/EUR/Eco/Scriptie/Data/TxtArtikelen
Acushnet.txt
Atlassian.txt
AXAequitable.txt
BloomEnergy.txt
BlueApron1.txt
BlueApron2.txt
CampingWorld.txt
Elevate.txt
Etsy1.txt
Etsy2.txt
Etsy3.txt
Etsy4.txt
Fitbit1.txt
fitbit2.txt
Fitbit3.txt
GlobalBlood.txt
Houlihan1.txt
Houlihan2.txt
Livent1.txt
Livent2.txt
Match1.txt
Match2.txt
Match3.txt
Okta1.txt
Okta2.txt
Ollie'sBargains.txt
OriginBancorp.txt
PlanetFitness1.txt
PlanetFitness2.txt
Proteostasis.txt
Snap1.txt
Snap2.txt
Snap3.txt
Snap4.txt
Snap5.txt
Snap6.txt
Snap7.txt
Snap8.txt
Twilio1.txt
Twilio2.txt
UsFoods.txt
Virtu.txt
```

3.2: Tests for heteroskedasticity

Twitter:

Heteroskedasticity Test: White

Null hypothesis: Homoskedasticity

F-statistic	7453.938	Prob. F(17.70017)	0.0000
Obs*R-squared	45109.76	Prob. Chi-square(17)	0.0000
Scaled explained SS	1132858	Prob. Chi-Square(17)	0.0000

Newspaper articles:

Heteroskedasticity Test: White

Null hypothesis: Homoskedasticity			
F-statistic	3.097225	Prob. F(17.70017)	0.0056
Obs*R-squared	26.92928	Prob. Chi-square(17)	0.0293
Scaled explained SS	15.59655	Prob. Chi-Square(17)	0.4094

4.1: Average first-day returns for every year in the sample.

Year	Average first-day Return
2015	40.23%
2016	27.02%
2017	25.42%
2018	19.62%

4.2: Average first-day returns for every industry in the sample.

Industry	Average first-day Return
Consumer Products and Services	5.59%
Energy and Power	66.67%
Financials	9.51%
Healthcare	46.40%
High Technology	46.29%
Materials	-0.18%
Retail	32.00%