Emerging Technologies in the Last Decade

BAN432 Final Project – Fall 2019

Candidate no: 6, 10, 55, 90, 91

19 November 2019

Abstract

In this report, we try to address the following research question (RQ) using textual analysis: What are the emerging and successful technologies during the last years? Using textual analysis to answer this question has an empirical appeal, namely letting the data speak by itself.

1 Introduction

In order to answer our research question, we will look at documents surrounding Initial Public Offerings (IPOs). The economic rationale is that firms which go public have successfully created a business which has growth potential in the eyes of investors. In other words, if a new successful industry emerges, many firms in this field will go public. Hence, comparing textual information of these firms to firms that went public a decade ago, will inform us about new emerging technologies. To solve the task, we will work with a sample of S-1 and S-1/A forms that the issuer files to the Securities and Exchange Commission (SEC) shortly before the IPO date. The prospectus describes, among others, the proposed business and the risk factors. The sample that we use covers documents that were filed in the 2nd quarter of the years 2008 to 2019, but not all forms from that period are part of the sample.

2 Preprocessing and Cleaning

The sample of S-1 and S-1/A files is classified as unstructured data. Some of the irrelevant information such as information about the company, the file or attachments has already been deleted by the lecturers. However, we had to carry out further cleaning steps. First, we removed punctuation and digits which are not relevant to our word-based RQ. Second, the words were converted to lowercase because it is only of interest what word is used and not how. Later, we also removed stop words, as this word category often occurs in textual data, but does not provide any information relevant to our RQ. We removed stop words in a later step because it was easier to code. Stemming was not a part of our cleaning steps because the algorithm does not always work well, and it is sometimes difficult to find out the meaning of stemmed words. We have cleaned up both provided file versions (the data file that contains all words classes and the one that only consists of nouns and adjectives) and performed our analysis on both versions. The results are nearly the same, which is expected, since technology related words are mostly nouns.

3 A First Indication of Emerging Industries

In general, the SEC's industry classification is not fully trustworthy as it is subject to human judgment. Moreover, a company can change its business, but the classification may not change at all. However, unlike old companies, IPO companies often have a limited business field, which makes it easier to make a human judgment about industry affiliation. In addition, these companies are young and therefore often still active in their industry. Thus, a look at the IPO companies' industries can provide some information.

We have collected the SEC's industry classification for the IPO companies in our sample using the EDGAR API. Unfortunately, the API is unable to collect information about all IPO companies in our sample. We even tried live API calls on the EDGAR website, but it does not work there either. As a result, we were only able to collect information of about two-thirds of the companies, but that should be enough to gain initial insights. The following table shows the most common industries among the IPO companies in our sample. Most companies emerge in the healthcare, oil and computer industry. This may give a first indication of the

industries where technology trends are taking place and about some keywords which we can expect in our word-based analysis.

```
##                                    Industry Description Frequency
## 1                             Pharmaceutical Preparations       157
## 2                           Services-Prepackaged Software        80
## 3                                            Blank Checks        67
## 4  Services-Business Services, Not Elsewhere Classified        55
## 5                         Crude Petroleum and Natural Gas        50
## 6                                            Metal Mining        44
## 7                                  State Commercial Banks        40
## 8     Services-Computer Processing and Data Preparation        35
## 9       Biological Products, (No Diagnostic Substances)        31
## 10       Surgical and Medical Instruments and Apparatus        29
## 11 Services-Computer Programming, Data Processing, Etc.        22
## 12              Services-Management Consulting Services        22
## 13             Savings Institution, Federally Chartered        21
## 14                           Retail-Miscellaneous Retail        17
## 15                                  Gold And Silver Ores        16
```

4 Keywords and Keyness Analysis

As IPO companies go public, in the S-1 forms they write a lot about companies´ business, financial issues and perspectives in order to give potential investors insights and to comply with SEC´s regulations. Therefore, an analysis based on word frequencies will mostly return terms that are related to financial issues. This can be seen in the table below which shows the most frequent words in our study corpus (explained later). As we are not interested in these words, an analysis just based on the IPO files itself is not sufficient to answer our RQ.

```
##           Token Frequency
## 1         notes    380891
## 2           may    279634
## 3         stock    252625
## 4       company    224708
## 5          will    209573
## 6        issuer    205025
## 7        senior    193390
## 8      interest    173995
## 9  indebtedness    166924
## 10   securities    165189
## 11         lien    163072
## 12      million    155622
## 13       amount    150341
## 14         date    142450
## 15   restricted    140240
```

One way to perform this analysis is to compare a study corpus with a reference corpus. Our study corpus consists of 500 files collected from different years. The files were chosen by chance. Even though this is probably not a representative sample from a statistical point of view, we think that the sample is able to answer our RQ. Our study corpus consists of 3.4 million tokens and 52,000 types and is thus big enough to reveal important insights. We have limited the corpus due to performance issues and due to the Heaps' law, which states that with more text (tokens) there are diminishing returns of new vocabulary (types). A reasonable reference corpus should be able to cover a similar language, i.e. it should contain a lot of words about business and finance. We tested two different reference corpora. The first one consists of the 10-K-files from 2018/2019 from the second home assignment. The second one consists of S-1 and S-1/A files from 2004/2005. Since both cover mainly financial language, the results are nearly similar. One might argue that

trends are also present in old companies and thus question the use of a corpus that is up to date. In fact, most trends are emerging in new companies. However, for an empirically elaborate analysis, both reference corpora might be not the best choice because they only cover files from two years and only from one source. The later analysis is based on the 10-K corpus which consists of 11.4 million tokens and 76,000 types. We have chosen this one because it might contain more topics.

We conducted an analysis based on keywords and keyness to identify those words that occur relatively more often in our study corpus than one would expect, in comparison to our reference corpus. The frequency difference of words between the two corpora must be statistically significant. As a statistical measure of association, we used the log-likelihood ratio. This analysis can identify words that stand out. Thus, it is a useful technique for term extraction. In order to achieve good results, we have set several bounds in this analysis. In creating the Document Term Matrix, we have set a global word bound. However, it is hard to find a reasonable bound because there is a trade-off. On the one hand, we do not want to consider words that appear in many documents because they are often financial terms and not relevant to our RQ. This implies that we need to set a low global word bound. However, since IPO companies often emerge in similar industries or use similar technologies, a too low bound would also eliminate relevant trend words too. In addition, we have also set a bound to the actual frequency in order to eliminate words that occur infrequently. In the last step we have only chosen those words with a high significantly log-likelihood ratio. As a result, we obtained a list of about 150 terms. We have selected 20 words that appear unusual, compared to everyday language, and are not financial terms. The words are shown in the table below. The same analysis based on the co-occurrence of words (n-grams) did not reveal additional information since technology words do not often appear in fixed phrases or specific collocations in comparison to financial terms.

```
##       Trend Word Freq Reference Corpus Freq Study Corpus         LL
## 1       hospital                   169             34386 36972.5943
## 2       drilling                   352             14080 13179.3863
## 3         gaming                   722              8522  5768.0910
## 4          cloud                  2937               286  3330.4048
## 5            oil                  2993              9289  1880.4759
## 6      physician                   188              2423  1700.4484
## 7       clinical                  2026              6558  1429.7894
## 8           cars                   145              1897  1339.1682
## 9        digital                  2367               814  1281.1319
## 10         cyber                  1400               301  1106.1743
## 11 semiconductor                  1070               166  1011.5250
## 12 cybersecurity                   938               106  1011.2111
## 13      networks                  1847               648   979.3932
## 14       devices                  2312              1011   945.8447
## 15     analytics                   823               104   850.4676
## 16     household                   120              1300   846.9850
## 17     platforms                  1317               434   743.6182
## 18    automotive                  1033               341   582.2867
## 19    statistics                   182              1171   559.6243
## 20        mobile                  2026              1256   457.4796
```

There are some similarities if we compare our analysis based on keywords and keyness with the SEC's industry classification. In our list of about 150 keywords, many of them are related to the same industries, namely healthcare, oil and the computer industry. We did not take financial words into account in our analysis, because we were not able to clearly differentiate whether there could be a trend in the financial industry or if the words are just related to the IPO. Nevertheless, we are convinced that there are a lot of developments going on in the field of finance due to the digitalization.

4.1 Keywords in Context

After we identified relevant keywords based on the keyness-analysis, we used the keyword in context method to examine how the words are used and whether they can represent any trends. For a given keyword, this method shows which words are commonly used in connection with that keyword. This is shown with a word cloud. We have identified 10 important trends, which are presented in the following table. In this chapter, we explain each trend in detail.

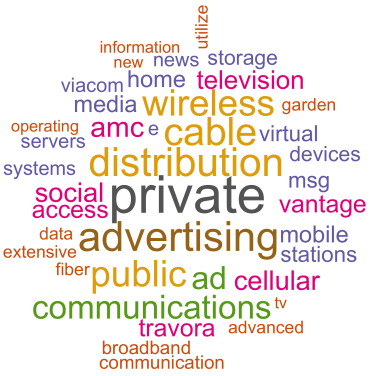
```
##  [1] "mobile"        "digital"       "networks"      "platforms"
##  [5] "cloud"         "analytics"     "automotive"    "oil"
##  [9] "semiconductor" "health"
```

4.1.1 Mobile & Digital Transformation

The word cloud on the left is more related to mobile applications by relating to apps in mobile phones or other mobile devices or platforms. The rapid growth of worldwide smartphone users shifts various applications towards mobile devices and hence also the traditional distribution channels. Besides, the mobile drilling devices units are quite often mentioned, which are not in the same context of mobile applications but are more related to the oil sector. The word cloud on the right shows a relation to the digital transformation in classical business processes such as advertising or marketing as well as in typical industries like automotive or media. As we know, the digitalization is quite often mentioned by many governments because it is regarded as the engine of the third industrial revolution during recent years. Therefore, we believe that there is a trend towards a much deeper digital transformation of companies and processes, as well as towards mobile applications or other similar digital products.

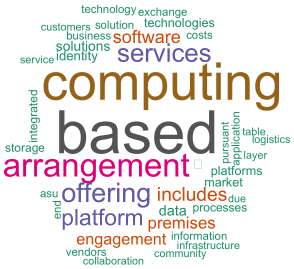
4.1.2 Networks & Platforms

The left word cloud shows a relation to network technologies such as 5G in cellular technology. In relation to this, the usage of private or public networks and the use of wireless and virtual networks are often mentioned. The rapid development of mobile telecommunication, big data analytics and internet of things (IoT) drives the growing demand towards a more efficient network. The right word cloud shows platform technologies. The development of digital platforms is affecting many traditional industries such as the aircraft industry or the maritime industry. For example, due to new environmental regulations the maritime industry is using big data and digital platforms to optimize their shipping routes, improve their energy efficiency and reduce emissions. Additionally, in comparison to the reference corpus the usage of social media platforms has shifted into the center of many companies' attention. Therefore, new networks and platforms technology are important for these companies to gain better access to the insights found on social media and might also be an indicator that telecommunication and IoT companies are restructuring to keep up with the mobile network demands of our increasingly digitalized society.



4.1.3 Cloud & Analytics

From the two word clouds below we find that cloud computing, cloud services, data analytics and information analytics are the main keywords under these topics. The word cloud on the left suggests that the usage of cloud service is widely discussed, such as cloud processes and cloud solutions. The word cloud on the right applies to our topic on data analytics and from there, we can see that analytics is applied in different areas such as marketing, advertising and management. As cloud services are often employed in analytics to provide companies with real-time information analysis on consumer behavior, business performance issues are another important topic for companies as it helps them to track their process and operational efficiency. Moreover, the cloud service and data analytics are not only an emerging area among these IPO companies, but also, they both play a crucial role in the trends mentioned earlier. Across time, data has been increasingly gathered and the amount of publicly available data has increased in the past decade. As companies start to realize the importance of big data to optimize business processes and analyze consumer feedback or sentiment, the demand for data analysts and data analytics has been on the rise. And at the same time, the big data analytics requires new solutions like cloud storage and cloud computing in order not to be limited by physical hardware. These could be one of the contributing factors to the upwards trend in the IPOs in both areas.



4.1.4 Automotive, Oil & Semiconductor

Across the past decade, there is a sharp increase in the awareness of environmental issues. As investors nowadays place a greater emphasis on environmental, social and governance issues, the companies whose business appeal to investors are more likely to perform well through strong investors' (funding) support. This may also be one explanation for the recent trend changes in the oil industry, as firms try to be more

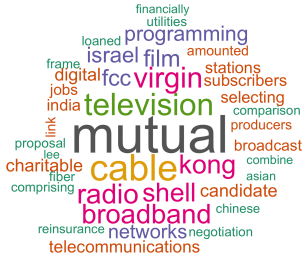
environmentally friendly. The shift towards a stronger preference for green and environment-friendly energy also has an impact on the automotive industry. It has, to a large extent, contributed towards the development of new cars and other mobility devices which are more energy efficient and environment-centric in design. The latest trend in the automotive industry involves the creation of self-driving cars. Led by Tesla, this shift in focus has created a new niche in the automotive industry that has allowed other automobile manufacturers to contribute to automation in the automotive industry. Hence, new firms can easily enter this area of research and development. Some relations to the analytics industry are also revealed, as the judgement of human experts and the collection of road usage data are essentially required in order to enhance and improve machine learning algorithms. Semiconductors are a core ingredient to all the electronic devices that people use. With the increasing trend seen in areas such as mobile, digital and automotive industry, demand for semiconductors will increase and thus this industry can be expected to be a part of the emerging trend in relation to IPO firms.

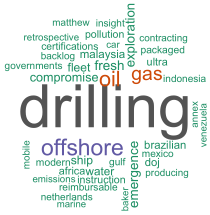
4.1.5 Healthcare Innovations

For the healthcare sectors, word clouds do not give us deep insights, but the high frequency of terms in our list (keyness analysis) in relation to the health sector (hospital, clinical, devices, etc.) indicate that there is something going on in this industry. This trend in the health sector may be linked to digitalization trends such as the increased use of personal electronic devices (mobile phones and smartwatches) to track personal health indicators (activity tracking). This also contributes to the increasing trend displayed in the mobile and digital space. The rising trend in the field of healthcare in the United States may also be partially credited to the aging population that the U.S. is facing, where the number of senior citizens above 65 years old are set to double across 2000 and 2040. This increase directly stimulates more demand on healthcare in the U.S. and contributes to the innovation trends in the health sector.

4.2 Topic Model

In addition to the keyword and keyness analysis, we also estimated a topic model that identifies and categorizes the topics that the IPO companies have talked about. For the topic model, we heavily restricted the Document Term Matrix that we used. We only used those words that have a statistically significant frequency difference compared to the reference corpus. Thus, it is based on the keyness analysis. We estimated a topic model because it can cluster a lot of terms identified by the keyness analysis, especially in relation to the health care industry and the oil sector. Thus, it may provide some additional insights regarding emerging trends. In total we estimated 12 topics. Among them the four key topics are: digitalization, telecommunication, healthcare and oil. The other topics are related to finance, manufacturing, gaming, households, etc. However, the topic model might overall not be the most convincing approach in identifying technological trends. We do not know how well the model assigned the topics. But the method can be regarded as a supplement to the previous analysis.





5 Conclusion

This report shows how textual data analysis can be applied to gather insights in the areas of business as well as how to identify emerging industries and technologies for IPO companies. It is determined that mobile & digital, networks & platforms, cloud & analytics, automotive, oil & semiconductors and healthcare are some of the industries and technologies relevant to emerging companies. To arrive at our analysis, we have used a keyness analysis to identify words that appear more frequently in our corpus than expected. After that the keywords in context method was used in order to identify the relevance of the chosen words. Based on this, a topic model was estimated for a further identification and clustering of the topics among emerging companies. However, there are also limitations on the relevance of the outcomes. Textual data analysis cannot specifically identify the technologies involved, rather it gives a broad overview about emerging industries and technology fields. This might be due to the way companies describe about their technologies, or a lack thereof, since some of these technologies might involve sensitive company information and are regarded as trade secrets. We also realized that the topic model might not be the optimal approach because the result of this model is challenging to evaluate.