UNIVERSITY OF CALIFORNIA,
IRVINE

Prediction of Firms' Annual and Quarterly Return Using NLP Techniques

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Electrical Engineering

by

Shima Nabiee

Committee:
Professor Nader Bagherzadeh, Chair
Associate Professor Matthew Harding
Professor Ender Ayanoglu

2020

Dedication

To my family who gave me all the love and support I needed.

To my awesome brother Ali.

Contents

Contents **iii**

List of Figures **v**

List of Tables **vi**

1 Introduction **1**
1.1 Related Works 2
1.2 My Contributions 4

2 Methodology **5**

3 Data **8**
3.1 EDGAR Dataset 8
3.2 CRSP Dataset 8
3.3 Compustat Dataset 10
3.4 Sentiment Category 10

4 Striping 10-K/Q Down to Text Files **12**
4.1 EDGAR Markup Tags 13
4.2 Parsing 13

5 Feature Selection **16**
5.1 Cosine Similarity 16
5.2 Jaccard Similarity 17
5.3 Minimum Edit Distance Similarity 18
5.4 Simple Similarity 19
5.5 Sentiment of Change 19
5.6 Change of CEO/CFO 19

6 Mechanism **20**
6.1 Fama-MacBeth Regressions 20

7 Results **23**
7.1 Statistical Significance 23
7.2 P Value and T Score 24
7.3 Market Indicators 26
7.4 Main Results 28

8 Conclusion **31**

Bibliography **32**

List of Figures

3.1 TEXT formatted filings 9
3.2 HTML formatted filings 10

7.1 T statistics and P value[8] 26

List of Tables

3.1 CRSP Table 11

7.1 Main Results [3] 28

Acknowledgments

I would first like to thank my thesis advisors Professor Nader Bagherzadeh and Matthew Harding. The door to Prof. Bagherzadeh office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this thesis to be my own work, but steered me in the right the direction whenever he thought I needed it. I would also like to thank Dr. Harding and Dr. Hersh who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for their supports and guidance.

ABSTRACT

Prediction of Firms' Annual and Quarterly Return Using NLP Techniques

By

Shima Nabiee

Master of Science in Electrical Engineering

University of California, Irvine, 2020

Professor Nader Bagherzadeh, Chair

In this thesis, we investigate the fact that changes in firms' annual and quarterly report language are useful signals to predict firms' future returns. A dictionary for all firms filings submitted to the Edgar website from 1993 to present is created. The filings then needed to be parsed and stripped to be able to extract useful information from each document. To measure the change in the language of reporting, several similarity measures are introduced to compute future return indicators based on those. Then, merged COMPUSTAT data, which has the firms' book value, with this information and created features to train the regressor. Also, to be able to train the regressor to predict future returns, we used CRSP data set, which provides each firms' return for each month. The idea is from Lazy Prices paper by Cohen, Malloy, and Nguyen.

Chapter 1

Introduction

Nowadays, most people have to work through so many repetitive tasks. They do this in a variety of ways, finding the most efficient approach, and then stick to that workflow. For instance, consider automatic renewal of magazine subscriptions or payment plans or email address auto-fill. The bias in behavior and the powerful role of inertia, breaking away from these default setting and way of doing these kinds of tasks, takes so much active effort.

When we are talking about the complex tasks, sticking with the existing state of doing these tasks is even more powerful. In corporate life, which is full of these complex repetitive tasks, even C-level officers (CFO, CCO, CEO. . .) face these kinds of tasks and responsibilities abundantly. Consider financial reports, complex multi-hundred-page financial statements that should be filed regularly. An easier and more efficient way to complete these reports is to use a template and have some choices set as default. Relying on a repeatable process is common in financial reporting, since it reduces the administrative burden.

Most of the time, templates and default language are consistent from a reporting period to the next one. There is no remarkable difference between firms' annual or quarterly reports, which means that the structure of these filings is almost the same during each period.

Especially the Management Discussion and Analysis ("MD&A") section and disclosed risk factors usually do not change much from one quarter to the next.

Now, what if there is a big change in the reporting language? First, let's consider what causes these changes. If agents depart from these standardized reporting templates, they can make major changes to the previously static language. Considering that these departures from known report language and structure require new insights and ideas, it suggests there may have been some material change in the firm. Thus, there may be some signals that can help us predict a firm's future returns.

A new working paper, "Lazy Prices" by Cohen, Malloy [3] and Nguyen, suggests that the alternations from previous non-changing reporting, have significant signals about firms' future stock return. They use text processing techniques and training Fama-Mcbeth regressor using similarity measures to predict future stocks returns associated with firms who change their reports ("changers") and compare these to the returns of firms who don't make such changes ("non-changers").

1.1 Related Works

During recent years, there are several growing literature in stock price predictability. Among them, several of them are most related to this work.
Firstly, several papers studied the topic of under-reaction in stock prices and how investor inattention will affect them. As for instance, [14] tests whether stock market investors appropriately distinguish new and old information about firms. They define the "staleness" of a news story as its textual similarity to the previous ten stories about the same firm. The result is tremendous: firms' stock returns respond less to stale news. Even so, a firm's return on the day of stale news negatively predicts its return in the following week. Individual

investors trade more aggressively on the news when the news is stale. This result shows that individual investors overreact to stale information, leading to temporary movements in firms' stock prices. In [4], they measure investor attention using search frequency in Google. They could capture investor attention in a more timely fashion than the previous attention measures and showed that an increase in search frequency predicts higher stock prices in the next 2 weeks and an eventual price reversal within a year. Similar work in [1] measures institutional attention using Bloomberg search activity and shows that stock price drift is most pronounced for stocks with the least amount of institutional attention.

By contrast, what Lazy Prices documents, is an acute form of investor inattention that is centered on the most important firm disclosure that they make, which leads to large return predictability. Using novel data from SEC firms' filings and variations in attention to this exact same item (in annual reports), helps them to predict variation in return patterns. In addition, they believe that the nature of this inattention is because of the fact that investors simply cannot interpret meaningful changes to these documents.

Secondly, As a result of increased computing power and advances in the field of natural language processing, in order to answer important questions in finance, many papers started to apply AI and NLP to automatically analyze financial documents. [10] is one of the most relevant papers to this study. He captures text complexity as a function of syllables per word, words per sentence, and the length of the document; and demonstrated that the annual reports of firms with lower earnings are harder to read and firms with annual reports that are easier to read have more persistent positive earnings. In addition, [6] shows that a positive tone in the MD&A section in SEC log files, is associated with higher current and future returns and that an increasingly negative tone is associated with lower contemporaneous returns.

Lastly, most similar to this paper is [2], which introduces a measure to find the degree to which a particular section (MD&A–firms' Management Discussion and Analysis) differs from the previous 10-K filings, and provides several findings. They showed that the magnitude of stock price responses to 10-K filings is positively associated with the MD&A modification score. However, the decline in the MD&A modification score, made a weaker signal to predict the stock price variations, meaning the that variations in this section are not as useful as it was previously.

In this paper; however, they showed that by isolating each item in this filing and using different similarity signals, they predict large negative returns in the future. The reason that predicted price changes linked with document changes are getting less than what it was before, is basically "because investors are missing these subtle but important signals from annual reports at the time of the releases, perhaps due to their increased complexity and length."

1.2 My Contributions

I was NLP engineer analyst in this work, my responsibility was to implement the text processing and create a dataset of features and labels, which are the firms' monthly returns. More specifically, my contributions and responsibilities were to:

- Create a dataset from all of the Edgar filings from 1993-2018.
- Stripping out the data from these filing– meaning that stripping all the data from HTML coded filings, isolating different items from filings, creating a JSON file for each filing.
- Implementing similarity measures codes, test them and save the data in Pandas DataFrames.
- Merging Compustat dataset with CRSP, to create the labels to be able to train the regressor using them.

Chapter 2

Methodology

There are various sources used in this work to be able to have any possible meaningful factors to train the model. All complete 10-K and 10-Q filings are downloaded from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website from 1993 to 2018.

EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission (SEC). Its primary purpose is to increase the efficiency and fairness of the securities market for the benefit of investors, corporations, and the economy by accelerating the receipt, acceptance, dissemination, and analysis of time-sensitive corporate information filed with the agency. [11]

HTML formatted 10-K and 10-Q filings contain a collection of exhibits, graphics, XBRL files, PDF files, and Excel files [7]. This information is submitted by each filing, which makes it very large and with unnecessary data that are not the kind of texts that are needed. Textual contents are extracted similar to [13]. We removed all the unneeded information such as HTML tags, XBRL tables, exhibits, ASCII-encoded PDFs, graphics, XLS, and other binary files. Also, tables that have numeric characters more than %20 are removed.

We provide two primary data sources associated with 10-X filings on the Security and Exchange Commission's (SEC) EDGAR website. The first stage essentially cleans each filing document of extraneous materials and is described in detail in this section. Separately stage two parses each document into tokens and tabulates various word counts.

A substantial portion of an EDGAR filing's content consists of HTML code, embedded PDF's, jpg's and other artifacts not typically of interest. The complete file size for some of the largest filings exceeds 400MB. The parsing process can be made orders of magnitude more efficient by extracting these items and creating compressed versions of the filings. For example, after the first stage, the largest file is less than 5KB. These considerations are most relevant for the annual and quarterly filings of firms, which is the focus of this process.

Monthly stock returns are obtained from the Center for Research in Security Prices (CRSP), which is a provider of historical stock market data. CRSP maintains some of the largest and most comprehensive proprietary historical databases in stock market research. This research-quality stock database contains 10 years of monthly history for active and inactive securities. The full dataset, delivered in a single Excel workbook, contains the time series and event history data for which CRSP is known.

Two important factors to predict a firm's future return is the firm's book value and earning per share. Those monthly values as well as CUSIP (to merge the data with CRSP dataset) are obtained using the Compustat dataset. Compustat is a database of financial, statistical and market information on active and inactive global companies throughout the world. It includes several databases, namely, monthly and daily pricing data, standard Poor's and other leading Index Data, qualitative content including business descriptions, officer information, and executive compensation, etc .

To be able to calculate the positive/negative sentiment of the changes in filings, we utilized sentiment category identifiers from Loughran and McDonald Master Dictionary. It's an Excel file containing each of the LM sentiment words by category– Negative, Positive, Uncertainty, Litigious, Strong Modal, Weak Modal, Constraining.

Quarter-on-quarter similarities between 10-Q and 10-K filings are measured using four different similarity measures taken from the literature in linguistics, textual similarity, and natural-language processing (NLP): i.) cosine similarity, ii.) Jaccard similarity, iii.) minimum edit distance and iv.) simple similarity. We describe each measure and its respective calculation in Feature Selection section.

Lastly, after I provided these data to the statistics and financial engineering group, they utilized Fama-MacBeth regression to train the regressor based on data provided. Mechanism section introduces Fama-MacBeth regression, and in the Result section, we will explain basic terms and statistical concepts to analyze the results of the Lazy Prices paper.

Chapter 3

Data

In this section, we will describe each dataset, and the process to gather them altogether.

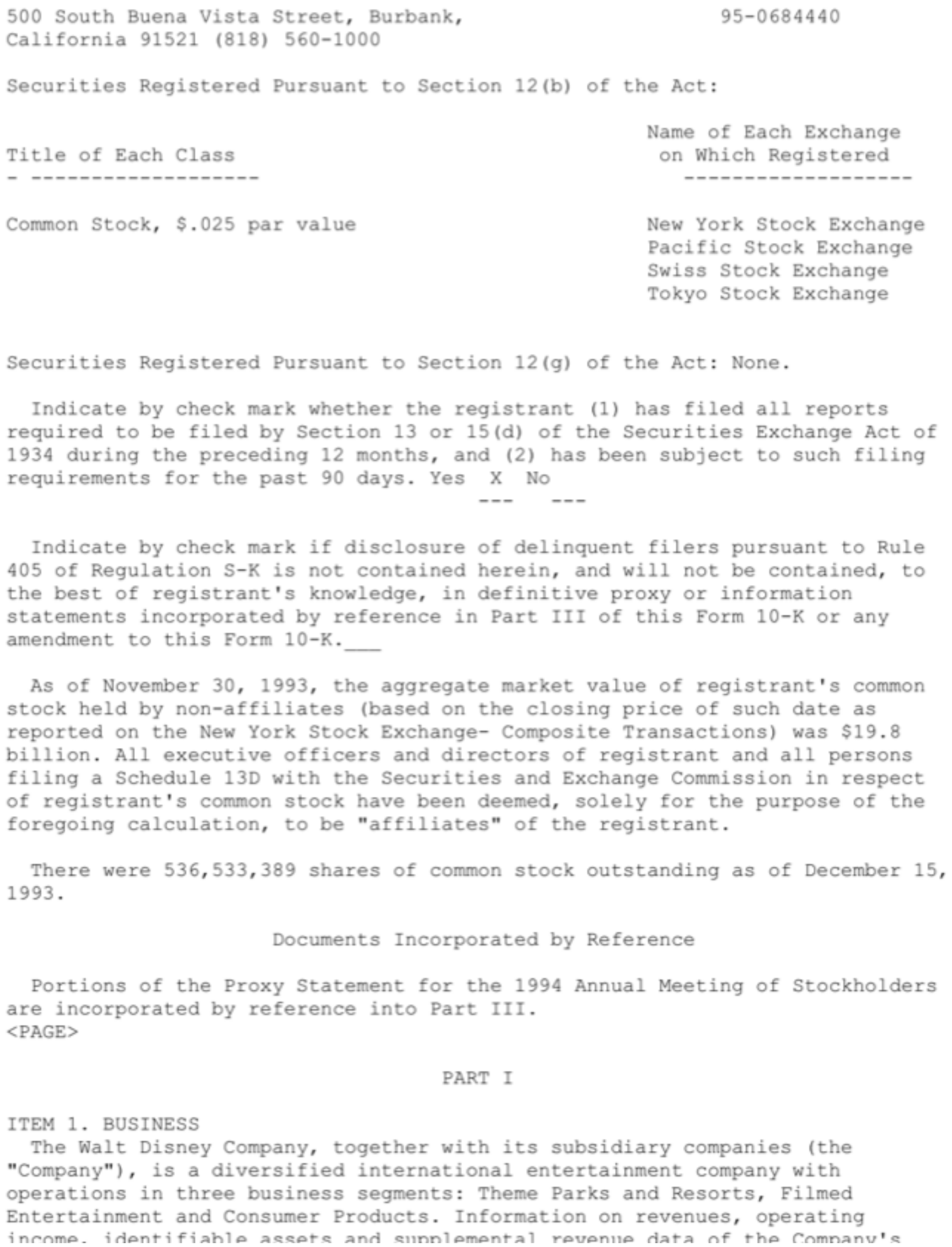
3.1 EDGAR Dataset

All complete 10-K, 10-K405, 10-KSB and 10-Q filings from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) from 1993 to 2018 are downloaded. The first challenge, was that some filings where in HTML format, and some of them in a simple text format. Therefore, extracting pure text was so challenging. Besides, we had to extract each part (item) separately. Figure 3.1 and Figure 3.2 shows the data.

Items that can be in each filing, are given in table 3.1. Tricky part is that in each part, there may be mentions of the other parts, so scanning text and separating text by keyword "Part" doesn't work.

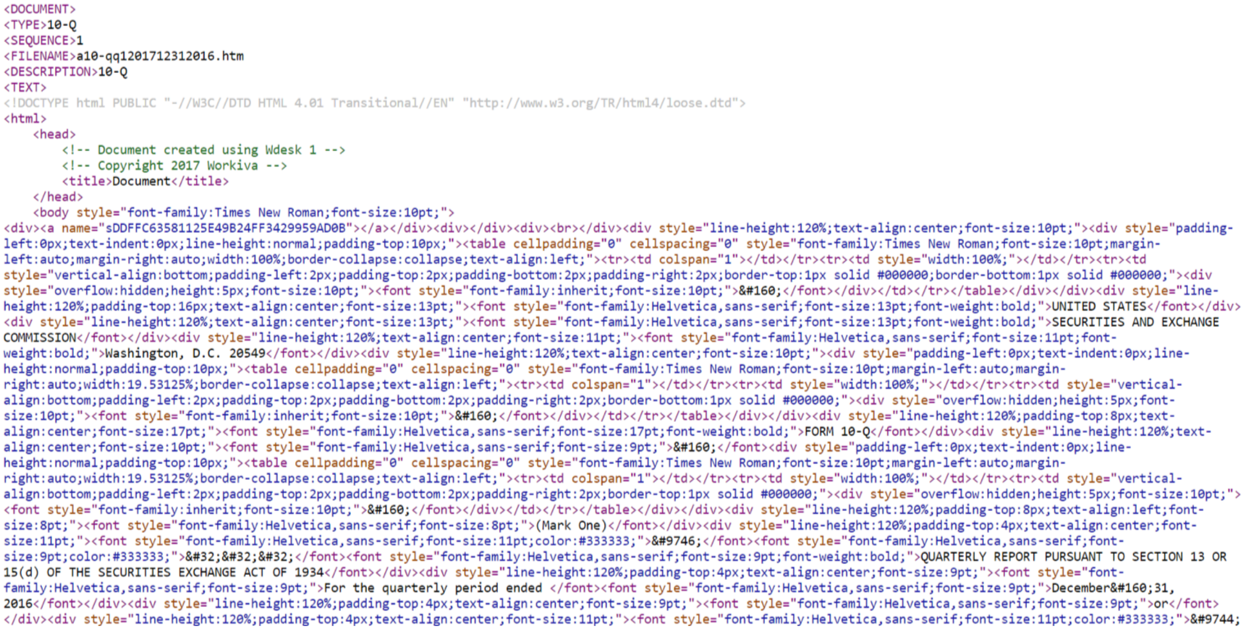
3.2 CRSP Dataset

In order to obtain stock return for training, we downloaded monthly stock returns from the Center for Research in Security Prices (CRSP). These data provide a key map from EDGAR



```
500 South Buena Vista Street, Burbank,                     95-0684440
California 91521 (818) 560-1000

Securities Registered Pursuant to Section 12(b) of the Act:

                                                     Name of Each Exchange
Title of Each Class                                   on Which Registered
- -------------------                                     ------------------

Common Stock, $.025 par value                        New York Stock Exchange
                                                     Pacific Stock Exchange
                                                     Swiss Stock Exchange
                                                     Tokyo Stock Exchange


Securities Registered Pursuant to Section 12(g) of the Act: None.

  Indicate by check mark whether the registrant (1) has filed all reports
required to be filed by Section 13 or 15(d) of the Securities Exchange Act of
1934 during the preceding 12 months, and (2) has been subject to such filing
requirements for the past 90 days. Yes  X  No
                                       ---   ---

  Indicate by check mark if disclosure of delinquent filers pursuant to Rule
405 of Regulation S-K is not contained herein, and will not be contained, to
the best of registrant's knowledge, in definitive proxy or information
statements incorporated by reference in Part III of this Form 10-K or any
amendment to this Form 10-K.___

  As of November 30, 1993, the aggregate market value of registrant's common
stock held by non-affiliates (based on the closing price of such date as
reported on the New York Stock Exchange- Composite Transactions) was $19.8
billion. All executive officers and directors of registrant and all persons
filing a Schedule 13D with the Securities and Exchange Commission in respect
of registrant's common stock have been deemed, solely for the purpose of the
foregoing calculation, to be "affiliates" of the registrant.

  There were 536,533,389 shares of common stock outstanding as of December 15,
1993.

                      Documents Incorporated by Reference

  Portions of the Proxy Statement for the 1994 Annual Meeting of Stockholders
are incorporated by reference into Part III.
<PAGE>

                                    PART I

ITEM 1. BUSINESS
  The Walt Disney Company, together with its subsidiary companies (the
"Company"), is a diversified international entertainment company with
operations in three business segments: Theme Parks and Resorts, Filmed
Entertainment and Consumer Products. Information on revenues, operating
income, identifiable assets and supplemental revenue data of the Company's
```

Figure 3.1: TEXT formatted filings



```
<DOCUMENT>
<TYPE>10-Q
<SEQUENCE>1
<FILENAME>a10-qq1201712312016.htm
<DESCRIPTION>10-Q
<TEXT>
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
    <head>
        <!-- Document created using Wdesk 1 -->
        <!-- Copyright 2017 Workiva -->
        <title>Document</title>
    </head>
    <body style="font-family:Times New Roman;font-size:10pt;">
<div><a name="sDDFFC63581125E49B24FF3429959AD0B"></a></div><div></div><div><br></div><div style="line-height:120%;text-align:center;font-size:10pt;"><div style="padding-
left:0px;text-indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;margin-
left:auto;margin-right:auto;width:100%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td
style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;border-top:1px solid #000000;border-bottom:1px solid #000000;"><div
style="overflow:hidden;height:5px;font-size:10pt;"><font style="font-family:inherit;font-size:10pt;">&#160;</font></div></td></tr></table></div></div><div style="line-
height:120%;padding-top:16px;text-align:center;font-size:13pt;"><font style="font-family:Helvetica,sans-serif;font-size:13pt;font-weight:bold;">UNITED STATES</font></div>
<div style="line-height:120%;text-align:center;font-size:13pt;"><font style="font-family:Helvetica,sans-serif;font-size:13pt;font-weight:bold;">SECURITIES AND EXCHANGE
COMMISSION</font></div><div style="line-height:120%;text-align:center;font-size:11pt;"><font style="font-family:Helvetica,sans-serif;font-size:11pt;font-
weight:bold;">Washington, D.C. 20549</font></div><div style="line-height:120%;text-align:center;font-size:10pt;"><div style="padding-left:0px;text-indent:0px;line-
height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;margin-left:auto;margin-
right:auto;width:19.53125%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-
align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;border-bottom:1px solid #000000;"><div style="overflow:hidden;height:5px;font-
size:10pt;"><font style="font-family:inherit;font-size:10pt;">&#160;</font></div></td></tr></table></div></div><div style="line-height:120%;padding-top:8px;text-
align:center;font-size:17pt;"><font style="font-family:Helvetica,sans-serif;font-size:17pt;font-weight:bold;">FORM 10-Q</font></div><div style="line-height:120%;text-
align:center;font-size:10pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;">&#160;</font><div style="padding-left:0px;text-indent:0px;line-
height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;margin-left:auto;margin-
right:auto;width:19.53125%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-
align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;border-top:1px solid #000000;"><div style="overflow:hidden;height:5px;font-size:10pt;">
<font style="font-family:inherit;font-size:10pt;">&#160;</font></div></td></tr></table></div></div><div style="line-height:120%;padding-top:8px;text-align:left;font-
size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">(Mark One)</font></div><div style="line-height:120%;padding-top:4px;text-align:center;font-
size:11pt;"><font style="font-family:Helvetica,sans-serif;font-size:11pt;color:#333333;">&#9746;</font><font style="font-family:Helvetica,sans-serif;font-
size:9pt;color:#333333;">&#32;&#32;&#32;</font><font style="font-family:Helvetica,sans-serif;font-size:9pt;font-weight:bold;">QUARTERLY REPORT PURSUANT TO SECTION 13 OR
15(d) OF THE SECURITIES EXCHANGE ACT OF 1934</font></div><div style="line-height:120%;padding-top:4px;text-align:center;font-size:9pt;"><font style="font-
family:Helvetica,sans-serif;font-size:9pt;">For the quarterly period ended </font><font style="font-family:Helvetica,sans-serif;font-size:9pt;">December&#160;31,
2016</font></div><div style="line-height:120%;padding-top:4px;text-align:center;font-size:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;">or</font>
</div><div style="line-height:120%;padding-top:4px;text-align:center;font-size:11pt;"><font style="font-family:Helvetica,sans-serif;font-size:11pt;color:#333333;">&#9744;
```

Figure 3.2: HTML formatted filings

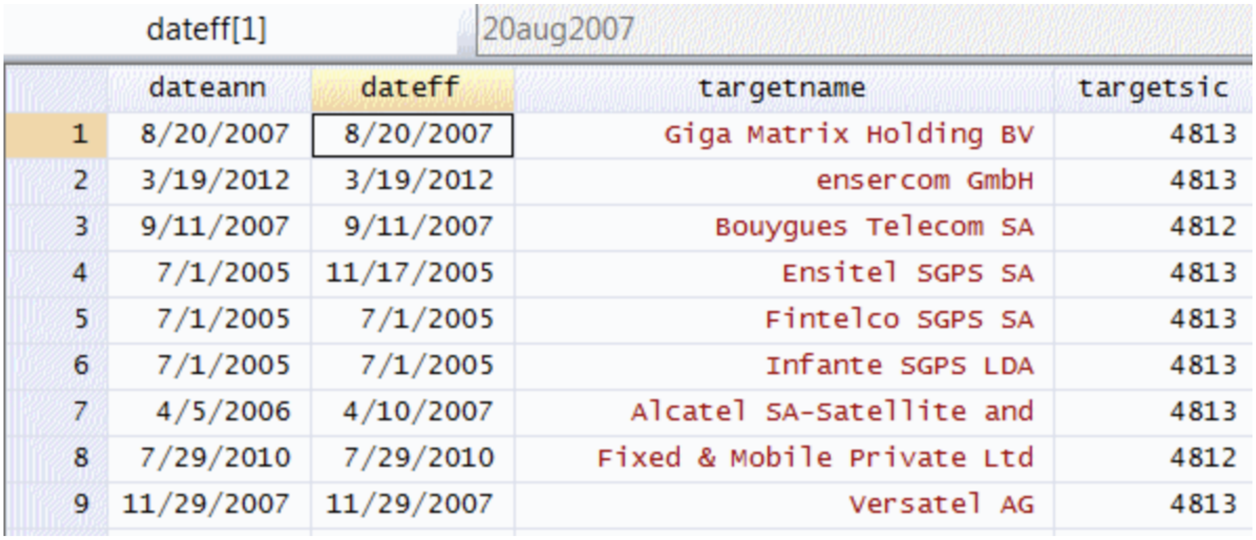
and Compustat as well as each firm's monthly return. To access these information we had to code an SQL script. Figure 3.1 demonstrate an example of this table.

3.3 Compustat Dataset

These dataset provides firms' book value of equity and earnings per share. These data will be used as a feature in the regressor. These data were available in pandas table, and needed to be merged with the data mentioned above.

3.4 Sentiment Category

Sentiment category identifiers from Loughran and McDonald (2016)'s Master dictionary [12] is to label each word in the filings with one of these sentiments: Negative, Positive,



dateff[1] 20aug2007

	dateann	dateff	targetname	targetsic
1	8/20/2007	8/20/2007	Giga Matrix Holding BV	4813
2	3/19/2012	3/19/2012	ensercom GmbH	4813
3	9/11/2007	9/11/2007	Bouygues Telecom SA	4812
4	7/1/2005	11/17/2005	Ensitel SGPS SA	4813
5	7/1/2005	7/1/2005	Fintelco SGPS SA	4813
6	7/1/2005	7/1/2005	Infante SGPS LDA	4813
7	4/5/2006	4/10/2007	Alcatel SA-Satellite and	4813
8	7/29/2010	7/29/2010	Fixed & Mobile Private Ltd	4812
9	11/29/2007	11/29/2007	Versatel AG	4813

Table 3.1: CRSP Table

Uncertainty, Litigious, Modal, Constraining. Then, for each two consecutive filings,we can find the sentiment of change (words that are deleted/added to next filing).

Chapter 4

Striping 10-K/Q Down to Text Files

All 10-K and 10-Q SEC complete text document filings are downloaded for each year and quarter. The text and html version of SEC filings is an integration of all information provided in the browser-friendly files. These types are listed in [7]. For example, IBM's 10-K filing on 02/28/2012, except from the main HTML formatted file, has four jpg (graphics) files, ten exhibits, and six XBRL files; which are also included in a single text file with the embedded HTML, the ASCII-encoded graphics, XBRL, exhibits.

In the mentioned filing, about %7.5 of the characters contain useful information, consisting of tables and exhibits. %55 of the filing is the HTML coding and %33 is only for XBLR tagged characters. The remaining 27% of the file is attributable to the ASCII encoded graphics. In this case, from about 48,250,000 characters, about only 3,618,750 are useful. Given the fact that in most of the filing, there are so many other ASCII encoded files, such as pdf, xls and all other files that can be encoded to binary, which can account for more than 90% of the document.

Since doing textual analysis and language processing demands textual content of a document, there is a need to strip the filing from all unnecessary data. This implies that we need to exclude markup tags, ASCII-encoded files, and tables which has more than %20 numerical characters over all tokens ratio. These kind of data might be useful for other purposes; however, in this work the stripped filings help to perform textual analysis in a more efficient way. The stripping procedure has been done in two stages as in [13]: removing HTML markup tags and parsing data to extract useful texts.

4.1 EDGAR Markup Tags

All of the original markup language tags (HTML, XBRL, XML) are removed from the downloaded documents. Heading of each filing includes submission date, information about company code (CIK), name of the company, address, fiscal year end, etc. All of these heading information are removed and saved as dictionary in a Json file. All of the exhibit tags are removed and textual data of the exhibits are added to the dictionary and removed fro filing.

After cleaning the HTML code from these special markup tags, and encapsulating exhibits, data should be extracted from each filing to obtain raw3 text file along with Json file for each filing to be able to perform natural language processing on textual data. Following parsing methods is used to create needed data for each filing downloaded from EDGAR.

4.2 Parsing

1. GRAPHICS, EXCEL, PDF, ZIP, and JSON ASCII encoded entities are removed. These documents are inside <TYPE> tags. The reason for including these ASCII encoded files is to make it possible to transfer across various hardware platforms more

effectively. After removing this type of files, the size of the filing will be so much less than the original one, since even a small image file, has a very large ASCII encoded part.

2. XMLs are removed. These were tagged inside <XML>.

3. XBRL are removed. These were tagged inside <XBRL>.

4. <DIV>, <TR>, <TD>, and <FONT> tags are removed. For the sake of efficiency, even though some of these information may be helpful, we removed them from the filing to get raw texts.

5. &AMP and &#38 are replaced with "&" and &NBSP and &#160 are replaced with a blank space. .

6. All other HTML numeric codes from **??** are replaced with its character.

7. Remove all remaining extended character references ([7] section 5.2.2.6.)

8. All SEC headers and footers are removed. All characters from the beginning of the original file inside </SEC-HEADER> or </IMS-HEADER> tags are deleted from the file. As discussed in 4.1, useful information in the heading are saved in Json file along with raw text. Besides, the footer "—–END PRIVACY-ENHANCED MESSAGE—–" at the end of each filing is deleted.

9. Since Item 7 and Item 8 are of the most importance inside EDGAR filings, there is a need to make sure they are not removed by mistake. In some cases, Item 7 or Item 8 of the filings starts with a table where the Item 7 or 8 are written within the table. Thus, any table containing Item 7 or Item 8 is not deleted.

10. Some of the tables are removed: for this parsing, only table strings where

    $$numeric\,chars \;/\; (alphabetic \;+\; numericchars) > \%20$$

    are removed.

11. Exhibits are removed from the raw text as stated in EDGAR Markup Tags.

12. All the remaining markup tags are removed(i.e., $<\ldots>$).

Chapter 5

Feature Selection

In this section, I'll describe each feature and how I implemented each of them. Firstly, each filling is converted to a dictionary, in which for each word we have the position of the word in document, number of occurrence and sentiment category.

5.1 Cosine Similarity

Cosine similarity between document D_1 and document D_2 is computed as follows. Let set of terms occurring in D_1 and D_2 be D_{s1} and D_{s2}, respectively. Union of D_{s1} and D_{s2}, T, contains all the term included in both documents. Let t^i be the i^{th} element of T . Term frequency vectors of D_1 and D_2 are defined as follows:

$$D^{TF}{}_1 = [nD_1(t_1), nD_1(t_2), ...] \tag{5.1}$$

$$D^{TF}{}_2 = [nD_2(t_1), nD_2(t_2), ...] \tag{5.2}$$

Number of times term t_j is repeated in D_i is denoted as $nD_i(t_j)$. Dot product of these vectors divided by the product of Euclidean norm of both documents are computed in order to obtain cosine similarly.

$$Cos\_sim \quad = \quad \frac{D^{TF}{}_1 \,.\, D^{TF}{}_2}{||D^{TF}{}_1|| \qquad ||D^{TF}{}_2||} \tag{5.3}$$

The more similar 2 filings are, the more closer cosine similarity is to 1. Cosine similarity equal to 0, implies that the documents has no words in common.

5.2 Jaccard Similarity

Using the same term-frequency vectors as in cosine similarity, Jaccard similarity can be obtained using following expression:

$$Jac\_sim \quad = \quad \frac{|D^{TF}{}_1 \ \cap \ D^{TF}{}_2|}{|D^{TF}{}_1 \ \cup \ D^{TF}{}_2|} \tag{5.4}$$

Formula in 5.4, implies that the Jaccard similarity is the size of the intersection divided by the size of the union of the two term frequency sets. Despite the fact that cosine similarity measure includes number of times a word has stated in each document, The term-frequency vectors used in 5.4, is binary, meaning that each word is counted only once in a given set.

5.3 Minimum Edit Distance Similarity

Another similarity measure utilized in this work is Minimum Edit Distance, MinEdit_sim, which is computed by counting the smallest number of operations required to transform one document into the other one. This measure takes a long time, since average words per document is around 13k. I used dynamic programming approach to find this measure.

More specifically, minimum edit distance can be computed by creating a matrix of size $(|D^{TF}{}_1| + 1) \times (|D^{TF}{}_2| + 1)$, and each element can be found using following expression :

$$d_{i0} = \sum_{k=1}^{i} w_{del}(b_k), \qquad for \qquad 0 \leq i \leq m \tag{5.5}$$

$$d_{0j} = \sum_{k=1}^{j} w_{ins}(a_k), \qquad for \qquad 1 \leq j \leq n \tag{5.6}$$

$$d_{i,j} = \begin{cases} d_{i-1,j-1} & for \qquad a_j = b_i \\ min \begin{cases} d_{i-1,j} + w_{del}(b_i) \\ d_{i,j-1} + w_{ins}(a_j) \\ d_{i-1,j-1} + w_{sub}(a_j, b_i) \end{cases} & for \qquad a_j \neq b_i \end{cases} \tag{5.7}$$

In equations 5.5 to 5.7, m and n denote $|D^{TF}{}_1|$ and $|D^{TF}{}_2|$ respectively. Also, $d_{i,j}$ is element in row i and column j of the matrix. $w_{del}, w_{ins}, w_{sub}$ is the weight of deleting, inserting or substituting a character (in our case, a word). Equations 5.5 (and 5.6) fill the row 0 (and column 0) with numbers between 0 and m (n). Equation 5.7, fills the rest of the matrix, which finally calculates and insert the minimum edit distance in the bottom right element ($d_{n,m}$).

5.4 Simple Similarity

The last similarity measure is called Simple_sim, and uses a simple side-by-side comparison method. This method is same as word track changes option. I used the python difflib library to compare the old document D_1 with the new document D_2. The number of the changes, additions, and deletions are counted and then normalized t by the average size of the old document and the new document.

$$c = \frac{addition + deletion + changes}{\frac{SizeD_1 + SizeD_2}{2}} \tag{5.8}$$

5.5 Sentiment of Change

This feature is not accounted as a similarity measure, even though it is basically a similarity measure in the sentiment of two documents. Loughran and McDonald Master Dictionary is a sentiment category identifier and word list, which assign sentiment to thousands of word common in finance field, e.g. negative, positive, uncertain, etc. To obtain the sentiment of change, number of negative words is subtracted from number of positive words in the set of uncommon terms in ${D^{TF}}_1$ and ${D^{TF}}_2$. It is then normalized by the size of the change $(|({D^{TF}}_1 \cup {D^{TF}}_2) - ({D^{TF}}_1 \cap {D^{TF}}_2)|)$. Same procedure has been done on the litigious and uncertain words.

5.6 Change of CEO/CFO

Similar to sentiment of change, this feature is not accounted as a similarity measure in main results. This feature is a binary feature which is 1 if there is a turnover from CEO to CFO, and 0 otherwise.

Chapter 6

Mechanism

In order to predict future firms' return, Fama-MacBeth cross-sectional regression has been done on 4 similarity measures and some other useful indicators which we will introduce them in result section. In this section, a description of Fama-MacBeth regression is given.

6.1 Fama-MacBeth Regressions

Theories of asset pricing frequently need to test different factors to be able to predict asset returns. These factors can range from macroeconomic (for example, consumer inflation or the unemployment rate) to financial (firm size, etc). The Fama-MacBeth two-step regression is a practical way of testing how these factors describe portfolio or asset returns [5].

In the first step, each portfolio's return is regressed against time series of factors to determine the the factor exposures. In the second step, for each time step, portfolio returns are regressed against the factor exposures, to give a time series of risk coefficients for each factor. Then the average of these coefficients are computed, once for each factor, to find the expected value for a unit exposure to each risk factor over time.

In equation form, for n portfolio or asset returns and m factors, the first step is to obtain the factor exposures, βs, by calculating n regressions, each one on m factors:

$$\begin{aligned} R_{1,t} &= \alpha_1 + \beta_{1,F_1} F_{1,t} + ... + \beta_{1,F_m} F_{m,t} + \epsilon_{1,t}, \\ &. \\ &. \\ &. \\ R_{n,t} &= \alpha_n + \beta_{n,F_1} F_{1,t} + ... + \beta_{n,F_m} F_{m,t} + \epsilon_{n,t}. \end{aligned} \tag{6.1}$$

In above equations, $R_{i,t}$ is the return of portfolio i at time t, $F_{j,t}$ is the factor j at time t, β_{i,F_j} are the factor exposures, which describe how returns are exposed to the factors. Since the purpose is to obtain the exposure of the return of each portfolio or asset to a given set of factors (features), all of the regressions use the same factors.

The second step is to regress the returns on the m estimated βs, to find the exposure of n returns to m factor exposures. Therefore, we use the same βs for all of the regressions. This step basically gives an intuition about the randomness of calculated βs, meaning that whether the larger factor exposures mean higher return or not.

$$
\begin{aligned}
R_{i,1} &= \gamma_{1,0} + \gamma_{1,1}\beta_{1,F_1} + ... + \gamma_{1,m}\beta_{m,F_m} + \epsilon_{i,1},\\
&.\\
&.\\
&.\\
R_{i,T} &= \gamma_{T,0} + \gamma_{T,1}\beta_{1,F_1} + ... + \gamma_{T,m}\beta_{m,F_m} + \epsilon_{i,T}.
\end{aligned}
\tag{6.2}
$$

where the returns R are the same as those in 6.1, γs are regression coefficients that are later used to calculate the risk premium for each factor, and in each regression i goes from 1 through n (n portfolios or assets). After the second step is done, there will be $(m+1)*T$ matrix for γs. Assuming that ϵs are i.i.d, risk premium γ_m for factor F_m is obtained by averaging γs over time (T). Using these coefficients (βs and γs), we can analyse the results in the next section.

Chapter 7

Results

The next and last step is to run monthly Fama-MacBeth cross-sectional regressions on given features. In order to be able to analyze the results, a description of P-value, T-stat, and predictors is given.

7.1 Statistical Significance

One of the well-known terms in statistics is statistical significance. Statisticians do complex operations that yield a result which should be evaluated to make sure whether those results prove their point. A straight forward idea, which is built on a few simple ideas and is not a complex phenomenon, is statistical significance. The bases of this concept are hypothesis testing, the normal distribution, and p values. In this section, all of these concepts will be briefly discussed.

The first idea needs to discussed is hypothesis testing, a technique for evaluating a theory using data. The "hypothesis" refers to the researcher's initial belief about the situation before the study. This initial theory is known as the alternative hypothesis and the opposite is known as the null hypothesis. Hypothesis tests are one of the foundations of statistics and are used to assess the results of most studies.

The testing part of hypothesis tests allows us to determine which theory, the null or alternative, is better supported by the evidence. There are many hypothesis tests and we will use one called the z-test. The second building block of statistical significance is the normal distribution. Assuming the data is normal, we have a Gaussian distribution with parameters μ and σ.

Instead of characterizing any point in terms of standard deviations from the mean, in statistics we use z-score, which just represents the number of standard deviations a point is from the mean. If we do Z transformation to all the data points, the new distribution is called the standard normal.

$$Z \quad transformation : Z = \frac{X - \mu}{\sigma} \tag{7.1}$$

The higher or lower the z-score, the more unlikely the result is to happen by chance and the more likely the result is meaningful. To quantify just how meaningful the results are, we define p-value.

7.2 P Value and T Score

A p-value is the probability of observing results at least as extreme as those measured when the null hypothesis is true [9]. Therefore, the lower the p-value, the more meaningful the result, because it is less likely to be caused by chance or noise. The result can be considered as statistically significant if the p-value is less than from what we established for significance before we begin the experiment.

The choice of alpha (p-value) depends on the situation and the field of study, but the most commonly used value is 0.05, corresponding to a 5% chance the results occurred at random. Let us consider a z-score of 2.045. If the p-value is 0.0294, we can reject the null hypothesis. There is statistically significant evidence that the result we got is meaningful. The p-value shows there is a 2.05% chance that our results occurred because of random noise.

Now, since we explained the concepts for statistical significance in general, we should explain how we are gonna evaluate the result in this specific study. Usually in statistics, if we are not able to access the ground truth (in our case, FMB coefficients), T-test with a T Statistic is used instead of z score.

As mentioned in the Mechanism section, after training the model, there are m + 1 series γ (including the constant in the second step) for every factor, each of length T. Assuming i.i.d ϵ, γ_m is calculate for factor F_m by averaging the m_{th} γ over T, and also get standard deviations and t-stats. T-stats for the m_{th} risk premium are:

$$\frac{\gamma_m}{\sigma_{\gamma m}/\sqrt{T}} \tag{7.2}$$

Using this t-statistics and t-distribution of coefficients, we can find the p-values and figure out whether the results are statistically significant or not. Figure 7.1 illustrates the relation between p-value and t-stats. As you can see, if we set the alpha to %2.5 beforehand, and the p-value for selected t-stat is less than %2.5, then the null hypothesis is accepted (and true hypothesis get rejected).

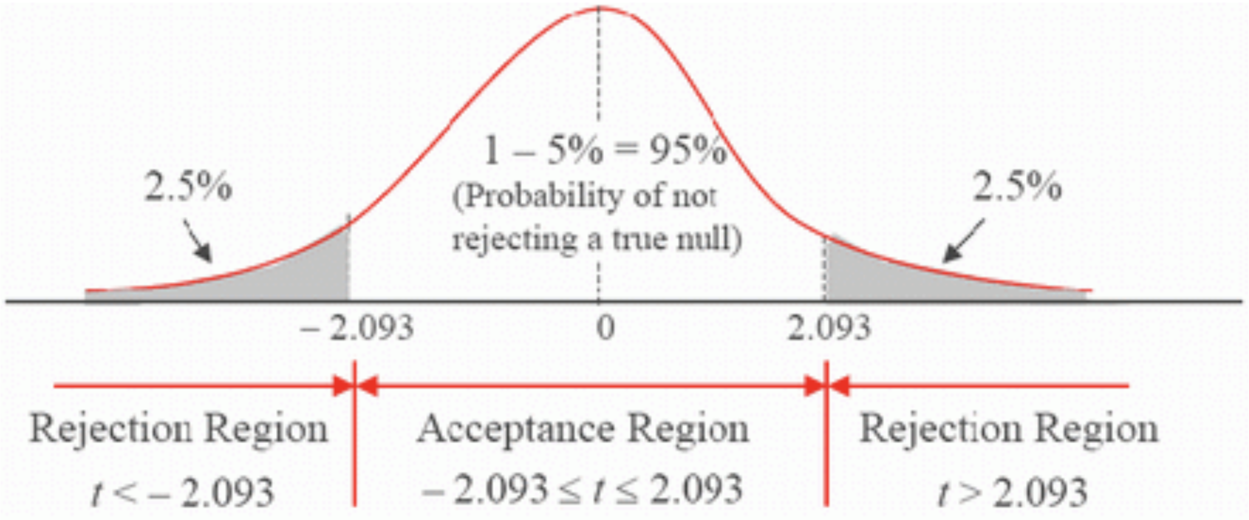


Figure 7.1: T statistics and P value[8]

With the provided descriptions, now we can statistically analyze the result from the paper. However, before doing so, we should briefly go through the well-known market indicator used in training.

7.3 Market Indicators

Market indicators are used in technical analysis to forecast market trends. Market indicators are ratios and formulas that explain current gains and losses in stocks and indexes, and furthermore, indicate if an index will experience short-term or long-term gains or losses.

In this work, some of the most important indicators are used to train the regressor. Here is the indicator name and its description :

- **Size:** This indicates the market value of the equity. Equity is the amount of money that would be returned to a company's shareholders if all of the assets were liquidated and all of the company's debt get paid off. Following illustrates the formula for the market value of equity:

$$Market \quad Value \quad = \quad Current \quad Stock \quad Price \quad \times \quad Total \quad Outstanding \quad Shares \tag{7.3}$$

- **Log(BM):** Book to Market (BM) value is the division of the book value to the market value of the equity. Book value can be found using the following formula:

$$Book \quad Value \quad = \quad Total \quad Assets \quad - \quad Total \quad Debt \tag{7.4}$$

- **SUE:** Standardized Unexpected Earnings (SUE) is the difference between the reported earnings and the expected earnings. The firm's expected earnings measurements include forecasts of the firm's profit and mathematical models of expected earnings based on the earnings of previous return periods. Investors tend to buy stocks with positive surprise and sell those with negative surprise.

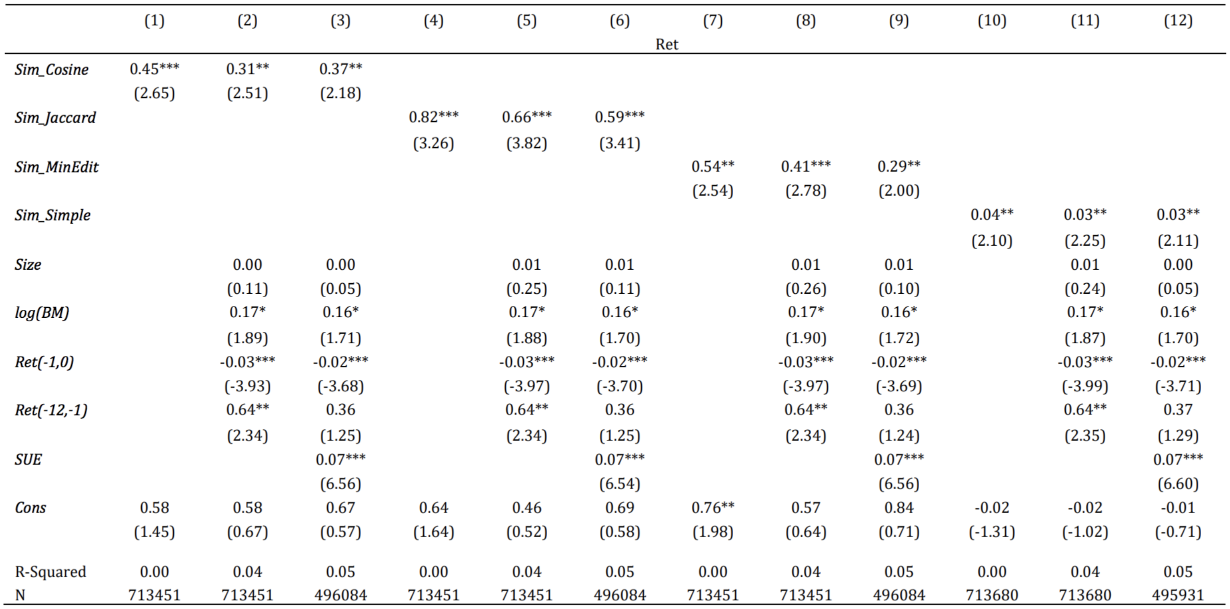
$$SUE_t \quad = \quad \frac{Q_t - E(Q_t)}{\sigma(Q_t - E(Q_t))} \tag{7.5}$$

  In equation 7.5, Q_t is the return in quarter t, and $E(Q_t)$ is the expected earnings.

  Large negative earnings surprises may have legal and reputational costs to firms. Shareholders can sue the firm for its fault in revealing negative earnings news promptly. On the other hand, by reason of the negative influence of withholding bad news on the managers' reputation, money managers may choose not to hold, and analysts may choose not to follow them.

- **Ret(i,j):** Also, the previous periods' returns can be a good indicator of future returns. In this paper, Ret(-1,0) and Ret(-12,-1) are used. They denote the previous month's return and the cumulative stock return from month t-12 to month t-2, respectively.

There are so many other return predictors that are out of the scope of this thesis.



	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Ret											
Sim_Cosine	0.45***	0.31**	0.37**									
	(2.65)	(2.51)	(2.18)									
Sim_Jaccard				0.82***	0.66***	0.59***						
				(3.26)	(3.82)	(3.41)						
Sim_MinEdit							0.54**	0.41***	0.29**			
							(2.54)	(2.78)	(2.00)			
Sim_Simple										0.04**	0.03**	0.03**
										(2.10)	(2.25)	(2.11)
Size		0.00	0.00		0.01	0.01		0.01	0.01		0.01	0.00
		(0.11)	(0.05)		(0.25)	(0.11)		(0.26)	(0.10)		(0.24)	(0.05)
log(BM)		0.17*	0.16*		0.17*	0.16*		0.17*	0.16*		0.17*	0.16*
		(1.89)	(1.71)		(1.88)	(1.70)		(1.90)	(1.72)		(1.87)	(1.70)
Ret(-1,0)		-0.03***	-0.02***		-0.03***	-0.02***		-0.03***	-0.02***		-0.03***	-0.02***
		(-3.93)	(-3.68)		(-3.97)	(-3.70)		(-3.97)	(-3.69)		(-3.99)	(-3.71)
Ret(-12,-1)		0.64**	0.36		0.64**	0.36		0.64**	0.36		0.64**	0.37
		(2.34)	(1.25)		(2.34)	(1.25)		(2.34)	(1.24)		(2.35)	(1.29)
SUE			0.07***			0.07***			0.07***			0.07***
			(6.56)			(6.54)			(6.56)			(6.60)
Cons	0.58	0.58	0.67	0.64	0.46	0.69	0.76**	0.57	0.84	-0.02	-0.02	-0.01
	(1.45)	(0.67)	(0.57)	(1.64)	(0.52)	(0.58)	(1.98)	(0.64)	(0.71)	(-1.31)	(-1.02)	(-0.71)
R-Squared	0.00	0.04	0.05	0.00	0.04	0.05	0.00	0.04	0.05	0.00	0.04	0.05
N	713451	713451	496084	713451	713451	496084	713451	713451	496084	713680	713680	495931

Table 7.1: Main Results [3]

7.4 Main Results

Table 7.1 illustrates the main results. Each element in the table has 3 information in it: the main number is the factor exposure (βs), the numbers in parenthesis are t-stats, and the asterisks show the statistical significance. *, **, *** denote the p-value less than %10 , %5 and %1 respectively.

Two points are necessary to indicate before analyzing the results: firstly, in finance and especially in stock return predictions, R-squared is very small (in this table the maximum R-squared is %0.05). Therefore, the other measures should be considered, such as t-stat, factor exposures, statistical significance, etc. Secondly, all of the selected features are tested separately, to be able to perceive their real effect on the results.

Columns 1-3, show the results of using cosine similarity feature with different host of indicators. Without adding any other feature, it has statistical significance of less than %1.

After adding other indicators (columns 2 and 3) the statistical significance decreased to %5; however, the R-Squared has increased, meaning that the data are closer to the fitted regression line.

Columns 4-6 only consider the Jaccard similarity measure amongst the 4 similarity measures. It indicates that not only the βs are increased, but also the statistical significance remains %1 while adding other indicators. Greater βs (comparing with other similarity measures) implies that for one-standard deviation decline/incline in a stock's document Jaccard similarity, returns are much more affected than the others.

The rest of the columns are considering minimum edit distance and simple similarity measures. The statistical significance for both of them is lower than the rest of the similarly measures. Considering the computational complexity of the minimum edit distance and . also very small βs of the simple similarity measure, I suggest not to waste resources to calculate them.

The market value of the firms does not have much effect in this case. Also, as provided in the table, last month's return is a very significant indicator even in the lack of other signals. It implies that this measure is the most reliable indicator of future return amongst others considered in this table. The small factor exposure for this signal, on the other hand, means that even big changes in last month return do not have a large impact for future return, meaning that low tolerate risk-takers could be safe using this signal.

Earning surprise, like last month's return, has high statistical significance, and it's almost constant for all considered similarity measures. As stated before, this measure is a reliable measure to buy or sell stocks based on that, however, there is always the risk of lack of honesty in disclosing news or error in computation. Considering its larger factor exposures and statistical significance in comparison to the other market predictor used here, it is less safe for low-risk tolerant investors to consider this kind of stocks based on earning surprise only.

All in all, comparing similarity measures together, Jaccard similarity seems to be the best indicator for future return. It basically measures the number of common words used in both documents, divided by the size of the bag of words. Intuitively talking, if the policy of the company to write the report or the reporter has been changed, then the words used in the documents will be changed more than what it was previously. The size of the bag of words will be larger and the size of the common words set will be lowered.

Last, but not least, based on the fact that all of the factor exposures for the similarity measures are positive, and also considering the fact that they are standardized to be from 0 to 1, starting from no similarity to highest similarity (basically comparing two documents which are the same), non-changers will have more positive return than the changers.

Chapter 8

Conclusion

In this thesis we used textual analysis introduced in [3] on annual and quarterly filings submitted by each company to Electronic Data Gathering, Analysis, and Retrieval system on SEC website, in order to predict firms' future returns. These techniques include designing several similarity measures to calculate the resemblance between two consecutive filings. Using these similarity measures and several well-known market indicators, yields to factor exposures which are statistically significant. In order to find factor exposures, Fama-MacBeth regressor is trained on the mentioned features. Results imply that the Jaccard similarity measure has the highest p-value amongst other features. Besides, earning surprise is a highly valuable signal about future returns, as it shows the difference between the expected return and real return. In case it is positive, it implies that the reported earning is greater than what expected. Finally, considering the sign of factor exposures of similarity measures, it can be perceived that non-changers (similarity measures closer to 1) have greater returns than those who have more changes (similarity measures closer to 0).

Bibliography

[1] Azi Ben-Rephael, Zhi Da, and Ryan D Israelsen. "It depends on where you search: Institutional investor attention and underreaction to news". In: *The Review of Financial Studies* 30.9 (2017), pp. 3009–3047.

[2] Stephen V Brown and Jennifer Wu Tucker. "Large-sample evidence on firms' year-over-year MD&A modifications". In: *Journal of Accounting Research* 49.2 (2011), pp. 309–346.

[3] Lauren Cohen, Christopher Malloy, and Quoc Nguyen. *Lazy prices*. Tech. rep. National Bureau of Economic Research, 2018.

[4] Zhi Da, Joseph Engelberg, and Pengjie Gao. "In search of fundamentals". In: *AFA 2012 Chicago Meetings Paper*. 2011.

[5] IHS EViews. *Fama-MacBeth Two-Step Regression*. 2014. URL: `http://didattica.unibocconi.it/mypage/dwload.php?nomefile=fama-macbeth20141115121157.pdf`.

[6] Ronen Feldman et al. "Management's tone change, post earnings announcement drift and accruals". In: *Review of Accounting Studies* 15.4 (2010), pp. 915–953.

[7] EDGAR Filing. "Filer Manual–Volume II". In: (2008).

[8] *Hypothesis Tests Concerning the Mean: T-test vs. Z-test*. URL: `https://analystnotes.com/cfa-study-notes-hypothesis-tests-concerning-the-mean-t-test-versus-z-test.html`.

[9] IWill Koehrsen. *Statistical Significance Explained*. 2018. URL: `https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687`.

[10] Feng Li. “Annual report readability, current earnings, and earnings persistence”. In: *Journal of Accounting and economics* 45.2-3 (2008), pp. 221–247.

[11] MultiMedia LLC. *lectronic Data Gathering, Analysis, and Retrieval*. URL: `https://www.sec.gov/edgar.shtml` (visited on 2018).

[12] Tim Loughran and Bill McDonald. “Textual analysis in accounting and finance: A survey”. In: *Journal of Accounting Research* 54.4 (2016), pp. 1187–1230.

[13] McDonald. *Stage One 10-X Parse Data*. URL: `https://sraf.nd.edu/data/stage-one-10-x-parse-data/` (visited on 2019).

[14] Paul C Tetlock. “All the news that’s fit to reprint: Do investors react to stale information?” In: *The Review of Financial Studies* 24.5 (2011), pp. 1481–1512.