A comparative study of regression analysis with and without search query data as a representation of public opinion

MATTIAS LARSSON

DAN KIRICHENKO
A comparative study of regression analysis with and without search query data as a representation of public opinion

En undersökande studie av regressions analys med och utan sökdata som en representation av folkopinion

Dan Kirichenko and Mattias Larsson

Degree Project in Computer Science DD143X
Supervisor: Alexander Kozlov
Examiner: Örjan Ekeberg

CSC, KTH 2016-05-11
Abstract
Stock prediction models using search query data is a modern phenomena and a relatively unexplored subject which potentially yields improvements to currently established prediction algorithms. This thesis will strive to improve an autoregressive prediction model by analyzing concurrent search query data to conclude whether or not taking such data into account will improve the prediction model. Multiple alternatives for sources of search query data has been analyzed and Google Trends was concluded as the most suitable candidate. The thesis found no strong indicator that amplifying the autoregressive algorithm with Google Trends data would produce better stock predictions. There remains to be found an elegant solution to improving prediction models using Google Trends data.

Referat
Contents
List of Figures .............................................. 4
List of Tables ............................................. 4
Acronyms .................................................... 5
1 Introduction ............................................. 6
   1.1 Problem statement .................................... 6
   1.2 Scope of study ....................................... 7
   1.3 Disposition of the report ............................ 7
2 Background .............................................. 8
   2.1 The stock market and prediction methods .......... 8
   2.2 Data mining .......................................... 8
      2.2.1 Google Trends .................................. 8
      2.2.2 Programming language - R .................... 9
   2.3 Statistics ............................................ 10
      2.3.1 Regression Analysis ........................... 10
3 Method .................................................. 11
   3.1 Literature study .................................... 11
   3.2 Choice of method ................................... 11
   3.3 Data collection ..................................... 12
      3.3.1 Google Trends and R .......................... 12
      3.3.2 Yahoo finance ................................. 12
   3.3 Manipulating collected data ........................ 12
4 Results ................................................ 13
   4.1 Base data ........................................... 13
   4.2 Regression models .................................. 13
      4.2.1 The Coca-cola company ........................ 17
      4.2.2 British Petroleum ................................ 17
      4.2.3 Wal-Mart Stores Inc ........................... 18
   4.3 Mean absolute error ................................ 19
5 Discussion ............................................. 20
   5.1 Discussion of results ............................... 20
   5.2 Discussion of method .............................. 21
6 Conclusion ............................................. 22
7 Bibliography .......................................... 23
List of Figures

2.1: Model of the map of a google trends search for “adwords”, “social media” and “seo”. Source https://www.hallaminternet.com/2014/google-trends-introduction-business/ 9
4.1: Graph of the three listed companies stock values closing prices 2010-07-24 to 2015-07-24 14
4.2: Graph of the three listed companies stock values closing prices 2010-07-24 to 2015-07-24 14
4.3: Graph of relative search amount of queries related to the Coca-Cola company. The y axis is an index of the total search hits as explained in the Google Trends section in the background. 15
4.4: Graph of relative search amount of queries related to the British Petroleum company. The y axis is an index of the total search hits as explained in the Google Trends section in the background. 16
4.5: Graph of relative search amount of queries related to the Walmart company. The y axis is an index of the total search hits as explained in the Google Trends section in the background. 16
4.6: Graph showing the mean error with and without Google Trends for the different intervals for the three companies that has been tested on the two models. 20

List of Tables

4.1: Table showing the linear regression without Google Trends on a 10 year interval for The Coca-Cola Company and the Coefficients are a, b and c in Model 1. 17
4.2: Table showing the linear regression with Google Trends on a 10 year interval for The Coca-Cola Company and the Coefficients are a, b, c and d in Model 2. 17
4.3: Table showing the linear regression without Google Trends on a 5 year interval for The Coca-Cola Company and the Coefficients are a, b and c in Model 1. 17
4.4: Table showing the linear regression with Google Trends on a 5 year interval for The Coca-Cola Company and the Coefficients are a, b, c and d in Model 2. 17
4.5: Table showing the linear regression without Google Trends on a 10 year interval for British Petroleum and the Coefficients are a, b and c in Model 1. 18
4.6: Table showing the linear regression with Google Trends on a 10 year interval for British Petroleum and the Coefficients are a, b, c and d in Model 2. 18
4.7: Table showing the linear regression without Google Trends on a 5 year interval for British Petroleum and the Coefficients are a, b and c in Model 1. 18
4.8: Table showing the linear regression with Google Trends on a 5 year interval for British Petroleum and the Coefficients are a, b, c and d in Model 2. 18
4.9: Table showing the linear regression without Google Trends on a 10 year interval for Wal-Mart Stores Inc and the Coefficients are a, b and c in Model 1. 19
4.10: Table showing the linear regression with Google Trends on a 10 year interval for Wal-Mart Stores Inc and the Coefficients are a, b, c and d in Model 2. 19
4.11: Table showing the linear regression without Google Trends on a 5 year interval for Wal-Mart Stores Inc and the Coefficients are a, b and c in Model 1. 19
4.12: Table showing the linear regression with Google Trends on a 5 year interval for Wal-Mart Stores Inc and the Coefficients are a, b, c and d in Model 2. 19
4.13: Table showing the mean error with and without Google Trends for the different intervals for the three companies that has been tested on the two models. 20
Acronyms
AI Artificial Intelligence
ANN Artificial Neural Network
API Application Program Interface
AR Autoregressive
BP British petroleum
IMRAD Introduction, Method, Results and Discussion
MSE Mean Squared Error
URL Uniform Resource Locator

Definitions
CSV Comma-separated values is a common file extension to save data of different sorts.
R A programming language established by the R Core Team
1. Introduction

The worldwide stock market is the collection of the buyers and sellers of shares from publicly traded corporations. Its history is deeply rooted in the development of organized trade in post-medieval Europe (Braudel, 1983). Trading stocks has historically been achieved at the stock exchange, a century old establishment which has garnered much research throughout its history. With the development of the modern western society, the stock market has transformed from the occupation of a few men to encompass the life savings of millions. In turn, the fluctuations of the market is not dictated by the few, but by the many. Additionally, in our current digital world, the transactions which change the value of a stock price are easily registered and available online for anyone to see.

These recent developments pose an interesting scenario in which the ability to do research on the stock market, individual corporations stock prices and the annual reports for each fiscal year is openly available to the public. It is well within the possibilities to construct an algorithm which predicts individual stock value changes based upon stock market data from previous days. In an environment in which the edge of stock trading lies with the one who most successfully interprets data, traders seek out to gain an advantage by considering other sources of data to better make predictions.

Amongst these advantages are data on public opinion, of which the opinion of stockholders are a significant subset. These public opinions can be measured and analyzed through different mediums (Matias, 2012). Research has been conducted which confirms online platforms such as Google (Preis, Moat, & Stanley, 2013), Wikipedia (Moat, Curme, Avakian, Kenett, Stanley, & Preis, 2013), Twitter (Bollen, Mao, & Zeng, 2010) and Facebook (Asur & Huberman, 2010) can provide data which can aid in creating prediction models. These companies provide services to gather and analyze data from their respective platforms. Research specifically geared towards Google Trends search query data suggests a correlation between the entire stock market index and search terminology (Preis, Moat, & Stanley, 2013) as well as research contradicting the effect of offline media as having a correlation to the Dow Jones stock market index (Mitchell & Mulherin, 1994).

1.1. Problem statement

Successfully predicting the stock market potentially yields significant profit while a failure to accurately do so is the bane of investments. Obtaining a reliable prediction will provide a natural advantage against your competitors. Research suggests data on user search queries can positively impact prediction algorithms. If the same procedure can be replicated and applied to the stock market it raises interesting questions. This investigative report will determine if it is possible to improve a stock prediction model based on stock data gathered from Yahoo by taking Google Trends search query data into account.

1.2. Scope of study

The stock market is a very wide and complex mechanism. Even within the boundaries of but one country, or one stock exchange, exists thousands of listed corporations with stock data
sometimes as far back as 45 years ago. However, our first and foremost limitation is that of Google Trends, which has only recorded search query data since 2004. To investigate the problem statement we believe it to be necessary to restrict our study to a time period of 5 years and a handful of individually listed corporations. These corporations will be chosen based upon what terminology would most likely be used to construct the query and how unique it is. In order to avoid ambiguity, corporations whose name or identifying trademarks are based upon common words, names or locations must be avoided. Search queries which contain the same words but are intended for a different meaning pose a risk as the data submitted might not be properly interpreted. An example of such a search query is ‘Tesla’, which could both be intended for Tesla Motors, Inc., the car manufacturer or Nikola Tesla, the scientist whose name is the basis for Tesla Motors, Inc.

A significant obstacle in this investigation is the form in which Google Trends presents its data. Google Trends does not reveal the total amount of searches for a specific query. The data presented is instead normalized on a scale of 0 to 100 according to region of origin and submission time. It is thus impossible to conclude the total amount of searches on a subject (Google Inc, n.d). The implementation of an algorithm which takes Google Trends data into account will have to solely rely on changes in the pattern of the data taken from Google Trends and as a result can not take scale between corporations into account. It will thus remain unknown whether or not a more searched upon subject has more or less of an impact on the prediction algorithm.

1.3. Disposition of the report

This report follows the standard academic structure (IMRAD) (Sollaci & Pereira, 2014), with slight modification. The following segments will initially cover the background of the subject, in regards to previous research on stock market prediction, functionality of Google Trends and in depth analysis of prediction methods. After the background, chapter three goes in depth of the method used to study the subject and our reasoning in developing and employing the method. Chapter four details the results acquired through the method. Lastly, chapter five and six interpret the results and discuss what conclusions, if any, can be drawn from the results, accumulating in a final statement on the result of the report and its validity. Besides these main segments of the report, we include chapters and segments dedicated to larger graphics, defining acronyms and an appendix of sources.
2. Background

In the background several concepts of great importance to understanding this thesis will be brought up to give the reader everything they need to understand the report. The background will bring up topics as stock market, Google Trends, common statistical methods and finally machine learning that all serve a role in providing an answer to our problem statement in our thesis.

2.1. Stock market

There exist several ways for companies to start up or expand their operations but all have one denominator in common which is the need to raise capital from an external source. Companies can take a loan from a bank but banks aren’t always so easy to convince to invest. Therefore most companies sells a part of the company to investors to acquire capital through public means instead. That part is sold in the form of shares that makes the investor become a part owner of the company that gives the person in question a part of the profit and assets (Teweles & Bradley, 2010).

Considering that selling shares in companies to get capital is quite common there’s no surprise that a market for selling these stocks evolved, the stock market. The stock market is a place where anyone can both sell and buy shares and stock. There are several reason why the stock market is a positive thing for the economy, one of them is that it prompts people to invest their savings in companies. This leads to that people don’t hold on to their savings because they stand to make a profit on their saving which lead to that more money get invested into companies which leads to higher employment rates (The Editors of Encyclopedia Britannica, 2014).

2.2. Data mining

Data mining is the science of extracting useful information from a wide range of datasets, often in the form of large deposits which are not necessarily organized (Chakrabarti et al., 2006). Data mining acts as an umbrella-term which covers several processes of gathering data which can, but do not necessarily, take on a fully automatic form. Alternatively, semi-automatic methods of data mining are usually prefered above fully manual data mining as with large data deposits the manual labor to extract data becomes very time consuming and prone to mistakes.

2.2.1. Google Trends

Google Inc. was founded in 1998 and today possess the largest market share of search engines world wide. Google launches Google Trends, based on Google Search, in 2006, as a way to capture and visualize data gathered from its search engine (Google Inc, 2014). Google Trends allows the user to access the statistics on the quantity of specific search terms used in the search engine in a given time interval. The statistics are presented most commonly as a graph detailing the amount of times a search word has been applied across the time axis.
Google Trends provides query data in proportion to overall searches and adjusted according to regionality. The data is displayed by geographical origin of the search data. Google Trends provides additional information, such as the geographical origins of search data, which it presents as a global map highlighting regional proportionality and quantity. Figure 2.1 is an example of Google Trends ranking and highlighting the search history and amount of a specific search word by region. Google Trends allows to further specify the search data by adding additional search terms and by limiting the search by region, time frame, categories and type of search (Spiegel, 2015). The queries applied to Google Trends is limited to common search phrases and will not display uncommon search terms.

Figure 2.1: Model of the map of a google trends search for “adwords”, “social media” and “seo”. Source: https://www.hallaminternet.com/2014/google-trends-introduction-business/

Google Trends has the ability to manipulate the searches by formatting the search query in different ways, there are 4 different possibilities of doing this (Google Inc, n.d). One approach is to write the keywords in any order and Google Trends will return results where the keywords appear together in any order. To get results for a specific order of keywords you place quotation marks around the sentence and it will return all results with that order. Previous mentioned techniques only returns results were all the keywords are present, if instead an addition sign is added between the keywords Google Trends will return all searches with one or more of the keywords. If a subtraction sign is instead added it will remove all results that contains the keyword written after the sign (Spiegel, 2015).
2.2.2. Programming language - R

The programming language R is one of the most used programming languages within the field of statistical computing. The language provides an environment for both statistical computing, graphics and data analysis (The R Foundation, n.d). R provides a large library with different functions to gather data, modify data and present data and enables the system of gathering or processing of this data to be automatic.

2.3. Statistics

In this chapter the statistical learning methods that are used in stock predictions will be presented and it will also include definitions and examples of the different methods that will be used during the rest of this report.

2.3.1. Regression Analysis

In the field of mathematical statistics, regression is the most widely used method to find a pattern among one or multiple independent or dependent variables and estimate a relationship. Regression analysis is used mostly for prediction and forecasting of data in a lot of different fields of studies where it also often is connected to machine learning. There exist a vast amount of different techniques for analyzing and then modelling these variables. The most universally known are the linear regression and ordinary least squares regression (Sykes, 1992).

The most basic technique of regression analysis is the linear regression that estimates a linear relationship between the provided variables. The simplest method of linear regression is

\[ y_i = a_1 + a_2 x_i + \epsilon_i \]

where \( a_1 \) and \( a_2 \) is the coefficients and \( x_i \) is an independent variable and \( \epsilon \) the error term (Sykes, 1992). Here the coefficients are easily estimated by using the formulas below.

\[
\begin{align*}
    b &= \frac{\sum x_i \cdot y_i}{\sum x_i^2} \\
    a &= \bar{y} - b \bar{x}
\end{align*}
\]

This model can of course also easily be developed by providing more variables to match a function of degree 2.

\[ y_i = a_1 + a_2 x_i + a_2 x_i^2 + \epsilon_i \]

To measure how well a regression analysis has performed compared to the actual data first of use the residual data that is the difference between the models value \( y'_i \) and the actual value \( y_i \).

\[ \epsilon_i = y'_i - y_i \]
Next step is to take out the average value from the sum of residuals which is also called the Mean Squared Error and as short MSE.

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} y_i' - y_i
\]

Another type of regression analysis is the seasonal model past data of the \( y_i \) variable is used to predict what the value will become in the future (PennState Eberly college of Science, n.d). One of the common ways to set up a seasonal model is by using the data from a month before and also from 12 months before.

\[
y_i = a + by_{i-1} + cy_{i-12} + \epsilon_i
\]

With multiple variables like in this case the calculation of the coefficients \( a, b \) and \( c \) becomes a bit more complicated but it follows the formulas below (Brannick, n.d).

\[
a = \bar{y}_i - b\bar{y}_{i-1} - c\bar{y}_{i-12}
\]

\[
b = \frac{\left( \sum y_{i-12}^2 \right) \left( \sum y_{i-1} \star y_i \right) - \left( \sum y_{i-1} \star y_{i-12} \right) \left( \sum y_{i-12} \star y_i \right)}{\left( \sum y_{i-1}^2 \right) \left( \sum y_{i-12}^2 \right) - \left( \sum y_{i-1}^2 \star y_{i-12}^2 \right)}
\]

\[
c = \frac{\left( \sum y_{i-1}^2 \right) \left( \sum y_{i-12} \star y_i \right) - \left( \sum y_{i-1} \star y_{i-12} \right) \left( \sum y_{i-1} \star y_{i-12} \right)}{\left( \sum y_{i-1}^2 \right) \left( \sum y_{i-12}^2 \right) - \left( \sum y_{i-1}^2 \star y_{i-12}^2 \right)}
\]

3. Methods

This chapter will first cover what methods have been used to gather the information and the data found in this report. It will then continue to describe how this data has been processed to get the results presented in the next chapter. It will also explain why certain methods have been chosen instead of other similar methods.

3.1. Literature Study

In the thesis several different and quite complex areas has been studied and by studying academic articles, online articles, papers and books within the related fields of the thesis the information in the report has been backed up. The fields studied are for example stock market prediction, ANN models, statistical mathematics and machine learning. The sources used in the thesis have their origin from Google Scholar, KTH library database, Universities and from a couple of well known sites Google.com/trends, Investopedia.com and nature.com.

3.2. Choice of method

To answer the thesis problem statement there’s a need to compare two different models where one takes into account the data provided from Google Trends and one does not. The choice of model was chosen due to previous work done in the area of predicting sales price of
products for companies using a seasonal AR model with and without google trends data (Choi & Varian, 2011). The research done in that area was successful and the result was an improvement of the model by using Google Trends for the prediction which this report hope to mirror but for stock prediction instead. There exists other prediction methods such as ANN for example but no research has been found to suggest that Google Trends would improve that or other known models except for the seasonal AR model. For this reason this thesis has investigate Google Trends impact on a seasonal AR models.

3.3. Data collection
To be able to construct a seasonal AR model there is a need for comparing data gathered from a company’s stock with the relative search amount on the Google search engine. Data has been collected from Yahoo by use of their URL API and from Google Trends by creating R scripts which collect and display data directly generated by Google Trends.

Search data for specific companies over a period of 5 years has been imported from Google Trends. Stock trading history from Yahoo has also been imported to be used together with the data from Google Trends to analyze if it’s possible to improve existing prediction method of the stock market with the help of Google Trends.

3.3.1. Google Trends and R
Google Trends was chosen for the reason that the search data provided from Google Trends is collected from the largest search engine in the world, Google (Zhu et al., 2012). Data was then generated by making use of R. R was chosen to interpret data above the alternative Microsoft Excel due to the excess use of manual input required by Excel which would risk data becoming compromised by human error. R allows through the use of gtrendsR (Massicotte & Eddelbuettel, 2016), an extension created by a third party, to directly scrape data and construct CSV files instead of manually downloading them as a resource from the Google Trends website (Smith, 2015). The gtrendsR extension contained the very same functionality provided by Google Trends by modifying the script.

3.3.2. Yahoo finance
Yahoo Finance has an API that allows anyone to get access to the raw data that are displayed on their website on the different companies listed. This is done by modifying the base url http://finance.yahoo.com/d/quotes.csv with adding different letters depending on what data is desired. What letter that change what are explained by Kelly Elias in Using the Yahoo Finance API for CSV (Elias, 2012). When the csv urls had been made there was a need for downloading them automatically and then to manipulate the data, this was done by using built in libraries in R (Rickert, 2015). There are similar sites that can provide the same stock data as Yahoo but due to the well documented API of Yahoo finance and the relatively easy implementation in R.
3.4. Manipulating the collected data

The collected data from Yahoo finance contains the data need to create a seasonal autoregressive model. By using the closing price \( y_i \) from the Yahoo stock data for a past date and the closing price from one month \( y_{i-1} \) respectively twelve months earlier \( y_{i-12} \) a model is created.

Model 1:

\[
y_i = a + by_{i-1} + cy_{i-12} + \epsilon_i
\]

The variable a, b and c is decided in the same way explained in the seasonal autoregressive model explanation in the background chapter. By using the data gathered there are two models created for each company one that uses the above formula directly with decided coefficients from data. The other one is equal to the other except that it takes into account a third variable, the google trends data that can be seen below.

Model 2:

\[
y_i = a + by_{i-1} + cy_{i-12} + dG_i + \epsilon_i
\]

The two model will be compared to one another and be tested with both data over a 5 and 10 year interval to measure the impact of amount of data available to the models.

4. Results

In the result chapter the base data used to create the seasonal autoregressive models is first presented that includes what companies that have been investigated and the data from the stock and Google Trends from each. In the section following that the actual created models is presented and how well they performed compared to one and another.

4.1. Base data

Initially stock data was gathered from the site Yahoo Finance about the three companies that have been investigated which are The Coca-Cola Company, British Petroleum and Wal-Mart Stores Inc. To test the difference between data of different time intervals both data from the latest 5 and 10 years was gathered. In figure 4.1 the graph shows the stock closing value ranging from 2010-07-24 to 2015-07-24 and in figure 4.2 the graph shows the same but for the time interval 2005-07-24 to 2015-07-24 instead for the three companies earlier mentioned.
Figure 4.1: Graph of the three listed companies stock values closing prices 2010-07-24 to 2015-07-24

Figure 4.2: Graph of the three listed companies stock values closing prices 2005-07-24 to 2015-07-24
Figure 4.1 demonstrates difference in stock value worth between the three example companies. The difference over time suggest dramatic short term change for BP which was one of the reasons why it was chosen to test how the models perform with quick changes. Wal-Mart Stores Inc stock is more steady growing which provides a good example for the developed model to be very accurate. The Coca-Cola Company has a more steady line, sometimes it goes up and then down again following almost a horizontal line in average.

Figure 4.3: Graph of relative search amount of queries related to the Coca-Cola company. The y axis is an index of the total search hits as explained in the Google Trends section in the background.

Figure 4.3 suggests a clear average of searches for queries related to The Coca-Cola Company with high and low spikes at certain times. The figure demonstrates annual tendencies, such as low search rates during the very end of the fiscal year. Spikes which do not appear frequently provide an insight into possible correlations between search history and stock market index history.
Figure 4.4: Graph of relative search amount of queries related to the British Petroleum company. The y axis is an index of the total search hits as explained in the Google Trends section in the background.

Figure 4.5: Graph of relative search amount of queries related to the Walmart company. The y axis is an index of the total search hits as explained in the Google Trends section in the background.
4.2. Regression models

As presented in the method two models has been compared and analyzed which has been tested on a 5 and 10 year time interval.

Model 1:
\[ y_i = a + by_{i-1} + cy_{i-12} + \epsilon_i \]

Model 2:
\[ y_i = a + by_{i-1} + cy_{i-12} + dG_i + \epsilon_i \]

The two models for each company will be presented down below and also the result when comparing the models.

4.2.1. The Coca-Cola Company


<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quota</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4.575933946</td>
<td>2.36011227</td>
<td>2.108345289</td>
<td>0.037134847</td>
<td>0.301355269</td>
<td>9.650012623</td>
<td>0.301355269</td>
</tr>
<tr>
<td>b</td>
<td>0.923760882</td>
<td>0.039584958</td>
<td>23.330124783</td>
<td>7.822864646</td>
<td>0.843649067</td>
<td>1.002156854</td>
<td>0.843649067</td>
</tr>
<tr>
<td>c</td>
<td>-0.03272915</td>
<td>0.03648481</td>
<td>-0.571257635</td>
<td>0.567563705</td>
<td>-0.10125074</td>
<td>0.557924424</td>
<td>-0.10125074</td>
</tr>
</tbody>
</table>

Table 4.1: Table showing the linear regression without Google Trends on a 10 year interval for The Coca-Cola Company and the Coefficients are a, b and c in Model 1.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quota</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3.766214847</td>
<td>2.569550989</td>
<td>1.45547657</td>
<td>0.14549627</td>
<td>-1.323912967</td>
<td>8.856342662</td>
<td>-1.323912967</td>
</tr>
<tr>
<td>b</td>
<td>0.519304547</td>
<td>0.039722069</td>
<td>23.136623</td>
<td>2.8185645</td>
<td>0.840358983</td>
<td>0.997709111</td>
<td>0.840358983</td>
</tr>
<tr>
<td>c</td>
<td>-0.01130708</td>
<td>0.040751157</td>
<td>-0.2774665</td>
<td>0.7819158</td>
<td>-0.092019883</td>
<td>0.06940572</td>
<td>-0.092019883</td>
</tr>
<tr>
<td>d</td>
<td>0.175946663</td>
<td>0.140524045</td>
<td>1.17855357</td>
<td>0.24096888</td>
<td>-0.119667779</td>
<td>0.471555105</td>
<td>-0.119667779</td>
</tr>
</tbody>
</table>

Table 4.2: Table showing the linear regression with Google Trends on a 10 year interval for The Coca-Cola Company and the Coefficients are a, b, c, d in Model 2.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quota</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4.590171739</td>
<td>3.614570077</td>
<td>1.26987</td>
<td>0.20529</td>
<td>-2.64380625</td>
<td>11.82842977</td>
<td>-2.64380625</td>
</tr>
<tr>
<td>b</td>
<td>0.923570812</td>
<td>0.058120905</td>
<td>15.8905</td>
<td>1.33-22</td>
<td>0.80718571</td>
<td>1.039955952</td>
<td>0.80718571</td>
</tr>
<tr>
<td>c</td>
<td>-0.016834457</td>
<td>0.061309444</td>
<td>-0.2746</td>
<td>0.78463</td>
<td>-0.1396045</td>
<td>0.10593555</td>
<td>-0.1396045</td>
</tr>
</tbody>
</table>

Table 4.3: Table showing the linear regression without Google Trends on a 5 year interval for The Coca-Cola Company and the Coefficients are a, b and c in Model 1.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quota</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2.243414802</td>
<td>3.5036953</td>
<td>0.6403</td>
<td>0.524589</td>
<td>-4.77533029</td>
<td>9.26215899</td>
<td>-4.77533029</td>
</tr>
<tr>
<td>b</td>
<td>0.860290208</td>
<td>0.058335502</td>
<td>14.65255</td>
<td>8.2E-21</td>
<td>0.744228535</td>
<td>0.97995188</td>
<td>0.744228535</td>
</tr>
<tr>
<td>c</td>
<td>-0.05102363</td>
<td>0.057983176</td>
<td>0.535</td>
<td>0.594768</td>
<td>-0.14718791</td>
<td>0.08514604</td>
<td>-0.14718791</td>
</tr>
<tr>
<td>d</td>
<td>1.943832875</td>
<td>0.681344231</td>
<td>2.861084</td>
<td>0.005925</td>
<td>0.584863677</td>
<td>3.314279383</td>
<td>0.584863677</td>
</tr>
</tbody>
</table>

Table 4.4: Table showing the linear regression with Google Trends on a 5 year interval for The Coca-Cola Company and the Coefficients are a, b, c, d in Model 2.
4.2.2. British Petroleum


Table 4.5: Table showing the linear regression without Google Trends on a 10 year interval for British Petroleum and the Coefficients are a, b and c in Model 1.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.27557695</td>
<td>1.81956039</td>
<td>0.701036</td>
<td>0.484661</td>
<td>-2.327648</td>
<td>4.87860189</td>
<td>-2.32764785</td>
</tr>
<tr>
<td>b</td>
<td>0.91927607</td>
<td>0.03795827</td>
<td>24.21812</td>
<td>1.342-47</td>
<td>0.8411035</td>
<td>0.9944576</td>
<td>0.84110355</td>
</tr>
<tr>
<td>c</td>
<td>0.05064097</td>
<td>0.03913485</td>
<td>1.294012</td>
<td>0.198188</td>
<td>-0.206567</td>
<td>0.1281482</td>
<td>-0.2065685</td>
</tr>
</tbody>
</table>

Table 4.6: Table showing the linear regression with Google Trends on a 10 year interval for British Petroleum and the Coefficients are a, b, c, d in Model 2.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.239982261</td>
<td>1.83709411</td>
<td>0.70763</td>
<td>0.480595</td>
<td>-2.3365141</td>
<td>4.93957634</td>
<td>-2.33614111</td>
</tr>
<tr>
<td>b</td>
<td>0.917258112</td>
<td>0.03906401</td>
<td>23.4809</td>
<td>6.85E-46</td>
<td>0.83988693</td>
<td>0.99429279</td>
<td>0.839886925</td>
</tr>
<tr>
<td>c</td>
<td>0.050873202</td>
<td>0.03987967</td>
<td>1.275092</td>
<td>0.204825</td>
<td>-0.0281492</td>
<td>0.129885556</td>
<td>-0.028149152</td>
</tr>
<tr>
<td>d</td>
<td>0.0096856</td>
<td>0.04838801</td>
<td>0.200165</td>
<td>0.841702</td>
<td>-0.086129</td>
<td>0.105524149</td>
<td>-0.08612948</td>
</tr>
</tbody>
</table>

Table 4.7: Table showing the linear regression without Google Trends on a 5 year interval for British Petroleum and the Coefficients are a, b, c in Model 1.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>11.5706621</td>
<td>5.722309</td>
<td>2.022027</td>
<td>0.0478762</td>
<td>0.11193691</td>
<td>23.02938735</td>
<td>0.111936912</td>
</tr>
<tr>
<td>b</td>
<td>0.7243067</td>
<td>0.097531</td>
<td>7.426411</td>
<td>6.193E-10</td>
<td>0.52900385</td>
<td>0.919609532</td>
<td>0.529003863</td>
</tr>
<tr>
<td>c</td>
<td>0.00802909</td>
<td>0.06389</td>
<td>0.125671</td>
<td>0.9004343</td>
<td>-0.11990755</td>
<td>0.13596574</td>
<td>-0.11990756</td>
</tr>
</tbody>
</table>

Table 4.8: Table showing the linear regression with Google Trends on a 5 year interval for British Petroleum and the Coefficients are a, b, c, d in Model 2.
4.2.3. Wal-Mart Stores Inc


<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.86778747</td>
<td>1.376576816</td>
<td>0.593595</td>
<td>-1.8584508</td>
<td>3.59402572</td>
<td>-1.85845078</td>
<td>3.59402572</td>
</tr>
<tr>
<td>b</td>
<td>0.941276716</td>
<td>0.038124716</td>
<td>24.68996</td>
<td>3.22E-48</td>
<td>0.88579391</td>
<td>1.01680191</td>
<td>0.88579381</td>
</tr>
<tr>
<td>c</td>
<td>0.049336278</td>
<td>0.044756478</td>
<td>1.102372</td>
<td>0.272553</td>
<td>-0.0392996</td>
<td>0.13797614</td>
<td>-0.03929558</td>
</tr>
</tbody>
</table>

Table 4.9: Table showing the linear regression without Google Trends on a 10 year interval for Wal-Mart Stores Inc and the Coefficients are a, b and c in Model 1.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.355267637</td>
<td>1.60783435</td>
<td>0.749664</td>
<td>0.454795</td>
<td>-2.22537608</td>
<td>4.935911357</td>
<td>-2.22537608</td>
</tr>
<tr>
<td>b</td>
<td>0.940436249</td>
<td>0.038130985</td>
<td>24.59319</td>
<td>3.25E-48</td>
<td>0.86460276</td>
<td>1.016357736</td>
<td>0.86460276</td>
</tr>
<tr>
<td>c</td>
<td>0.046207477</td>
<td>0.045535283</td>
<td>1.014762</td>
<td>0.312331</td>
<td>-0.04398089</td>
<td>0.136395831</td>
<td>-0.04398089</td>
</tr>
<tr>
<td>d</td>
<td>-0.0140847</td>
<td>0.03369261</td>
<td>-0.41804</td>
<td>0.676694</td>
<td>-0.0081716</td>
<td>0.052647757</td>
<td>-0.0081716</td>
</tr>
</tbody>
</table>

Table 4.10: Table showing the linear regression with Google Trends on a 10 year interval for Wal-Mart Stores Inc and the Coefficients are a, b, c, d in Model 2.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3.859108424</td>
<td>2.500705454</td>
<td>1.543208</td>
<td>0.128314</td>
<td>-1.14846872</td>
<td>8.86668472</td>
<td>-1.14846872</td>
</tr>
<tr>
<td>b</td>
<td>0.888937692</td>
<td>0.064142304</td>
<td>13.85396</td>
<td>6.65E-20</td>
<td>0.760544764</td>
<td>1.01743062</td>
<td>0.760544764</td>
</tr>
<tr>
<td>c</td>
<td>0.064221249</td>
<td>0.055888803</td>
<td>0.97915</td>
<td>0.331644</td>
<td>-0.05711806</td>
<td>0.19556055</td>
<td>-0.05711806</td>
</tr>
</tbody>
</table>

Table 4.11: Table showing the linear regression without Google Trends on a 10 year interval for Wal-Mart Stores Inc and the Coefficients are a, b and c in Model 1.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Meanvalue</th>
<th>t-quata</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Higher 95%</th>
<th>Lower 0.95%</th>
<th>Higher 0.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4.296673746</td>
<td>3.25133619</td>
<td>1.321512</td>
<td>0.191705</td>
<td>-2.216530297</td>
<td>10.809888</td>
<td>-2.2165303</td>
</tr>
<tr>
<td>b</td>
<td>0.889211098</td>
<td>0.064694794</td>
<td>13.74471</td>
<td>1.37E-19</td>
<td>0.75961186</td>
<td>1.01881033</td>
<td>0.75961186</td>
</tr>
<tr>
<td>c</td>
<td>0.061745144</td>
<td>0.067155777</td>
<td>0.919447</td>
<td>0.361806</td>
<td>-0.072783034</td>
<td>0.19627531</td>
<td>-0.072783034</td>
</tr>
<tr>
<td>d</td>
<td>-0.01897497</td>
<td>0.08998847</td>
<td>-0.21323</td>
<td>0.833922</td>
<td>-0.197240296</td>
<td>0.1592904</td>
<td>-0.197240296</td>
</tr>
</tbody>
</table>

Table 4.12: Table showing the linear regression with Google Trends on a 10 year interval for Wal-Mart Stores Inc and the Coefficients are a, b, c, d in Model 2.

4.3. Mean absolute error

When comparing the different models prediction values vs the real stock price at the time under the periods specified the mean absolute errors for each was found.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Wal-Mart Stores Inc</th>
<th>British Petroleum</th>
<th>The Coca-Cola Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOGTrends - 10 years</td>
<td>2.06661655</td>
<td>2.898369232</td>
<td>2.175372252</td>
</tr>
<tr>
<td>GTrends - 10 years</td>
<td>2.078911598</td>
<td>2.902489076</td>
<td>2.223799433</td>
</tr>
<tr>
<td>NOGTrends - 5 years</td>
<td>2.23848106</td>
<td>2.208717474</td>
<td>2.637606669</td>
</tr>
<tr>
<td>GTrends - 5 years</td>
<td>2.24433019</td>
<td>2.207053585</td>
<td>2.684339547</td>
</tr>
</tbody>
</table>

Table 4.13: Table showing the mean error with and without Google Trends for the different intervals for the three companies that has been tested on the two models.
5. Discussion

The discussion will take up vital information that might have change the result to the worse or to the better if changes had been done in the method. The chapter will also contain a general discussion around the given results.

5.1. Discussion of results

The most essential thing to bear in mind when interpreting the final result is that a search query submitted to Google is highly subjective. We can not empirically establish a connection between search terminology and intention behind the search query. It is thus only in the grand scheme of things that Google Trends data can be considered of any value. Whatever the intention of a specific search query relating to a corporation, its increase or decrease in frequency symbolizes a change in the desired outcome by those who submit said query.

Initially, the thesis was intended to analyze differences between search terminology related to a specific corporation and compare the differences in regards to changes in the stock price of said corporation. Unfortunately the very low amount of searches, in proportion to neutral search queries, on the Google search engine with positively or negatively charged terminology made it impossible to measure any distinct changes. In the case of specific positively or negatively charged terminology alongside a corporation did not meet the criteria to be saved by Google Trends and was thus inaccessible to us.
Furthermore, the results are limited to the few corporations we have taken upon ourselves to examine. These examples have been chosen based upon not only the criteria mentioned in the scope of study but also in regards to variety amongst both search query data and stock price data. Our examples bring up special circumstances in which there have been significant changes in both search query data and stock price data. The Deepwater Horizon oil spill disaster led to a substantial public reaction at BP plc which were at the center of the scandal. BP search query data faced enormous spikes during this period and can clearly be identified by taking a look at the graph results. With the model acting over a period of ten years, it covers the events of the BP search query spike. Both the five and ten year analysis covers the events of a tremendous BP stock drop. Such significant changes in the data gathered does not appear to affect the prediction models despite their obvious correlation. This suggests it may be of importance to select corporations which differ from our example corporations, with factors such as corporation size, popularity and online presence playing a part.

5.2. **Discussion of method**

The methods we have conducted to scrape data from Google Trends and Yahoo Finance are very simple in their nature and do not leave much room for debate. The popularity and size of Google Trends and Yahoo Finance suggest reliability and we consider these sources perfectly trustable for their respective purposes. Considering that previous research has been conducted on alternative platforms that might represent public opinion, we believe that the choice of Google Trends is well motivated not only in prospective results, but also for a hindsight analysis of differences between Google Trends and other sources of search query data. We believe this lends the potential to create a discussion of differences between search engines, social media platforms and data sharing websites, as different sources of data representing public opinion may spawn different conclusions.

There are alternatives to autoregressive algorithms to predict stock market movements. A specific algorithm we are particularly interested in is the ANN method of machine learning which can be constructed to predict stock price changes and how it would be impacted by taking search query data into account.
6. Conclusion

This thesis has sought out to investigate the possibility to improve stock prediction algorithms by modifying the models in accordance to concurrent search query data from Google Trends. Search query data has derived from Google Trends has been implemented into an autoregressive prediction algorithm and applied over a five and a ten year time interval. The comparison of the unmodified and modified prediction models show little to no difference and suggest that Google Trends search query data has not had a significant impact on the prediction model. The differences in prediction results between the two models are very small on the example corporations and demonstrate that further investigation must be applied to conclude whether or not the search query data can have a positive impact on the autoregressive algorithm. The conclusion is that the hypothesis remains unconfirmed and that an autoregressive prediction model does not improve prediction of stock prices.
7. Bibliography


