How can Google Trends and sentiment analysis of TripAdvisor and Facebook predict visitor numbers to the United Kingdom and Canada from 2010 - 2015?

MSc Research Project
Data Analytics

Emmet Hutchin
x15018695

School of Computing
National College of Ireland

Supervisor: Mr. Tony Delaney

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Emmet Hutchin
Student ID:	x15018695
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Mr. Tony Delaney
Submission Due Date:	22/08/2016
Project Title:	How can Google Trends and sentiment analysis of TripAdvisor and Facebook predict visitor numbers to the United Kingdom and Canada from 2010 - 2015?
Word Count:	6,167

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	13th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

How can Google Trends and sentiment analysis of TripAdvisor and Facebook predict visitor numbers to the United Kingdom and Canada from 2010 - 2015?

Emmet Hutchin
x15018695
MSc Research Project in Data Analytics

13th September 2016

Abstract

Internet search data and sentiment analysis are becoming standard prediction methodologies and have been used to predict everything from the stock market to cinema box office revenues, they have been seen as separate tools to carry out analysis. This paper attempts to bridge the gap in the existing research by using search and sentiment analysis to predict future visitor numbers to the UK and Canada. Sentiment analysis using the OL and HGI lexicons will be performed on hotel reviews for the UK and Canada scraped from TripAdvisor and comments extracted from hotel Facebook pages over a five-year period. This data will be matched with Google Trends data. It also aims to establish the best way to combine the data and using regression analysis. Some promising results are achieved by using multiple regression with Google TripAdvisor and Facebook data achieving over twice the correlation with visitor numbers than the highest scoring individual variable.

1 Introduction

This paper will examine how data from Google Trends, sentiment analysis of TripAdvisor reviews and Facebook likes and shares can be used to predict the number of visitors to Canada and the United Kingdom (UK) from 2010 - 2015. Connections have been established between public sentiment and Internet search volume and changes to the stock market and to successfully predict movie, game and music revenue (Curme et al. (2014), Asur and Huberman (2010) and Choi and Varian (2012)). However, these predictions used sentiment analysis and search data separately and this paper will attempt to bridge the gap in the existing research by combining the two methods.

Before the widespread adaption of the Internet, if a business wanted find out how it was perceived it would have to conduct surveys which were costly and time consuming and by their nature could only reach a limited number of people. This was not a problem if the company was in a very niche area, however most business are not. Now the Internet offers many forms of data at the touch of a button enabling companies to effectively survey millions of potential customers for very little cost. This paper will examine two of the most popular forms of data, Internet search data in the form of Google

Trend data, the most popular search engine (Cooper et al. (2005)), sentiment analysis obtained from TripAdvisor, the most popular travel website (Jeacle and Carter (2011)) and Facebook, the most popular social network site in the world (Ortigosa et al. (2014)). Both individually and combined, these should provide the data necessary to predict the visitor numbers. As TripAdvisor does not provide reviews for cities, only hotels, tourist attractions etc. a random sample of data from these will be taken and as far as possible the same sample data will be used to obtain data from Facebook and Google Trends.

The Internet is becoming more and more prevalent in people's lives and even as far back as 2005, 64% of online travellers used search engines to plan their travel (Xiang and Gretzel (2010). If business can accurately predict customer numbers they can ensure the correct number of staff, facilities etc. are available to give their customers the best possible experience and also reduce costs/increase profits. It will also allow hotels/airlines/attractions to adjust their prices if demand is forecast to be low in order to encourage customers during these periods thus leaving them less prone to market fluctuations.

Internet search data and sentiment analysis has been used to successfully predict outcomes in a diverse range of areas, however they are generally not used together in studies. This paper will address this gap as part of its analysis and determine which method works best or if combinations of data work best. (Hutchin (2016))

2 Related Work

2.1 Introduction

This paper will examine how Google Trend data and sentiment analysis carried out on TripAdvisor reviews and Facebook comments can be used to predict visitor numbers to the UK and Canada.

People from countries with a higher Gross Domestic Product(GDP) have a propensity to use Internet searches to look forwards rather than backwards (Preis et al. (2012)) and almost 92% of Canada's overseas visitors in 2015 came from the United States, France, Germany, United Kingdom, Japan and Australia (*tourismsnapshot-dec2015_en-lowres* (n.d.)), (all areas with relatively high GDPs). Google searches for the cities of Miami, Tampa and Sarasota Florida were found to correlate to visitor numbers four weeks in the future (Choi and Liu (2011)). Using web search has been proven to successfully predict consumer behaviour, although it works best in the absence of other models (Goel et al. (2010)), this will not hinder this paper as each distinct tourism entity may have other data available to them individually i.e. an airline will know how many tickets have been booked but no other entities in the tourism market will have access to this (other than general forward guidance to the stock market). There will be no guide available to the market as a whole that will predict how it will perform in the future.

This review will examine using search data, sentiment analysis and the analytical techniques that can be used to predict visitor numbers to the UK and Canada from 2010 to 2015.

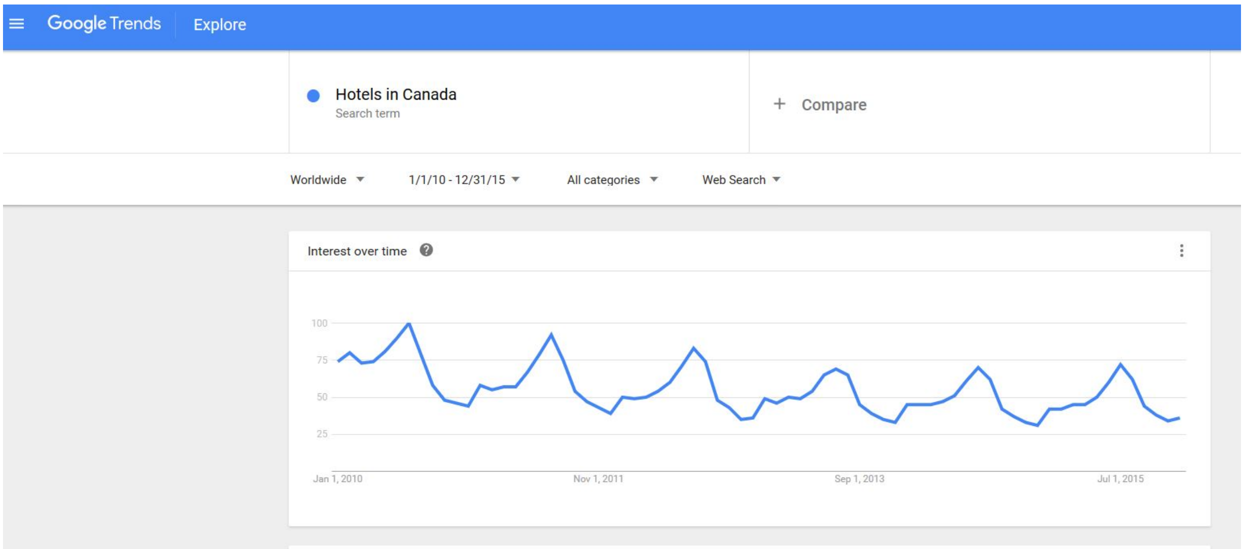
2.2 Search Data

Using Internet search engine data to predict future events became popular in the mid 2000's with the first paper published in 2005 (Choi and Varian (2012)), which used web search data based on job search data from the preceding weeks against the official U.S. monthly unemployment data to predict the rate of unemployment. Etteredge et al. use search data provided by Rivergold Associates, Ltd. in their WorldTracker's Top 500 Key Word Report. Two reports containing data from the previous twenty-four hours and sixty days respectively covering September 15th, 2001 to March 1st, 2003 were used. Ten weeks of data were missing which was extrapolated from the surrounding weeks. They selected six key terms used by job seekers, (Ettredge et al.; 2005, p. 88-89). The extrapolation of some data could potentially impact on the accuracy of the results.

Although other studies are trying to predict a wide variety of future events they all broadly follow the same data gathering techniques for obtaining data from search engines such as Yahoo! or Google. Yahoo! was the most popular search engine in the United States up until 2003, when it was overtaken by Google (Cooper et al.; 2005, p. 3).

Due to the decreasing popularity of Yahoo! more recent studies have generally used Google Trends to obtain search data, Google Trend provides an index of the volume of search queries on a given term over a period of time and geographic area as specified by the user and is based on query share which is the total query volume divided by the total number of queries during the parameters set by the user. As it is an index the maximum value is normalised to 100 and the query share at the start data is normalised to 0 (Choi and Varian; 2012, p. 212).

An example of a Google Trend Query can be seen in Figure 1 and the data can be downloaded as a csv file.



Hotels in Canada
Search term

Worldwide ▾ 1/1/10 - 12/31/15 ▾ All categories ▾ Web Search ▾

Figure 1: Google Trend Query for Canadian Hotels

It has been used to predict visitor numbers to Hong Kong over a 6-year time span. Google Trends data was broken down by country of origin and the average of the first two weeks' data was compared against the Hong Kong Tourism Board monthly data to give six weeks' lead time in terms of forecasting the visitor numbers (Choi and Varian (2012)). This study will attempt to predict visitor numbers from Google Trends data using an appropriate time lag and will also examine sentiment analysis and compare its performance to Google Trends data. The sentiment analysis is extremely important to the study as search engines have inbuilt limitations in presenting complex areas such as tourism. Social media and search engines have a symbiotic relationship whereby social media is updated frequently and contains a lot of hyperlinks to other sites, thus the search engines visit them frequently and move them up their rankings (Xiang and Gretzel; 2010, p. 181) as Google uses the number of hyperlinks as part of its Page ranking algorithm. When users enter search terms relating to holiday destinations they are more likely to be directed to a social media site than any other (Xiang and Gretzel; 2010, p. 185).

2.3 Sentiment Analysis

In contrast to web search studies which generally track data over a number of years (Preis et al. (2012), Ettredge et al. (2005) etc.), sentiment analysis studies generally track data over a few months analysing millions of records in the process. Asure et al. analysed nearly 3 million tweets relating to 24 movies over a three-month period in their study to predict box office revenues from an analysis of Twitter sentiment (Asur and Huberman; 2010, p. 493). Similarly, the study to forecast closing price of the Dow Jones Industrial Average (DJIA) analysed nearly 10 million tweets from February 28th to December 19th, 2008 (Bollen et al.; 2011, p. 2).

This paper will attempt to bridge the research gap by using sentiment analysis of Facebook and TripAdvisor rather than Twitter over a longer time frame (5 years) to predict the number of visitors to the UK and Canada and will allow comparison between the two most popular forms of prediction.

Facebook is the world's most popular social network site, in 2012 it had 550 million daily users and 1 billion monthly users. The advantage it has over other social network sites is that the relationships are based on friendships rather than business acquaintances etc. and this lends more weight to comments posted on the site as people are more likely to trust their friends (Ortigosa et al.; 2014, p. 528). TripAdvisor is a review website established in February 2000 and by 2011 attracted over 40 million visitors per month and covered over 125,000 visitor attractions and 450,000 hotels. The advantage it has over other travel websites is it contains reviews written by visitors to the attraction/hotel rather than professional reviewers. This allows a greater number of reviews and and the aggregate of these should give more accurate results as well as greater trust as readers will not see the reviewer as having something to gain from the review (Jeacle and Carter; 2011, p. 298).

Sentiment analysis involves extracting a collection of words, often from Twitter, newspapers or blogs, into a corpus and this is analysed against a preformed lexicon (dictionary) for positive, negative and sometimes neutral words or phrases to produce a score for the analysed data. The lexicon is so key to sentiment analysis that the first paper on the subject was exclusively concerned with generating the lexicon (Hatzivassiloglou and McKeown (1997)). Later studies such as Godbole et al built on Hatzivassiloglou et al's. work by improving on the lexicon development process. The paper analysed the computer dictionary WordNet and tracked how many positive and negative words were encountered when moving from the start to the end of a set of words e.g. good and bad (Godbole et al. (2007)).

Bollen et al. study to find correlations between Twitter sentiment analysis and movements in the DJIA observed Tweets over a 10 months. It also refined many of the ideas used to generate the sentiment analysis. They used OpinionFinder which automatically assigns a positive or negative score each day and Google-Profile of Mood States (GPOMS) to break down the daily sentiment analysis into 6 categories such as clam, happy etc. for more accurate tracking. They also improved on the pre-processing by removing stop words such as "and", "the" etc., punctuation and spam tweets i.e. tweets containing "www" or "http" and only included tweets containing a mood such as "I feel" to ensure the public mood was captured and improve the accuracy of the results (Bollen et al. (2011)). The idea of using multiple lexicons was further developed by Olivereira et al. during their research into predicting stock market changes using Twitter sentiment analysis. They used the Harvard General Inquirer (GI), Opinion Lexicon (OL), Macquire Semantic Orientation Lexicon (MSQL), MPQA Subjectivity Lexicon (MPQA), SentiWordNet (SWN), Emoticon Lexicon and a union of all the lexicons to create the positive/negative sentiment score. They also reduced the "noise" in the tweets by only selecting tweets with "cashtags" i.e. those with a $ before the stock ticker symbol as this is the twitter convention for discussing stock market related issues for companies increasing accuracy(Oliveira et al. (2013)).

2.4 Data Analysis

This paper will use linear regression to analyse the results of the Google Trends and sentiment data and to predict the number of visitors to the UK and Canada. Previous studies used clustering to try and predict consumer purchasing behaviour (Yadav et al. (2012)) however they used web log data rather than the trend and sentiment data this study will be using. As the search, sentiment analysis and tourism visitor data is all time bound it is appropriate to use linear regression and time series analysis. The Google Trends and visitor data can be analysed using linear regression to determine how close the correlation between the data sets is (Goel et al. (2010), Choi and Varian (2012), Cooper et al. (2005) and Ettredge et al. (2005)). Other approaches have been used to predict the future with Google Trends data such as generating a set of random stock market returns and calculating the standard deviation of the Google Trends data against these (Preis et al. (2012)), however the linear regression model works better in this paper. (Hutchin (2016))

3 Methodology

3.1 Methodology

The methodology used for this paper will be knowledge discovery and data mining (KDD). The steps are shown in Figure 2 below.
Selection:
The data selected for this paper will consist of visitor numbers by month for the UK and Canada from 2010 - 2015. These will be measured against TripAdvisor review data, Facebook and Google Trend data as described in section 3.2 below.
Pre-processing:
This stage consists of the extraction and cleansing of the data as described in sections 3.2 and 3.3 below.
Transformation:
This stage consists of loading the data into an appropriate file structure using R for full analysis.
Data Mining:
The data will be analysed in R as described in section 3.4 below.
Interpretation/Evaluation:

The results will be evaluated for any correlation between the review/trend data and changes to visitor numbers.

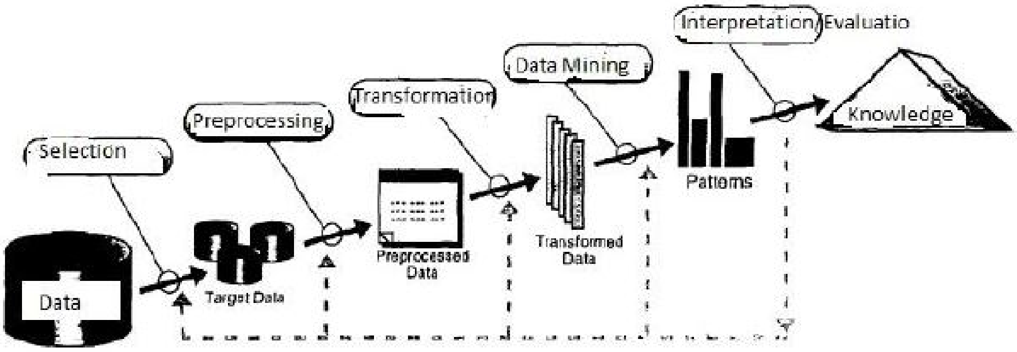


Figure 2: An Overview of the Steps Comprising The KDD Process.
Source: Fayyad et al. (1996)

3.2 Data Gathering

The data sets for this paper will consist of visitor numbers to the UK and Canada for 2010 - 2015 from the relevant authority, broken down by month.

The main focus of the data gathering phase will be obtaining data from TripAdvisor, Facebook and Google Trends. TripAdvisor will not provide an API key for academic or data research purposes, however, the data can be scraped from the website using the rvest package (Wickham and RStudio (2016a)) in the R programming language (*R: The R Project for Statistical Computing* (n.d.)). TripAdvisor do not have reviews for countries such as the UK and Canada, rather they allow users to review places within these countries. Review data will be scraped from a representative sample of hotels listed on TripAdvisor for the countries. The scraped data will contain the headline quote of the review, the number of stars given and the rest of the review and will be stored as a csv file.

Facebook data can be obtained by means of an API authorisation token and the rFacebook (Barbera et al. (2016)) package in the R programming language. To ensure consistency with the TripAdvisor data, only the hotels from the TripAdvisor sample found on Facebook will be used. As most hotels restrict postings on their pages to themselves, the posts will be downloaded and the number of likes and shares counted. Any user comments made on these posts will also be downloaded. These will be aggregated for use as another form of sentiment analysis.

For analytical consistency, data for the same hotels used in TripAdvisor will be extracted from Google Trends. This data will be downloaded and stored in csv files.

3.3 Data Cleaning and Processing

R can be used to extract TripAdvisor and Facebook data in such a way that it requires minimal cleaning, each of the separate files for hotels will be amalgamated into one file and punctuation marks etc. can be removed. R can also be used to pass the cleaned data into the data dictionaries (lexicons) and perform the sentiment analysis.

The Google Trend data files also have to be checked for date consistency and fixed if necessary in Excel as this is an ideal tool for working with smaller csv files. As search volume changes are being tracked, the files will have to be combined to give an indicative change in the search activity for the UK and Canada. The score will be divided by the number of hotels downloaded to maintain the 0-100 index.

3.4 Data Analysis

A random sample of 80% of the data will be used to train the regression model and the remaining 20% will be used to test how closely the predictions were to the actual figures. This will avoid any over fitting of the data to the model.

3.4.1 Linear Regression

Linear Regression will be used to examine the relationship between the individual variables on the visitor data. Multiple regression will be used to examine the relationship between the independent variables and the monthly visitor data. It will also be used to determine the benefit of using all three variables and combining them in one model.

3.4.2 Regression

Simple Linear Regression is the relationship between a dependent variable and a single independent predictor variable defined by the equation:
$y = \alpha + \beta x$
The intercept α describes where the line crosses the Y axis and the slope β describes the change in y given an increase in x (Lantz (2013)) e.g. the expected change in visitor numbers for a given increase in Google Trend scores.

Multiple Regression is used in cases of more than one independent variable and will be used to examine the relationship between the combined values of Google Trends, TripAdvisor and Facebook and their relationship with the visitor data.
$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$
With y being the sum of the intercept term and the product of the β values. ϵ has been added as an error term as the predictions are not perfect (Lantz (2013)). (Hutchin (2016))

4 Implementation

4.1 Environment

The experiments in this paper were carried out using the open source R programming package (*R: The R Project for Statistical Computing* (n.d.)) version 3.2.2. The Facebook data was extracted using the rFacebook package (Barbera et al. (2016)), TripAdvisor data was extracted using the rvest package (Wickham and RStudio (2016a)) and Google Trends data was extracted using the gtrendsR package (Massicotte and Eddelbuettel (2016)), sentiment analysis was carried out using the stringr package (Wickham and RStudio (2016b)) and the statistical analysis used the psych (Revelle (2016)) and car (Fox et al. (2016)) packages. The data was stored on a Dell Inspiron 5547 with an Intel(R) Core(TM) i7-4510U CPU and 16GB RAM running Windows 10 64- bit operating system.

4.2 Data Extraction

4.2.1 Visitor Data

As the visitor data for Canada is only available from the Canadian tourist authority, Destination Canada, in their monthly reports (*tourismsnapshot-dec2015_en-lowres* (n.d.)), a copy of each PDF report was downloaded and the visitor data was copied into a csv file. The monthly data for the UK was available as a download from the Office for National Statistics (for a full list of the data sources see appendix 8 and 9 in the user configuration manual). This was then stored as a csv file.

4.2.2 TripAdvisor

TripAdvisor does not provide reviews of locations, only hotels, tourist attractions etc. in those locations and they do not provide a data extraction API for academic research so the reviews from each hotel will be scraped using the rvest package in R. TripAdvisor provides a full list of Canadian hotels by province and the name and URL for all of hotels was extracted by province and stored in a csv file. A random sample of 10% of the hotels was taken in proportion to the number of visitors that began their trip in each province. This list was consolidated using R and the maximum number of review pages per hotel review was added to the file. The headline review, star rating, full review (or part of the review if it was a long review as the full long review could only be viewed if logged into TripAdvisor) and date of review were scraped and stored in a file which were amalgamated via R for use in sentiment analysis.

A similar approach was taken for the UK, however as it is not traditionally divided into provinces TripAdvisor provided a list of hotels per town. This was scraped and the list used to create a master list for the UK. A random sample of 5% of the hotels in proportion to the number of visitors to England, Scotland, Wales and Northern Ireland was used to scrape the reviews etc. as per Canada. A smaller sample was used for the UK due to the vast difference in the number of hotels on TripAdvisor in the two countries, approximately 5,400 in Canada and 12,000 in the UK.

4.2.3 Facebook

An R program used the sample list of hotels to obtain the Facebook ID and URLs of any hotels that had a Facebook page. The rFacebook package requires the ID to extract the information from the Facebook page via an API. The package extracted the message posted on the Facebook page, the date, number of likes and shares. As most of the pages were set up so that only the hotel could post messages, and these are likely to be overwhelming positive, any comments associated with the message and their date were also downloaded along with the number of likes each comment received. This was stored in a separate file and consolidated into a for sentiment analysis.

4.2.4 Google Trends

The list was also passed into an R program using the gtrendsR package to obtain the Google Trends score for each hotel (if there were enough searches carried out on that hotel to return data). This data was stored in separate csv files and amalgamated using R. This data is produced by week so an average value for all of the hotels was taken per week in Excel. This data was then converted into monthly values in excel (*Google-Trends-Excel-Template-Convert-Weekly-Data-to-Monthly.xlsx* (n.d.)). Data was returned for 332 hotels for Canada and 204 for the UK.

4.3 Sentiment Analysis

Two lexicons were used for sentiment analysis, the Opinion Lexicon (OL) (Hu and Liu (2004), Liu (2010)) which comprised of a negative word file and a positive word file and the Harvard General Enquirer (HGI) which is a spreadsheet of positive and negative words as well as different emotions. The words were separated into respective files. Each of these dictionaries was used by an R program to calculate the sentiment score for every comment/review for Facebook and TripAdvisor. The files were then manipulated in Microsoft Excel to format the dates and the scores for each month were calculated. For Canada, Facebook had approximately 150 hotels with 10,000 scoreable comments from a total of 37,000 comments. Were as TripAdvisor had 93,000 scoreable comments from approximately 100,000 reviews. For the UK Facebook had 138 hotels nearly 71,000 scoreable comments from 163,000 comments. TripAdvisor had 227,000 scoreable comments from a total of 244,000 comments.

4.4 Time Lag

The sentiment analysis scores for Facebook and TripAdvisor and the Google trend score for each month were added to a file containing the appropriate visitor data ready to be tested for regression analysis using R.

As this paper aims to achieve a predictive model, using data for January 2015 against visitor data for January 2015 cannot be used as some of the Google Trend data could have been taken after the visits had already taken place. Linear regression was used to determine the closest correlation between each of the independent variables and the visitor with a time lag from one to twelve months for each of the countries.

When the appropriate time lag for the independent variables was identified, 80% of the data was randomly selected using the Rand function in Excel and used for training the regression model. The remaining 20% was used to test the model.

5 Evaluation

Linear regression was carried out on each of the independent variables against the Canadian and UK visitor data respectively to determine the optimal time lag and the results can be seen for Canada in Table 1 and the UK in Table 2 below:

Variable	Time Lag (Months)	Correlation
Google Trend	12	0.8315886
TripAdvisor	3	0.598045
Facebook	6	0.598045

Table 1: Optimal Time Lag for Canada (in months)

The optimal time lag for Canada is 12 months for Google Trends, 3 months for TripAdvisor and 6 months for Facebook. Similarly, the UK is 7 months for Google Trends, 4 months for TripAdvisor and 1 months for Facebook.

Variable	Time Lag (Months)	Correlation
Google Trend	7	0.648983
TripAdvisor	4	0.6246374
Facebook	1	0.598045

Table 2: Optimal Time Lag for the UK (in months)

5.1 Canada

Linear/Multiple Regression was carried out for each of the independent variables, using the three variables and the mathematical total of their score. The best results were obtained using the average sentiment score for TripAdvisor and Facebook obtained from the HGI lexicon and the Google Trends data. The total is the mathematical total of the three independent variables and G+T+F is the multiple regression score for the three variables. The results for each of the variables can be seen in Table 3:

Variable	R-squared	Adjusted R squared	Correlation
Google Trend	0.32	0.3042	0.775888
TripAdvisor	0.202	0.1834	0.05481717
Facebook	0.1977	0.179	-0.4626158
Total	0.3503	0.3352	0.7816762
G+T+F	0.6983	0.6762	0.8627918

Table 3: Canadian Regression Score

The data has been tested for multi-collinearity (figure 3) and the correlation scores between the independent variables are not high enough to cause concern. This is re-enforced by the correlation ellipse which has a greater stretch the higher the correlation and the ellipses in the diagram are closer to ovals. The Variance Inflation Factor (VIF) of the data was also tested as values of 10 or more can indicate multi-collinearity. Scores of Google Trends 1, TripAdvisor 1.3 and Facebook 1.3 again show there is no dependency between the independent variables.

There is also a linear relationship between the independent variables and the visitors (Figure 4). The almost perfect correlation between the Google Trends score and the Total score is due to the Google Trends score making up a large part of the total score as the average sentiment scores are being used.

The Ceres plot in figure 5 shows the linear relationship between the independent variables when they are used together in the model unlike figure 4 where the linear relationship is shown for their individual use against the visitor data. Again the linear relationship between the independent and dependent variables can be seen.

To ensure there was not auto-correlation between the data the independent variables residuals were plotted and the results are in figure 6. A Durbin-Watson statistic of 1.9 backs up the residuals plot and shows there is no autocorrelation in the data.

Heteroscedasticity value of 1.7 and p-value of 0.19 show the data does not violate the rules of homoscedasticity.

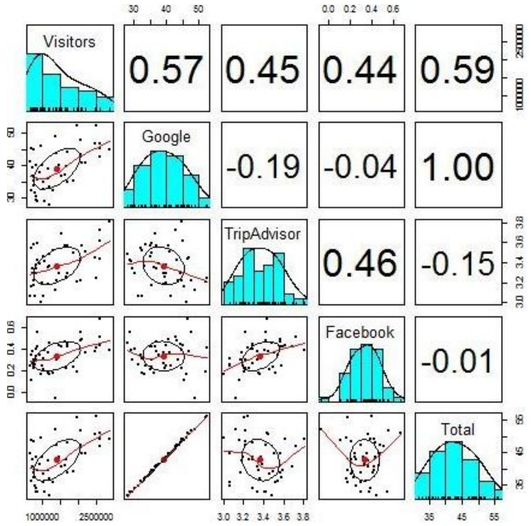


Figure 3: Correlation Between the Independent Variables And The Canadian Visitor Data

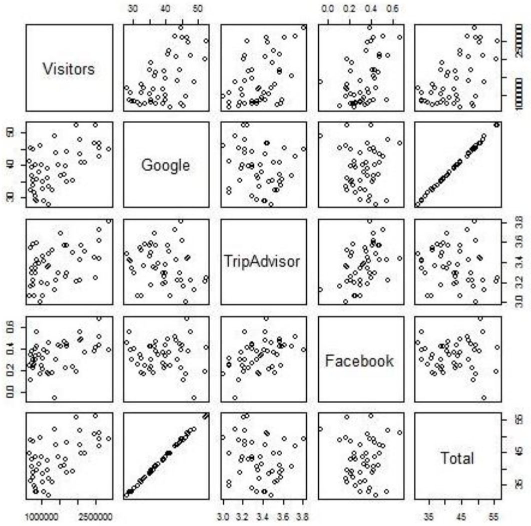


Figure 4: Linear Relationship Between the Independent Variables and the Canadian Visitor Data

5.2 The UK

However, for the UK the sentiment analysis score for Facebook for the two lexicons was the same so only the OL data was examined using Linear Regression. Linear/Multiple Regression was again carried out on the data as per Canada and the results are in Table 4:

The data has been tested for multi-collinearity and the results are in figure 7. While the correlation score between Google score TripAdvisor is high it's not high enough to cause concern. VIF scores of Google Trends 2.1, TripAdvisor 2.1 and Facebook 1.3 show there is no dependency between the independent variables.

The linear relationship between the independent variables and the visitor data is in Figure 8 and the linear relationship between the independent and dependent variables can be seen in the Ceres plot in Figure 9.

To ensure there was not auto-correlation between the data the independent variables

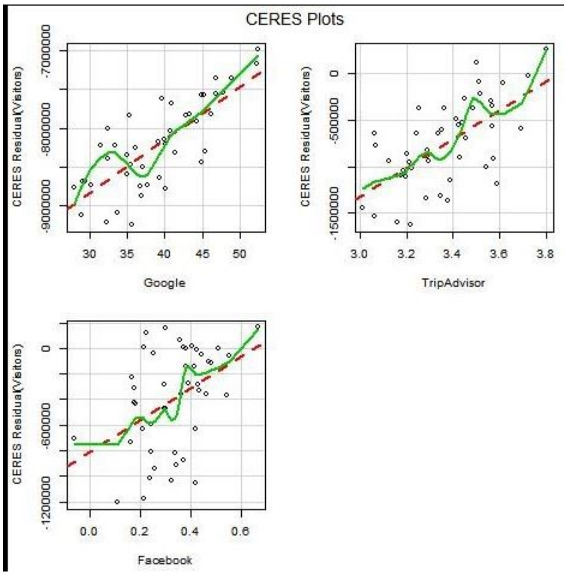


Figure 5: Linear Relationship Between the Independent Variables Used Together and the Canadian Visitor Data

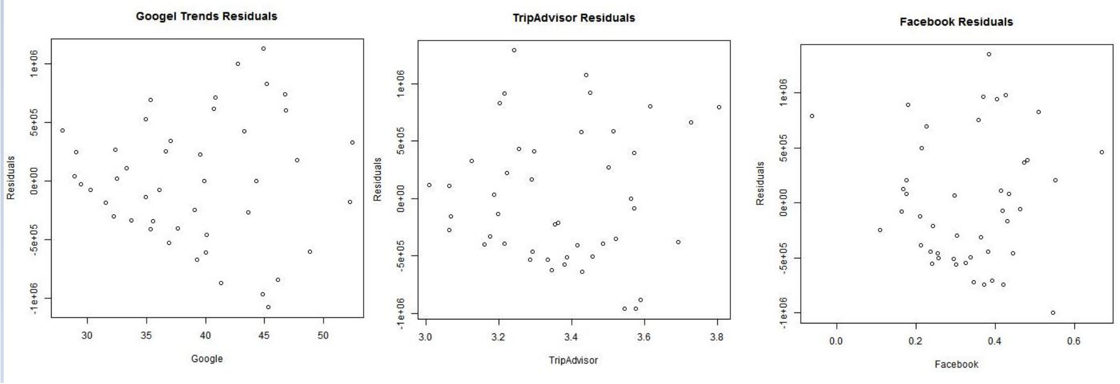


Figure 6: Residuals Plot for the Independent Variables

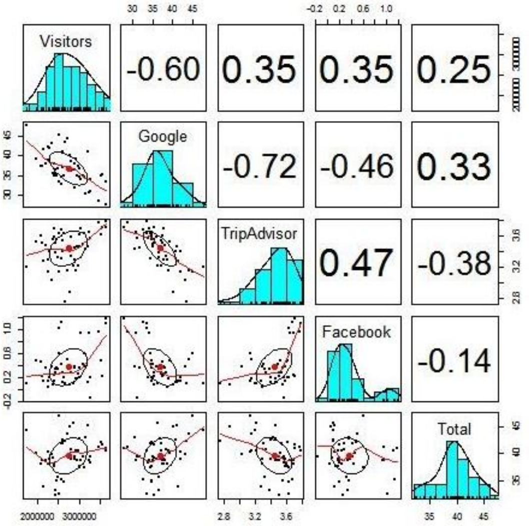


Figure 7: Correlation Between the Independent Variables and The UK Visitor Data

Variable	R-squared	Adjusted R squared	Correlation
Google Trend	0.3592	0.3459	0.6951864
TripAdvisor	0.1253	0.1071	0.4020472
Facebook	0.1237	0.1055	0.8036658
Total	0.0637	0.0442	0.2282867
G+T+F	0.3834	0.3432	0.8627918

Table 4: UK Regression Score

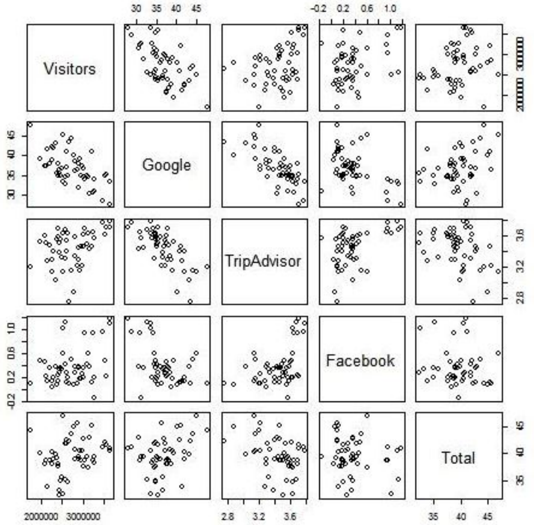


Figure 8: Linear Relationship Between the Independent Variables and The UK Visitor Data

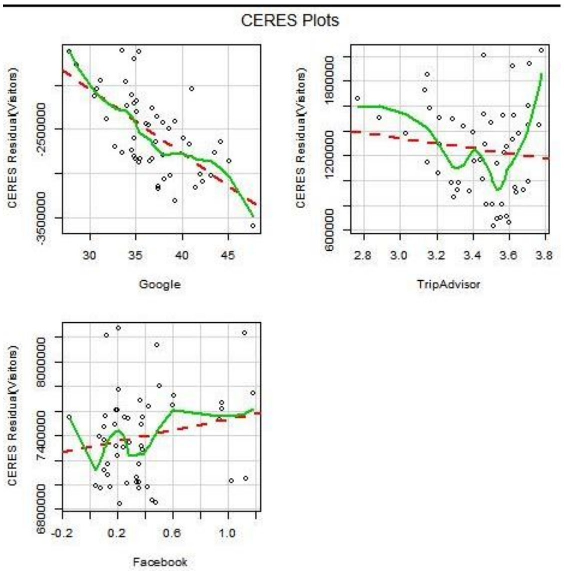


Figure 9: Linear Relationship Between the Independent Variables Used Together and The UK Visitor Data

residuals were plotted and the results are in Figure 10. A Durbin-Watson statistic of 2.1 backs up the residuals plot and shows there is no autocorrelation in the data.

The heteroscedasticity value of 1.7 and p-value of 0.19 show the data does not violate the rules of homoscedasticity.

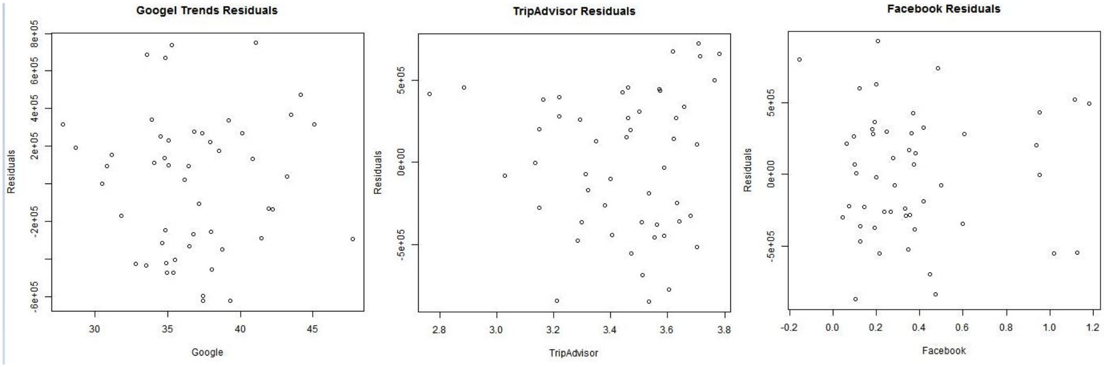


Figure 10: Residuals Plot for the Independent Variables

5.3 Discussion

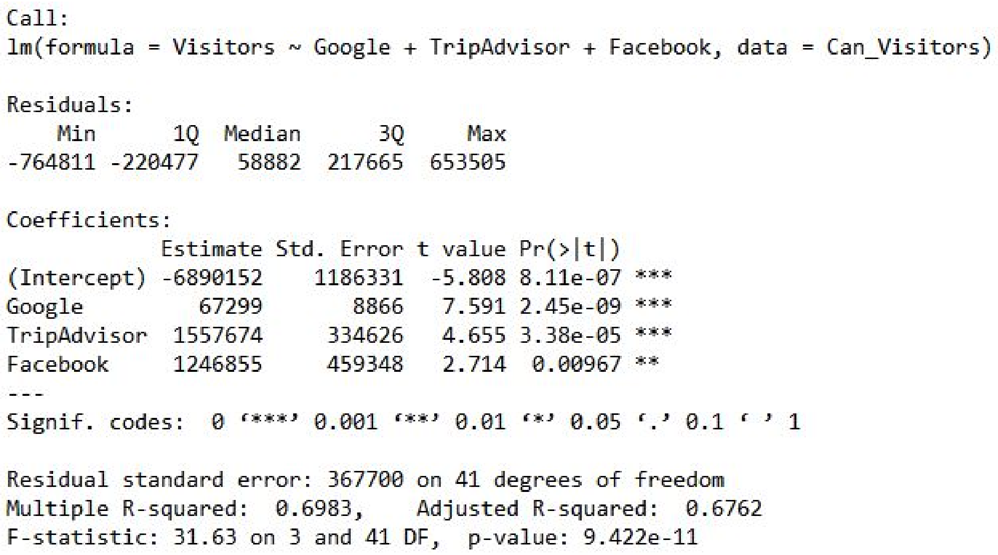
5.3.1 Canada

The paper is examining the use of the three independent variables as a predictor for visitor numbers and the R-squared values of the results for Canada (Table 3) show using the three independent variables (G+T+G) together in a multiple regression model is significantly better than using any of them on their own. Even though the Google Trends score has a high level of correlation, the R-squared score shows that very little of the model can be attributed it when used on its own. Adding up the values of the three independent variables (Total) is marginally better than using them individually with an R-squared score of .35 to the Google Trends score of .32, the G+T+F score of .69 is nearly twice as good as the Total score. Figure 11 shows the residual scores and although the maximum and minimum error values are quite large, 50% of the errors are between +/-200,000 visitors and this is quite acceptable as some months have over two million visitors. These scores are much better than the three independent variables or the Total score, except for the median error for the Total score which is only -5257, however all of the other results show this is not a good model to use.

Figure 11 also shows the coefficient scores for the G+T+F model which show that Google Trend and TripAdvisor scores have a significance level of 0 showing they are extremely unlikely to be unrelated to the visitor data and Facebook has a significance level of 0.001, again showing how unlikely it is not to be related to the model.

Multiple regression allows a test to be carried out using an interaction between the three independent variables rather than the three together in the G+T+F model above and although this interaction provides a higher correlation score of .88, R-squared score of 0.7475 and a marginally better adjusted R-squared score of 0.6998, the coefficients do not show any significance scores less than .11 meaning that the variables maybe be unrelated to the model and thus not a good model to use to predict the visitor data.

Overall the G+T+F model of multiple regression is the best predictor of the Canadian visitor numbers.

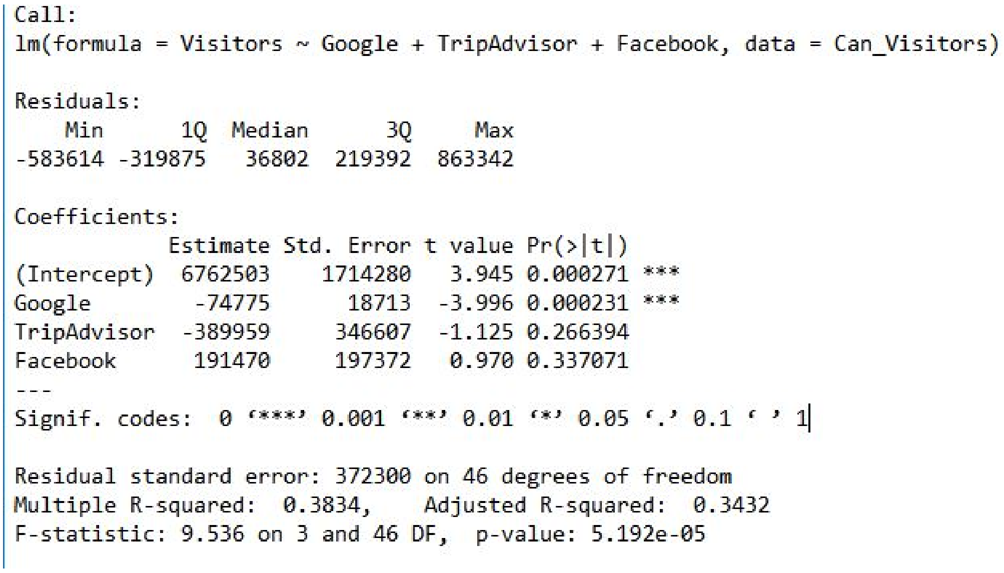


```
Call:
lm(formula = Visitors ~ Google + TripAdvisor + Facebook, data = Can_Visitors)

Residuals:
    Min      1Q  Median      3Q     Max 
-764811 -220477   58882  217665  653505 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6890152    1186331  -5.808 8.11e-07 ***
Google         67299       8866   7.591 2.45e-09 ***
TripAdvisor  1557674     334626   4.655 3.38e-05 ***
Facebook     1246855     459348   2.714  0.00967 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 367700 on 41 degrees of freedom
Multiple R-squared:  0.6983,	Adjusted R-squared:  0.6762 
F-statistic: 31.63 on 3 and 41 DF,  p-value: 9.422e-11
```

Figure 11: Canadian Residuals and Coefficient Scores Using the HGI Lexicon

5.3.2 The UK

For the UK results (Table 4) the R-squared score the G+T+F model is better than the individuals and the Total scores although in this case it is only marginally higher than the Google Trends score and the G+T+F adjusted R-squared score is .0037 lower, the adjusted r-squared score is not as relevant for Google Trends as it's an error correction measure to compensate for adding in additional variables in multiple regression.



```
Call:
lm(formula = Visitors ~ Google + TripAdvisor + Facebook, data = Can_Visitors)

Residuals:
    Min      1Q  Median      3Q     Max 
-583614 -319875   36802  219392  863342 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6762503    1714280   3.945 0.000271 ***
Google        -74775      18713  -3.996 0.000231 ***
TripAdvisor  -389959     346607  -1.125 0.266394    
Facebook      191470     197372   0.970 0.337071    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 372300 on 46 degrees of freedom
Multiple R-squared:  0.3834,	Adjusted R-squared:  0.3432 
F-statistic: 9.536 on 3 and 46 DF,  p-value: 5.192e-05
```

Figure 12: Canadian Residuals and Coefficient Scores Using the OL Lexicon

Figure 12 shows the residuals with a median error for the G+T+F model of only -36,802 which although Facebook has a median error of only -128 all of its other values are worse than the G+T+F model. This time only Google Trends shows a significance score of 0 it is still the best model available. Again the combined model has a higher R-squared score of 0.464 and adjusted R-squared of 0.3746, however, again none of the variables show any significance scores against the visitor data.

5.3.3 Overall

It has been shown that the best model for predicting visitor numbers is using the three independent variables in a multiple regression model.

The model does not perform as well for the UK as it does for Canada and this is due to the reduced sample size of the UK data is half that of Canada (5% and 10% respectively). Google Trends also returned less data for the UK than for Canada indicating that there may have been a lot of smaller non-chain hotels in the UK sample or people looked up an area rather than a specific hotel. A larger percentage sample would even out these differences.

There were many ways the sentiment analysis data could have been used and this study used the average score per month rather than the total score as this prevented one or two heavily reviewed hotels from unduly influencing the results. Also there was a huge increase in the number of reviews from 2010 to 2015 resulting in an increase in total scores over the years from several hundred in 2010 to several thousand in 2015.

The difference in time lags between (Table 1 and Table 2) Canada and the UK for people looking up Google etc., and visiting the country can be explained by the different time frames people use to book trips. The lag in Google Trends for Canada was 12 months, however a 1-month lag had nearly the same correlation. This could be due to large numbers of people in the highly populated areas of the USA (which makes up a significant percentage of Canada's tourism) such as New York or Chicago deciding to take a break at short notice. The twelve-month lag would be from visitors from further afield such as the UK or Japan which would incur greater time and expense in travelling and would thus take longer to plan such trips. The shorter time lags for TripAdvisor and Facebook are also due to how people plan their main holiday, starting off with Google and narrowing choice of location, hotels etc. nearer the departure data. This is more prevalent as people book flights and hotels over the Internet rather than using a travel agent.

Using multiple lexicons did not significantly enhance the process as the average score for both the OL and HGI lexicon was the same for the UK and the combined score for both for Canada was lower than the HGI lexicon score. However, using a larger representative sample have changed these results.

6 Conclusion and Future Work

6.1 Conclusion

This study set out to investigate the best way of using Google Trend, TripAdvisor and Facebook data to predict the visitor numbers to Canada and the UK as there was a gap in the existing research with regards to using Internet search and sentiment analysis data in the same prediction model. It established that using multiple regression on the three independent variables was the best method to achieve this goal. Simply adding the scores was shown to be worse than using multiple regression and although combining the data gave a better R-squared, the variables were not shown to have any significance on the visitor data in the model.

The size of the statistical sample is important with Canada having a 10% sample size showing much greater success in predicting visitors then the UK's 5% sample size.

Building a time lag into the model is important as although a month to month comparison provides better results it is impossible to know how much of the visits had already taken place during the month when the search/review took place and thus hindering the model predicting future visitor numbers.

6.2 Future Work

This research could be further developed by using different models such as Artificial Neural Networks (ANNs) as an alternative to multiple regression. Time series analysis could also be carried out to investigate if the pattern of visitors is a better indicator of future visitor numbers than Internet Search and sentiment analysis.

While the benefit of using two lexicons has proved inconclusive using several more could be of benefit, epically using an emoticon lexicon. Although only a hand full of TripAdvisor reviews and several hundred Facebook comments for Canada contained emoticons nearly a 65% of the Facebook reviews for the UK contained them. This would partially explain the poorer performance of the UK model and using an emoticon lexicon could greatly enhance sentiment analysis of social media. The sentiment analysis only examined positive and negative words and could be expanded to take the amount of likes a comment received into account e.g. if a comment received 4 likes its score could be multiplied by 4 as it resonated with readers. The score could also be refined by using a multiplier in connection with the emotions available in the HGI lexicon e.g. if a positive word had s strong emotion attached, its value could be increased.

This study only examined a sample of hotel reviews and could be expanded by using a much bigger statistical sample of hotels and extending the analysis to other information such as visitor attractions as holiday makers are more likely to base their choice of destination based on what an area has to offer and then look for accommodation.

Many other factors influence people's choice of destination such as the strength of the economy and currency fluctuations. A majority of visitors to Canada are from the US and an increase in the value of the US Dollar over the Canadian Dollar will allow Americans to buy more in Canada so there should be an increase in visitors. Conversely higher unemployment in the US should see a reduction in visitors to Canada. This economic data could be built into the model to examine if it improved its accuracy.

With both countries hosting the Olympics within the time frame of the study (Vancouver in 2010 and London in 2012) their impact could be studied as one of the main benefits of hosting the Olympics is claimed increased visitor numbers.

Acknowledgements

I would like to thank my supervisor Mr. Tony Delaney for his help, support and advice throughout the project.

References

Asur, S. and Huberman, B. A. (2010). Predicting the Future with Social Media, *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent*

Technology (WI-IAT), Vol. 1, pp. 492–499.

Barbera, P., Piccirilli, M. and Geisler, A. (2016). Rfacebook: Access to Facebook API via R.
URL: *https://cran.r-project.org/web/packages/Rfacebook/index.html*

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market, *Journal of Computational Science* **2**(1): 1–8.
URL: *http://www.sciencedirect.com/science/article/pii/S187775031100007X*

Choi, H. and Liu, P. (2011). Reading tea leaves in the tourism industry: a case study in the Gulf oil spill, *Available at SSRN 1893078* .
URL: $http://papers.ssrn.com/sol3/papers.cfm?abstract_id = 1893078$

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends, *Economic Record* **88**: 2–9.
URL: *http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/abstract*

Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A. and Peipins, L. A. (2005). Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003, *Journal of Medical Internet Research* **7**(3): e36.
URL: *http://www.jmir.org/2005/3/e36/*

Curme, C., Preis, T., Stanley, H. E. and Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves, *Proceedings of the National Academy of Sciences* **111**(32): 11600–11605.
URL: *http://www.pnas.org/content/111/32/11600*

Ettredge, M., Gerdes, J. and Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics, *Commun. ACM* **48**(11): 87–92.
URL: *http://doi.acm.org/10.1145/1096000.1096010*

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and others (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework., *KDD*, Vol. 96, pp. 82–88.
URL: *http://www.aaai.org/Papers/KDD/1996/KDD96-014*

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W., Winsemius, D., Zeileis, A. and R-Core (2016). car: Companion to Applied Regression.
URL: *https://cran.r-project.org/web/packages/car/index.html*

Godbole, N., Srinivasaiah, M. and Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs., *ICWSM* **7**(21): 219–222.
URL: *http://www.uvm.edu/ pdodds/files/papers/others/2007/godbole2007a.pdf*

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. and Watts, D. J. (2010). Predicting consumer behavior with Web search, *Proceedings of the National Academy of Sciences* **107**(41): 17486–17490.
URL: *http://www.pnas.org/content/107/41/17486*

Google-Trends-Excel-Template-Convert-Weekly-Data-to-Monthly.xlsx (n.d.).
URL: *http://connect.icrossing.co.uk/wp-content/uploads/2013/04/Google-Trends-Excel-Template-Convert-Weekly-Data-to-Monthly.xlsx*

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives, *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 174–181.
URL: *http://dx.doi.org/10.3115/979617.979640*

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 168–177.
URL: *http://dl.acm.org/citation.cfm?id=1014073*

Hutchin, E. (2016). Adopted from RIC Research proposal, NCI.

Jeacle, I. and Carter, C. (2011). In TripAdvisor we trust: Rankings, calculative regimes and abstract systems, *Accounting, Organizations and Society* **36**(45): 293–309.
URL: *http://www.sciencedirect.com/science/article/pii/S0361368211000420*

Lantz, B. (2013). *Machine Learning with R*, Packt Publishing.

Liu, B. (2010). Sentiment Analysis and Subjectivity., *Handbook of natural language processing* **2**: 627–666.
URL: *http://www.crcnetbase.com/doi/pdf/10.1201/9781420085938-c26*

Massicotte, P. and Eddelbuettel, D. (2016). gtrendsR: R Functions to Perform and Display Google Trends Queries.
URL: *https://cran.r-project.org/web/packages/gtrendsR/index.html*

Oliveira, N., Cortez, P. and Areal, N. (2013). Some Experiments on Modeling Stock Market Behavior Using Investor Sentiment Analysis and Posting Volume from Twitter, *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, ACM, New York, NY, USA, pp. 31:1–31:8.
URL: *http://doi.acm.org/10.1145/2479787.2479811*

Ortigosa, A., Martn, J. M. and Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning, *Computers in Human Behavior* **31**: 527–541.
URL: *http://www.sciencedirect.com/science/article/pii/S0747563213001751*

Preis, T., Moat, H. S., Stanley, H. E. and Bishop, S. R. (2012). Quantifying the Advantage of Looking Forward, *Scientific Reports* **2**.
URL: *http://www.nature.com/articles/srep00350*

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research.
URL: *https://cran.r-project.org/web/packages/psych/index.html*

R: The R Project for Statistical Computing (n.d.).
URL: *https://www.r-project.org/*

tourismsnapshot-dec2015_en-lowres (n.d.).
URL: *http://en.destinationcanada.com/sites/default/files/pdf/Research/Stats-figures/International-visitor-arrivals/Tourism-monthly-snapshot/tourismsnapshot-dec2015*$_e n - lowres.pdf$

Wickham, H. and RStudio (2016a). rvest: Easily Harvest (Scrape) Web Pages.
URL: *https://cran.r-project.org/web/packages/rvest/index.html*

Wickham, H. and RStudio (2016b). stringr: Simple, Consistent Wrappers for Common String Operations.
URL: *https://cran.r-project.org/web/packages/stringr/index.html*

Xiang, Z. and Gretzel, U. (2010). Role of social media in online travel information search, *Tourism Management* **31**(2): 179–188.
URL: *http://www.sciencedirect.com/science/article/pii/S0261517709000387*

Yadav, M. P., Feeroz, M. and Yadav, V. K. (2012). Mining the customer behavior using web usage mining in e-commerce, *2012 Third International Conference on Computing Communication Networking Technologies (ICCCNT)*, pp. 1–5.