Stony Brook University
Department Of Computer Science

CSE 545

BIG DATA ANALYTICS

Team Name - Bella Ciao
Prayag Goyal - 112961923
Srinivas Kandari - 112713946
Dheeraj Ramchandani - 112970301
Rohit Chaudhary - 112687906

Report Date: May 20, 2020

Introduction

The project, Economy and Tourism, is developed to predict the footfall in a region based on the Twitter data and Google Trends related to that region. This is important because tourism has a substantial impact on the economy and therefore, it would be helpful to know the number of visitors in a region beforehand. The economic aspect of the travel industry is often assumed to be related to only core hospitality and transportation services and hence, undermined. On the contrary, the importance of tourism extends well beyond these industries.

We achieved our goal by performing data analysis on textual data and on numerical data that we gathered from Twitter and Google Trends, respectively. The Twitter data is selected on the basis of names of the popular tourist places of a country/region. We have done real time analysis on the Twitter data and travel related trending words to find the correlation between this data and the number of tourists coming in at that particular location. The information regarding the number of tourists is obtained using Google Trends.

SDG: The objective of our project directly relates to SDG # 8. This SDG aims at implementing policies to promote sustainable tourism that creates jobs and promotes local culture and products as well. Our analysis will help achieve this goal by enabling governments and authorities to make informed decisions related to sustainable development of tourist hotspots to attract more people and hence, boost the economy of the region.

Background

There have been studies in the past to establish a correlation between tourism in a place and the economy of that place. One such study was carried out by World Travel Tourism Council's (WTTC) Economic Impact Research in 2018 which said that the travel industry alone employed 7 million new people and the growth of this sector was pegged at 4.6%. Another related study was done by the United Nations World Tourism Organization UNWTO in the same year which said that yearly international arrivals have increased 56 folds from 1950 to 2018, and so it creates the need to have a very well documented set of data and analysis to understand the growth and development of a region as tourism is playing a major role in economic growth in most of the countries around the globe.

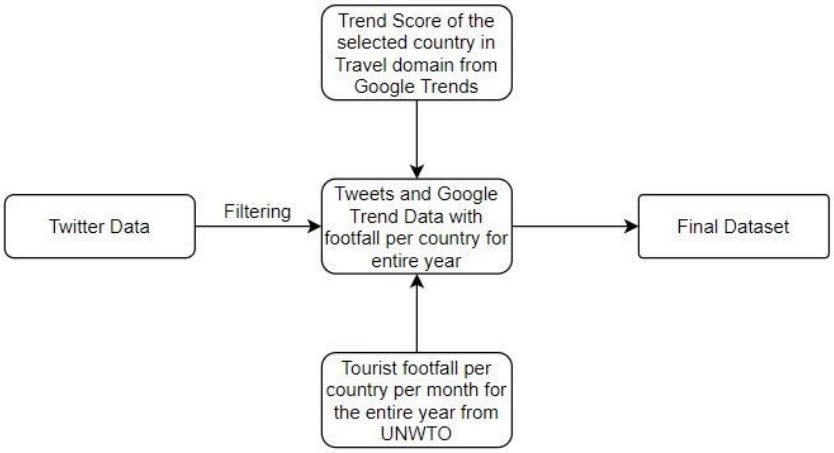
Apart from the above mentioned studies, [2] and [3] relate to economy and tourism and suggest that tourism directly contributes to the GDP for a country and helps in creating employment. It is also shown that the individual spending on tourism is also increasing. It provides a validation for us to focus on sustainable tourism for providing jobs and boosting the local businesses which therefore helps in rising the economy.

Data

We have used both textual and numerical data for the project. For textual data we have used tweets from Twitter and Google trend score from Google trends along with tourist footfall numbers from UNWTO which is the World Tourism Organisation.

Twitter: For data preparation, we used the 'GetOldTweets3' library in python to collect textual data from Twitter for more than 250 days. The data for each day was around 1.8 GB in size therefore we collected Twitter data filtered on certain keywords related to tourist attractions, country names and regions. The filtered data was around 15 GB in size. We picked mainly 5 labels/fields from each tweet text like userID, userLocation, domainType, tweetText and tweetDate and time. Filtering the tweets on the basis of location was very helpful as location is visible in only around one percent of the tweets.

Google Trends: We then collected the Google trend scores from Google trends for a particular selected country for each of the dates using the 'pytrends' library and merged it with the tourist footfall numbers to create the final dataset. In line with Tweets, here also the Trends were collected with default location as 'World' and domain number 67 which is related to 'Travel' industry.



Figure[1]

Each row in the final dataset is the combination of country and date, with columns as words extracted from tweet text combined with trend score, and resultant vector is the footfall vector. The final dataset was pre-processed to remove punctuation, stop words, emoji and website URLs and words were vectorized using TF-IDF to create feature columns. One important point was to divide the footfall numbers by length of the month the tourist footfall data was present on a per month basis and we predicted the future footfall on a per day basis.

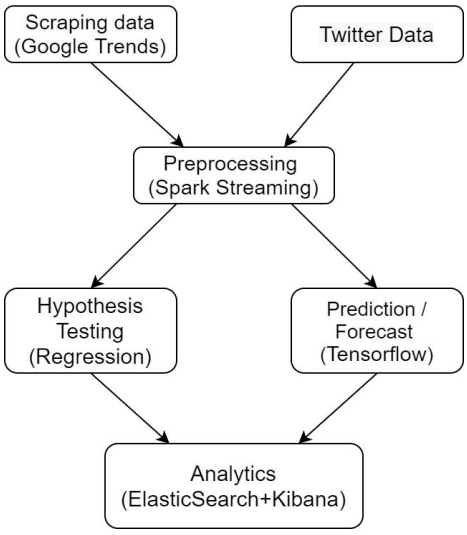
Method

Architecture Diagram:

As you can see from figure 2, all input data from Twitter is streamed through spark streaming where it is pre processed along with joining with Google Trends and Footfall dataset. After preprocessing the data is being sent for Hypothesis testing and forecasting. The results from hypothesis testing and forecasting are stored in Elasticsearch and visualized in Kibana. The two data pipelines which we are using here are :

1) Spark Streaming: Get input from all data sources (twitter, google trends) and preprocess and output to Tensorflow.

2) Analytics pipeline: Get input from Hypothesis testing and Prediction from Tensorflow and output to Elasticsearch and Kibana stack for real time visualizations.



Figure[2]: Architecture Diagram

Spark Streaming:

A Spark streaming flow has been established to receive tweets filtered on the basis of tourist destinations of different countries. The streaming module of the library Tweepy is used to connect with Twitter API over TCP/IP and capture tweets based on keywords. This stream of tweets is then passed onto Spark streaming module where it is further preprocessed to remove stopwords, punctuations, and URLs. The required fields from preprocessed data are then fed to further pipeline for Hypothesis Testing and Forecasting.

Hypothesis Testing:

We find the top 1000 words coming in the tweets for a particular country according to the overall word count . Then for these words we find the correlation with the footfall on a per day basis. For this we run multiple linear regressions for word counts of each word for a particular day to find the top positive and negative correlated words with footfall.

Forecasting:

We predict the footfall of tourism for future dates using a Neural Network model, for this regression task the features used are: Date, trend_score(from google trends) and word counts (words are output of hypothesis testing) and predict the footfall.
Initially, we trained Neural Network with many number of neurons (Layer 1- 75, act = ‘relu’; Layer 2- 25, act=’relu’) but this network didn’t perform well because there were too many parameters(weights) and few days(records) to learn from and the parameters were not optimised(as gradient descent updates parameters in tiny steps i.e, learning rate). We ended up using a modest Neural network with two deep layers(Layer 1- 10, Layer 2- 10); for which the weights optimised quickly. We trained the model with data from 228 days(June 2019 to Feb 2020) for 50 Epochs.

Analytics using Elasticsearch:

Results from hypothesis testing and forecasting models are being stored in Elasticsearch. Elasticsearch provides filters for date, country and all other fields. Real time visualizations can be viewed on Kibana which provides a dashboard for all data stored in elasticsearch. The forecasting model takes input from hypothesis testing which is queried from elasticsearch.

Results

Hypothesis Testing:

For hypothesis testing, we try to visualize it as a word cloud. For each country and time period we provide a word cloud for top positive correlated words.

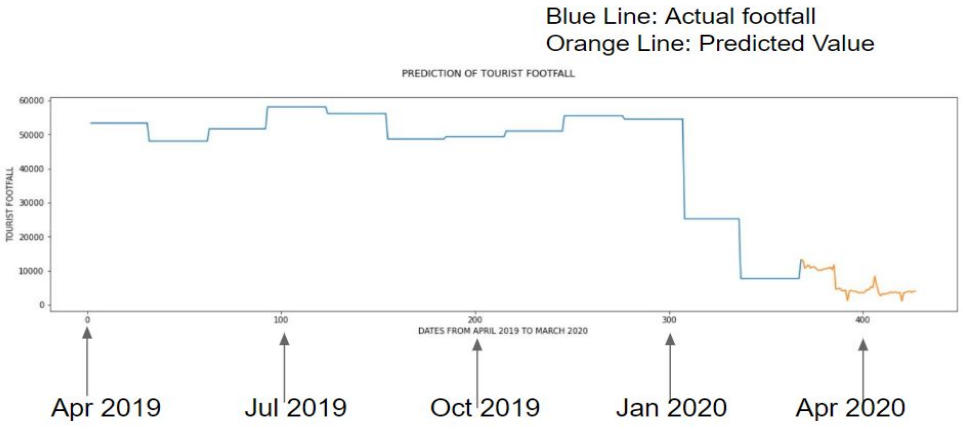
Figure[3]

As you can see in figure 3, there are few tourist destinations in the top positive correlated words with footfall in Singapore. These include Sentosa island, Clarke quay, Garden by the bay, Cloud forest etc which can be used as a good feature in predicting the footfall for Singapore.

Forecasting:

We have evaluated the predictions of neural Networks for the dates(In february 2020) for which we have observed values and found the RMSE being 3225. Considering the whimsical trends in tourism(due to corona pandemic) for these dates the observed RMSE can be classified as decent. As seen in Figure[4], the blue curve depicts the observed footfall in the respective months and the orange curve depicts the predicted footfall for the month April. It can be observed that the model has done a pretty good job at learning the underlying trend.

RMSE = 3225.753
Epochs = 50



Figure[4]

In figure[5] below, we have predicted the footfall for March 2020 and observed that the footfall prediction in the beginning of month was around 13000(which is expected). But for the dates after 16th march 2020 the predictions have fallen 3000-4000. As per the working of Neural Network it's features, i.e, the word counts for these were less as these words have less frequently appeared on twitter and these predictions are reasonable because many countries implemented lockdown from mid march. This actually proves our hypothesis that the words we had are pretty good features in predicting the footfall.

	date	predicted
0	2020-02-24	[13209.11]
1	2020-02-25	[13105.198]
2	2020-02-26	[10694.562]
3	2020-02-27	[11376.816]
4	2020-02-28	[11636.977]
5	2020-03-01	[10781.032]
6	2020-03-02	[11188.272]
7	2020-03-03	[11127.029]
8	2020-03-04	[10567.87]
9	2020-03-05	[10040.656]
10	2020-03-06	[10163.96]
11	2020-03-07	[10271.81]
12	2020-03-09	[10597.245]
13	2020-03-10	[10535.45]
14	2020-03-11	[10739.677]
15	2020-03-12	[11028.929]
16	2020-03-13	[10355.006]
17	2020-03-14	[11729.005]
18	2020-03-16	[4567.8286]
19	2020-03-17	[4680.124]
20	2020-03-18	[4931.3804]
21	2020-03-19	[4170.5684]
22	2020-03-20	[4181.454]
23	2020-03-21	[4274.2754]
24	2020-03-22	[1238.3114]
25	2020-03-23	[4128.469]
26	2020-03-24	[4185.9834]
27	2020-03-25	[3957.2964]
28	2020-03-26	[4029.403]
29	2020-03-27	[3709.3135]
30	2020-03-28	[3494.8513]

Figure[5]

Though the model did a pretty good job at predicting the footfall for future dates. The training could have been much more effective if we had trained it on more data. Also, the data has a significant number of outliers for dates pertaining to the year 2020 as the covid pandemic took place at this period and it disturbed the regular trend of tourism.

Conclusion

- Our project will help authorities in decision making for devising policies related to regulate sustainable tourism in a country.
- The top words analysis can also be used to generate popular hashtags/content on social media attracting large numbers of tourists.
- The forecasting model can help them to manage resources(Local businesses, jobs) for the influx of the tourist footfall.

Future work

- Other social media data can be used. For example: Facebook check ins or location based instagram posts.
- Major Event calendars can also be integrated to enhance the forecasting outcome.

References

1. Tourist footfall data taken from UNWTO: https://www.unwto.org/
2. “Tourism Economics Research: A review and assessment“ - https://www.sciencedirect.com/science/article/abs/pii/S0160738312000795
3. “Research on influence of Tourism on Economy” - https://academic.oup.com/ijlct/article/14/2/241/5430500
4. Tweets filtered on the basis of tourist attractions and for a selected time-frame taken from: https://github.com/marquisvictor/Optimized-Modified-GetOldTweets3-OMGOT
5. Google trends data collected using pytrends library with help from: https://medium.com/intro-to-python-wows/google-trends-4db836214868
6. Tweepy and Spark Streaming https://medium.com/@distillerytech/bringing-big-tools-to-big-data-spark-and-spark-streaming-ed93f5b478d7

Figures

1. Data preparation flow diagram.
2. Architecture Diagram
3. Hypothesis testing results as a word cloud of top correlated words.
4. Footfall forecasting output as a result of regression.
5. Predicted footfall as an output of forecasting model.