

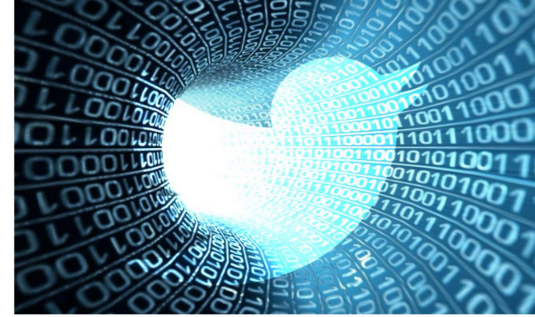
Economy and Tourism

Big Data Analytics (CSE 545)
Stony Brook University

By:
Rohit Chaudhary
Dheeraj Ramchandani
Prayag Goyal
Srinivas Kandari

INTRODUCTION

- The objective of this project is to predict the tourism in a particular country and its effects on the economy with the help of textual data analysis using Google trends and social media such as Twitter.
- We will do real time analysis of the tweets and trending words related to travel and check if they are correlated to the number of tourists coming in, number of people employed in tourism related industry.



INTRODUCTION



MOTIVATION:

- 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 the tourism extends well beyond these industries.
- Our analysis will help governments and authorities in decision making related to sustainable development of tourist hotspots to attract more people and hence, boost the economy of the region.

How does it relate to SDG?

Proper data analysis of flow of tourists and identifying trends helps in maximizing the contribution of tourism to poverty reduction and by making tourism work as a tool for development and promoting the inclusion of tourism in the development agenda. Effective correlation and prediction of travel trends will help to devise and implement policies to promote sustainable tourism that creates jobs and promotes local culture and products as well.

BACKGROUND

- According to a World Travel Tourism Council's (WTTC) Economic Impact Research in 2018:
 - ❑ Travel & Tourism created 7 million new jobs worldwide.
 - ❑ The growth of the sector was pegged at 4.6%.
- An assessment by the United Nations World Tourism Organization (UNWTO) in 2018 estimated that there has been a 56-fold increase (25 million to 1.4 billion) in international arrivals per year data from 1950 to 2018.
- There are various direct and indirect impacts of tourism on economy:
 - ❑ Direct Impact - lodging, restaurants, transportation, amusement and retail trade.
 - ❑ Indirect Impact - Change in sales, income and employment of a region.

Therefore, 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.

BACKGROUND

TOURISM ECONOMICS RESEARCH: A REVIEW AND ASSESSMENT¹

Haiyan Song
The Hong Kong Polytechnic University, Hong Kong
Larry Dwyer
University of New South Wales, Australia
Gang Li
Zheng Cao
University of Surrey, United Kingdom

ABSTRACT

This paper aims to provide the most up-to-date survey of tourism economics research and to summarise the key trends in its recent development. Particular attention is paid to the research progress made over the last decade in respect of approaches, methodological innovations, emerging topics, research gaps, and directions for future research. Remarkable but unbalanced developments have been observed across different sub-research areas in tourism economics. While neoclassical economics has contributed the most to the development of tourism economics, alternative schools of thought in economics have also emerged in advancing our understanding of tourism from different perspectives. As tourism studies are multi- and inter-disciplinary, integrating economics with other social science disciplines will further contribute to knowledge creation in tourism studies.

<https://www.sciencedirect.com/science/article/abs/pii/S0160738312000795?via%3Dihub>

Research on Infulence of Tourism on Economy

Dan Luo and Wuzhong Zhou

Institute of Tourism and Landscape Architecture, School of Arts, Southeast University,
Nanjing 210096, China
dana_luo@126.com

Abstract. Tourism can enhance economic growth and can also benefit from the growth. In this paper, in order to identify the mode of tourism-economy relationship, a multi-dimensional time series of both tourism and economy was analyzed with principal component analysis (PCA). Our results revealed that, in 1989, most of the European countries shared a similar mode, in which tourism had a great contribution to both GDP and employment, and the capital investment was at a high level. In 2000, the developing countries advanced so much that their tourism-economy mode was similar to that of the developed countries. In 2011, the modes diversified. The United States showed a distinct mode. The mode of China gradually resembles that of the United States. In China, the government and individual tourism spending is increasing.

<https://academic.oup.com/ijlct/article/14/2/241/5430500>

Few research is done on Economy and Tourism already, which suggests that tourism is increasing as a part of GDP for countries. 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 economy.

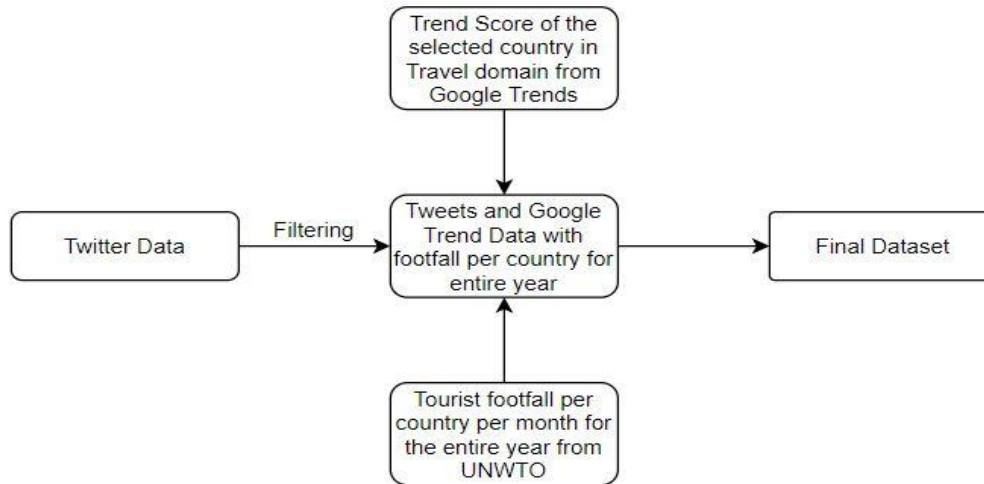
DATASETS USED

- Data was scraped from Twitter(Tweepy) and Google trends and combined to form the dataset.
- Tourist footfall per country per month:
<https://www.unwto.org/>
- Monthly tweets from Twitter from Twitter archive without filtering:
<https://archive.org/search.php?query=twitterstream&sort=-publicdate>
- All tweets filtered on the basis of tourist attraction and within a specified time frame using following API:
<https://github.com/marquisvictor/Optimized-Modified-GetOldTweets3-OMGOT>
- We used pytrends library in Python to get ranking of a word from Google trends:
<https://medium.com/intro-to-python-wows/google-trends-4db836214868>



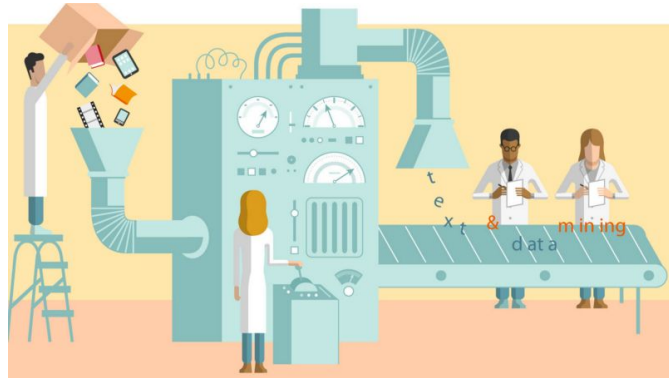
DATA PREPARATION

- We used the 'GetOldTweets' library to download tweet texts for more than 250 days. The tweet data for each day was around 1.8 GB, hence we filtered the data on keywords related to countries, tourist attractions and regions.
- This filtered tweet data is pushed into Elasticsearch.
- We then merged the textual Twitter data with numerical tourist footfall and Google trend score of that country on that date.
- We used the library pytrends to extract Google trend score.



DATA PREPROCESSING

- Data was preprocessed to remove punctuation and stop words. Special characters such as emoji's and website urls were removed
- We obtained the daily footfall by dividing monthly footfall by 30
- Thereafter, words were vectorized (using TF-IDF) and converted to features.



```
> Apr 29, 2020 @ 16:01:59.000 tweet: ""This weeks #wheretogowednesday is right across the water from Umass Boston! Marina Bay is home to beautiful boardwalks and amazing restaurants such as Siros and Port 305. They also have small boutique shops as well! A great place to check out so close to UMB! pic.twitter.com/nP14oUeMgM"" date: Apr 29, 2020 @ 16:01:59.000 country: Singapore _id: mFv9H3IBsc1XRJ4jry7p _type: _doc _index: tweets _score: -
```

```
> Apr 29, 2020 @ 11:16:00.000 tweet: """"The Marina Bay Sands" date: Apr 29, 2020 @ 11:16:00.000 country: Singapore _id: nFv9H3IBsc1XRJ4jtC7e _type: _doc _index: tweets _score: -
```

```
> Apr 29, 2020 @ 11:15:06.000 tweet: ""RT @travelust_bible: Amazing view from Marina Bay Sands roof top pool - Singapore #Travel #Holiday #Inspiration #Luxury #Hotel #Singapore #Wanderlust #Explore #LoveTravel #TheTravelustBiblepic.twitter.com/D82bg4XZ3v"" date: Apr 29, 2020 @ 11:15:06.000 country: Singapore _id: nlv9H3IBsc1XRJ4jtS7h _type: _doc _index: tweets _score: -
```

```
> Apr 29, 2020 @ 10:56:34.000 tweet: ""au Marina Bay?"" date: Apr 29, 2020 @ 10:56:34.000 country: Singapore _id: n1v9H3IBsc1XRJ4jti6y _type: _doc _index: tweets _score: -
```

Original tweets being stored in Elasticsearch

ARCHITECTURE DIAGRAM

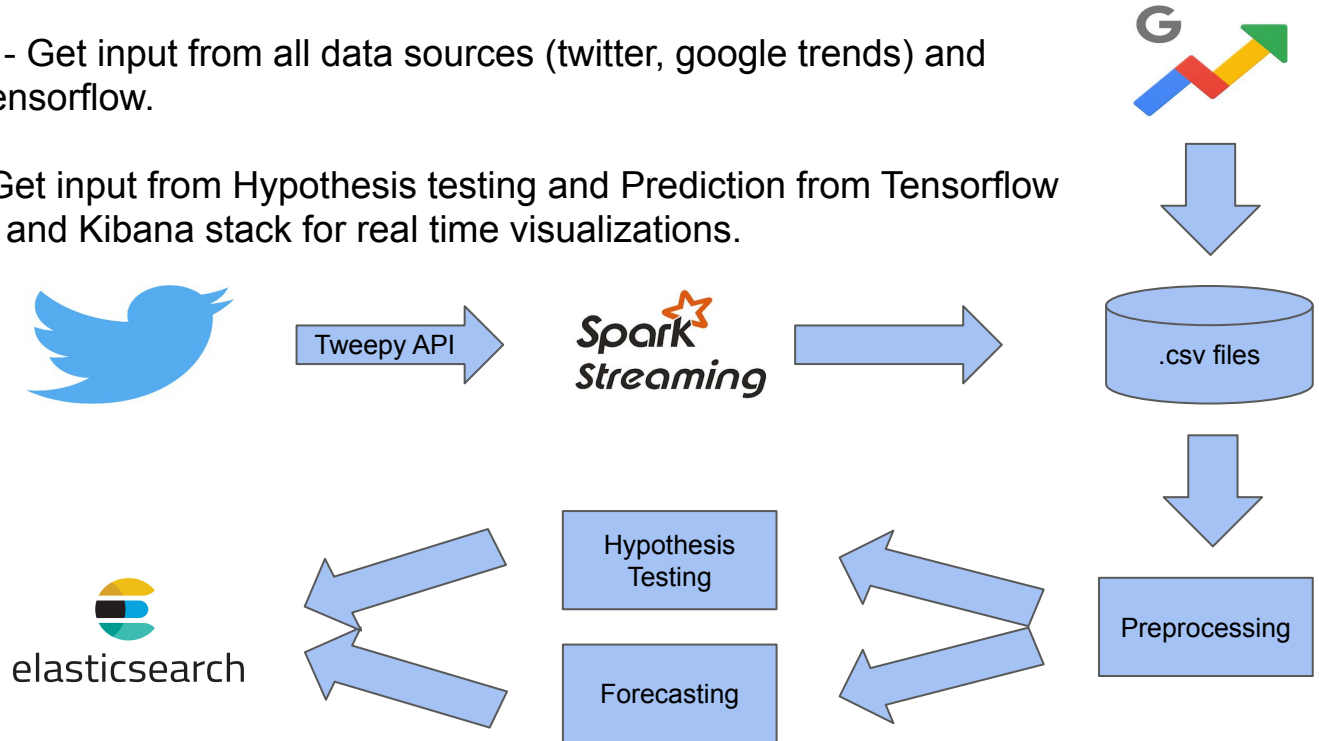
Data Pipeline:

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

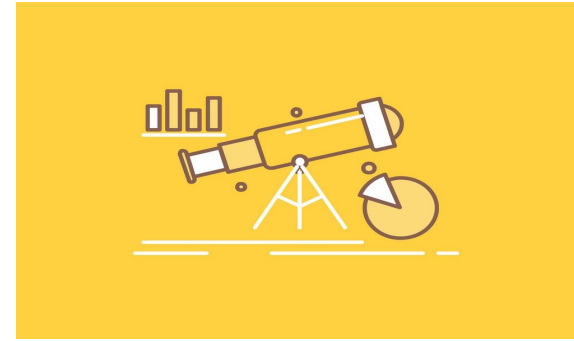
Data Analytics pipeline - Get input from Hypothesis testing and Prediction from Tensorflow and output to Elasticsearch and Kibana stack for real time visualizations.

Class concepts used:

- Hypothesis testing
- Linear Regression (Tensorflow)



METHOD



Hypothesis:

Out of all the tweets we take top 1000 more frequent words and find top 50 positively and negatively correlated words using linear regression.

Forecast :

- We query data from Elasticsearch for a specific country and use that data to train a two layer Neural Network. We use this model to predict the footfall for future dates.
- We also query streaming data from Elasticsearch. And use this new data to retrain the already trained model.
- We also have utility functions; `save_model` to save the model to disk and `load_model` to load the model from disk

TENSORFLOW TRAINING

Trials:

- Two layer deep learning model (Layer 1- 75, activation = 'relu', layer 2 - 25, activation = 'relu', epochs = 25)
- Same network as above without relu activation.

Problems:

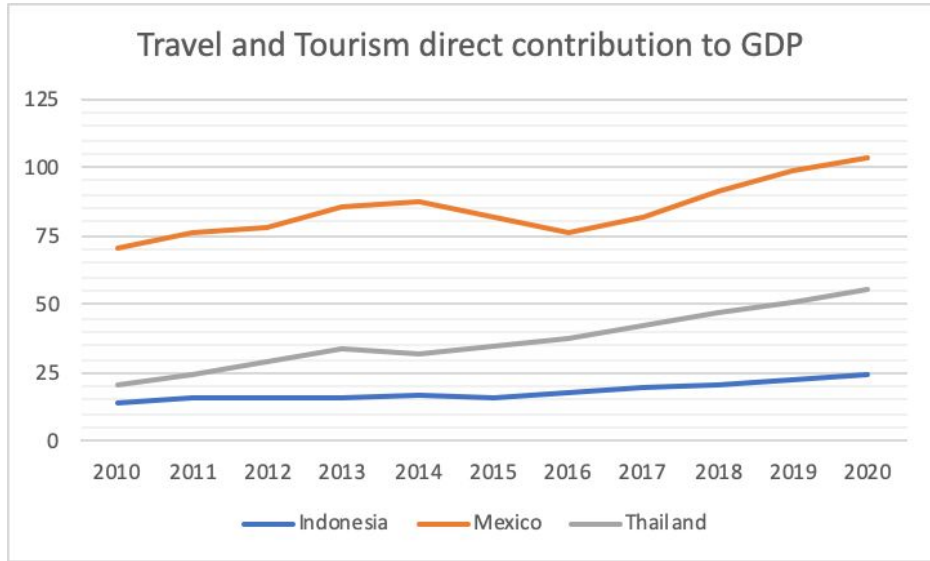
Too many parameters and only few dates to learn from. (didn't optimise)

Final architecture:

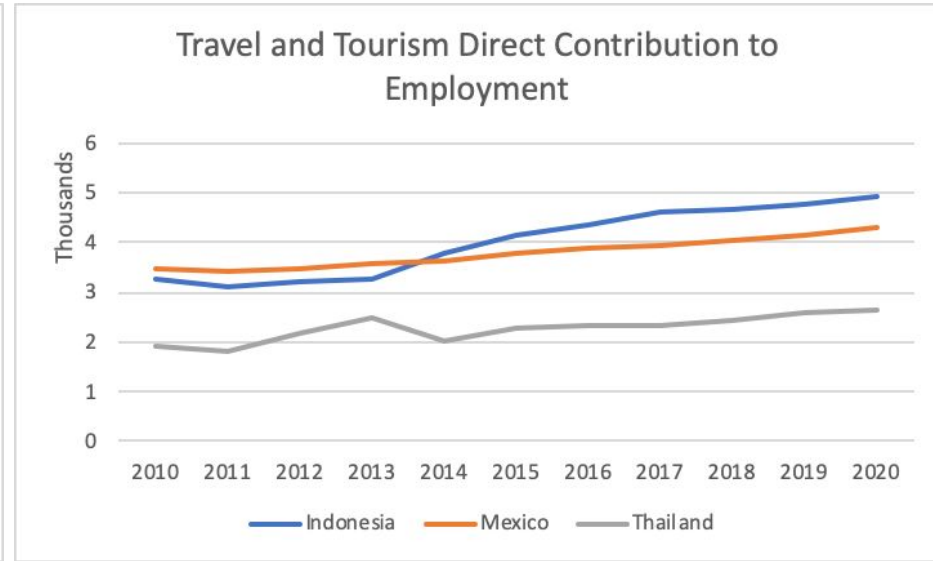
- Two layer deep learning model (Layer 1- 10, layer 2 - 10, epochs = 50)
- Weights optimised.

RESULTS

Background research papers linked Economy and Tourism, here are few visualizations to support it.



Time series graph for Travel and Tourism direct contribution to GDP from year 2010-2020



Time series graph for Travel and Tourism direct contribution to employment

HYPOTHESIS - WORD CLOUD



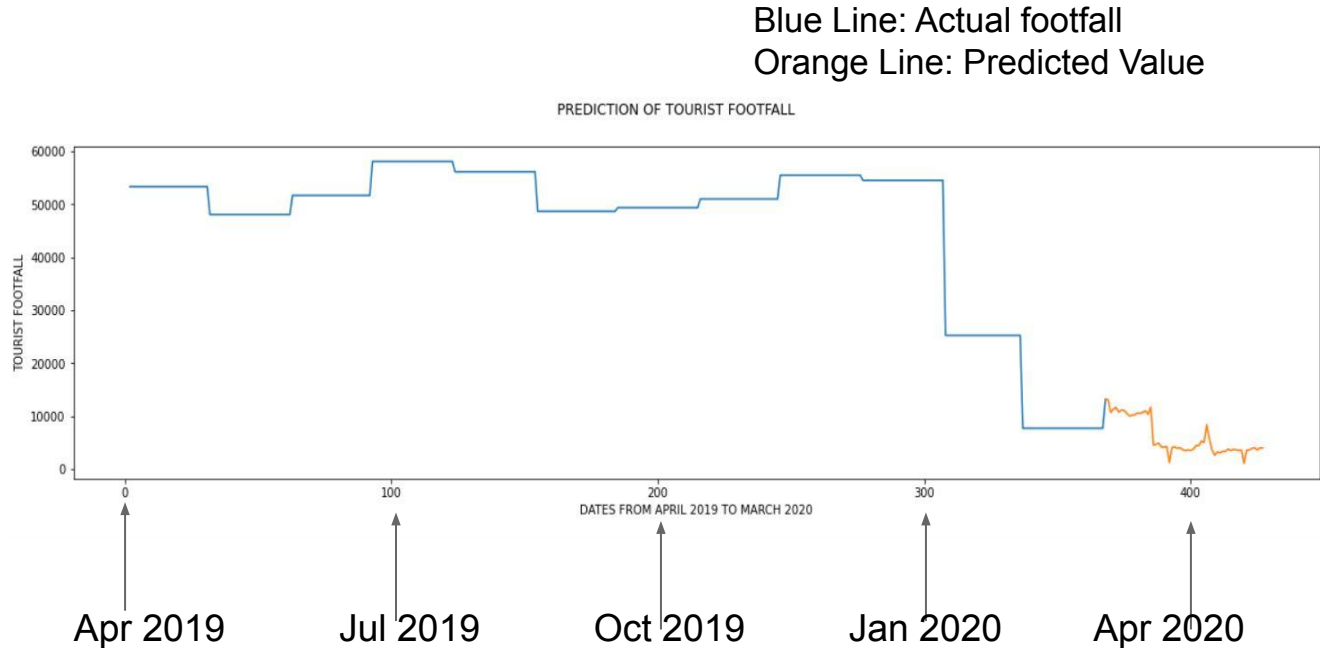
Contains few famous
tourist destinations
(marked with circles)

Word Cloud for most positive correlated words related to football numbers for Singapore

TIME SERIES FORECASTING - FOOTFALL

	date	predicted	observed
0	2020-02-12	[11620.953]	7738
1	2020-02-13	[11498.067]	7738
2	2020-02-14	[11379.776]	7738
3	2020-02-16	[10451.702]	7738
4	2020-02-17	[10462.803]	7738
5	2020-02-18	[9724.278]	7738
6	2020-02-19	[9359.856]	7738
7	2020-02-20	[10805.445]	7738
8	2020-02-21	[11397.614]	7738
9	2020-02-23	[11910.581]	7738


RMSE = 3225.753
Epochs = 50



TIME SERIES FORECASTING - FOOTFALL

Predicted values of footfall for Singapore

	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]



As we can observe the prediction for footfall for dates after lockdown has fallen down to 3000 - 4000, the reason being our words being mentioned less on twitter which proves our Hypothesis that the our word counts are good features in predicting footfall.

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 attracting large number of tourists.
- The forecasting model can help them to manage resources(Local businesses, jobs) for the influx of the tourist footfall.