A Comparative Study of Sentiment Expression on Twitter in Pune and Dubai

Aneri Patel[1] and Kunal Kulkarni[1]

[1]Center for Urban Science and Progress, New York University

December 15, 2021

Abstract

Tourism is one of the most important sources of economic activities in cities and a significant driver of the economy for many. While it contributes to 20% of the GDP for the small emirate of Dubai in the UAE, it contributes a minor fraction of 6.7% to the Indian GDP. It is crucial for governments to assess the impact of tourism-investment in these locations have on people around them - both tourists and residents. This raises the need to monitor the sentiment of visitors and residents both at tourist spots for developed and developing tourism locations using social media data. The explosion of available social media data allows for a more time-sensitive and geographically specific sentiment analysis than ever before. In this project, we perform a spatio-temporal sentiment analysis of data from the micro-blogging website Twitter to assess the general sentiment of both residents and visitors around these tourist destinations for the developed and developing (in the context of tourism) cities of Dubai and Pune respectively. This comparative study intends to identify differences in the patterns of sentiment distribution for both Dubai and Pune while also exploring any possible relationship between weather and these patterns.

1 Literature Review

Much of academic research on tourism remains grounded in economic analysis rather than the assessment of the socio-cultural impacts of tourism (Zaidan & Kovacs 2017). There is a wealth of information that hides behind unstructured social media data that can be used to analyze these impacts. The temporal and spatio-temporal effects on tourists' emotions when visiting a city's tourist destinations were studied by Padilla et al. However, there has been no concentrated effort to study the general public sentiment of both residents and tourists combined at tourist hotspots.

It is also interesting to identify the relationship between weather and public sentiment, especially at tourist hot-spots as it has already been established that all tourism destinations and operators are climate-sensitive to a degree Scott & Lemieux (2010).

2 Introduction

Traditionally, researchers rely on surveys and time-space travel diaries to gain an understanding of the trends of visitors' opinions. Sometimes, these data collection techniques are complemented with digital devices that track locations over time Padilla et al. (2018). While common, these techniques are costly, require active participation, and need to be repeated periodically to understand and measure changes over time Padilla et al. (2018). Considering that traditional survey methods are time-consuming and expensive, we need timely and proactive data sources to respond to the rapidly evolving dynamics of everyday life on the population's sentiment Valdez et al. (2020). Unstructured data from Twitter and other social media platforms represent a large and largely untapped resource for social data and evidence Ahmed (2017).Twitter remains the most popular platform for academic research, as it still provides its data via a number of Application Programming Interfaces (API) Ahmed (2017).

Dubai entices tourists with shopping opportunities and many other valuable, modern and ancient, attractions as well like its diverse culture and fundamental milestones symbolizing the Emirates' resurgence on the global arena. The Middle-Eastern hub has capabilities like the existence of infrastructure and modern means of communication and transportation, that enable it to tap into the potential of tourism to achieve economic growth *tourism* (n.d.). The Dubai tourism strategy of 2020, focused on attracting 20 million foreigners per year by 2020 Stephens et al. (2019). Tourism development strategies such as this one have repeatedly been successes, thus developing a tourism-dependent economy that was greatly deterred by CoVID-19.

Pune is a city that has now become a hot spot for business and leisure. Rapidly increasing purchasing power of the middle class is also one of the driving factors for the recent increase in tourism in Pune Deshpande & Deshpande (2016). It is dotted with historical temples, parks, forts, and caves that attract tourists from all over the country; thus attributing it with immense potential for a thriving tourism economy Deshpande & Deshpande (2016).

In this project, we study a) polarity of sentiment and b) frequency of expression on social media in cities and tourism hotspots. We further explore the influence of weather factors on these target features. With Pune lying on an elevated plateau and Dubai being a desert, a stark difference in sentiment towards precipitation is expected in both places with the population in Dubai having a positive outlook and that in Pune being negative. The study concluded that the average sentiment expressed on a rainy day in Dubai is 0.003 more than the average sentiment expressed on a normal day while it is 0.006 lesser in Pune.

3 Data

The TwitterSearchScrapper functionality of the snscrape package in python was used to scrape all English tweets within a 30km radius of the city center of Dubai (25.2048°N, 55.2708°E) and Pune (18.5204°N, 73.8567°E) between January 2019 and December 2021. The same functionality was used to specifically scrape tweets generated within a 0.5km radius of carefully selected tourist locations in both cities. The tourist destinations selected for Dubai are Burj Khalifa (25.1972° N, 55.2744°E), Burj Al Arab (25.1412°N, 55.1852°E), Dubai Marina (25.0805° N, 55.1403°E), Palm Jumeirah (25.1124°N, 55.1390°E), Dubai Museum (25.2635°N, 55.2972°E), and Dubai World Trade Center (25.2233° N, 55.2869°E). Selected hotspots for Pune are Dagdusheth Ganpati Temple (18.5164°N, 73.8561°E), Sinhagad Fort (18.3663°N, 73.7559°E), Shaniwar Wada Fort (18.5195°N, 73.8553°E), Aga Khan Palace (18.5525°N, 73.9015°E), and Khadakwasla Dam (18.4423°N, 73.7671°E).

A total of 3,776,770 tweets were scraped, with 1,093,198 tweets from Pune and 2,683,572 from Dubai. Around 14% of total tweets posted in Pune over approximately 3 years were found expressing negative sentiment while 43% of them were positive and a large fraction of 42% of them were neutral . A similar trend was found in tweets posted in Dubai, where the total number of tweets expressing negative sentiment is significantly less than that expressing positive sentiment. (Figure 1, Figure 2)

Reliable weather data was bought from VisualCrossing Weather API for both Dubai and Pune. Humidity, precipitation, wind speed, cloud cover, visibility, and minimum and maximum daily temperature and 'feels-like' values were used from this data. While the average precipitation in Pune was 18.87 mm , it was 0.23 mm in Dubai over the last 3 years.

For the twitter data, the data was scraped and stored in partitioned files at a daily level. Also, the data was aggregated on a daily as well as hourly level for a period of 3 years. Owing to the huge volume of unstructured textual data to be processed and number of files to be loaded, the analysis was conducted using a `DASK` cluster hosted on Google Colaboratory to enable performant Natural Language Processing of the Big Data.

4 Methodology

The approach followed in this project consists of four steps: data collection from twitter, preprocessing; tweet polarity identification; temporal and spatiotemporal analysis of the data. As the underlying data is textual, appropriate steps as illustrated in Figure 7 were followed to perform sentiment analysis.

The overall average sentiment expressed through tweets on any day of the year is positive for both cities as the frequency of positive tweets is consistently greater than the frequency of negative tweets. Thus, we analyze the relationship between weather features and a) positive sentiment expression b) negative sentiment expression separately.

4.0.1 Data Preprocessing

The tweets scraped from twitter were already filtered by location and language used. Tweet Preprocessing was performed using the `tweet-processor` package in python. Hashtags, usernames, and urls were removed. Punctuation was also removed for easier tokenization.

4.0.2 Tokenization

For qualitative analysis of the tweets, they first need to be broken down into words i.e. tokens. Tokenization transforms unstructured textual material into a set of words for further processing. Each tweet is read as a corpus for tokenization using the `textblob` library in python. As the textblob tokenization library uses *space* as a delimiter, it can produce erroneous outputs when there are typographical errors in the text.

4.0.3 Polarity and Subjectivity Calculation

The PatternAnalyzer implementation of the TextBlob package was used for calculation of polarity and subjectivity of every tweet. Polarity lies between [-1,1]. -1 defines a negative sentiment and 1 defines a positive sentiment. Thus, a tweet is defined as expressing negative sentiment if it has polarity < 0 and it is defined as expressing positive sentiment if polarity > 0 for it. TextBlob has semantic labels that help with fine-grained analysis. For example — emoticons, exclamation marks, emojis, etc Loria (2018). How these special characters are handled in intermediate and final results of the above steps are shown in Figure 3, Figure 4, Figure 5.

4.0.4 Predicting Polarity and Frequency of Tweets using Weather features

In order to explore and quantify any underlying relationship between the weather and expressed sentiment, a workflow was followed to predict the average sentiment polarity expressed on any given day using weather features for both Dubai and Pune separately.

Steps of the workflow are described below:

- Smoothing of time series using a rolling mean with a window of size 7.
- Identification of strongest period present in time series using periodogram.
- Multiplicative seasonal decomposition of time series using period calculated in the previous step (Figures 24, 23)
- Time series data preparation for input to Long-Short Term Memory (LSTM) model.
- Train test split of multivariate time series data keeping test-size as 0.2. Out of a 1052 day dataset, 847 are used for training and the rest for test.
- Evaluating model performance.

5 Results

While analyzing relationships between twitter sentiment and the chosen weather features, a correlation of strength 0.04 is observed between tweet frequency and visibility, while the strongest correlation of strength 0.08 is observed between average polarity of tweets and humidity. Figure 9 indicates that the average expressed sentiment in Dubai peaks at 3 am and steadily declines throughout the day until midnight, which is when it starts increasing again. Figure 8 indicates that the frequency of tweets peaks around 7 am and 5 pm while it experiences a mid-day slump. This trend could be attributed to high temperatures during the day and office hours. The trends for frequency of tweets per hour of the day are interestingly similar for both of the cities.

Figure 11 indicates that the average expressed sentiment in Pune peaks at midnight and steadily declines till 8 pm, which is when it starts increasing again. During the day, local maxima of expressed sentiment occur at 11 am and 6 pm. This trend could be attributed to lunch hours and end of work-day times.

Figure 10 indicates that the frequency of tweets peaks around 7 am and 5 pm while it experiences a mid-day slump. This trend could be attributed to high temperatures during the day and office hours.

Figures 12, 13, 14, 15 show that the correlation between weather data and expressed sentiment was higher at tourist spots when compared to the entire city for both Pune and Dubai. This means that weather has an amplified effect on expressed sentiment at tourist locations in both cities.

Results from Figure 4.0.4 are inconclusive as LSTM models require significantly more data to learn patterns Gasmi et al. (2018). The model architecture is described in Figure 20 was used to predict outputs as displayed in Figures 17, 18.

6 Conclusion

Through this analysis and modeling, insightful relationships between the weather data and sentiment data was uncovered. The results of this project can be scaled to conclude that changes in weather have a direct impact on the expressed sentiment expressed at various tourist spots across tourism developed and developing cities like Dubai and Pune respectively. Policy makers can use results built on this study to develop efficient strategies that channelize taxpayer money in their cities.

7 Future Scope

Better quality geo-tagged social data is required for high definition spatial analysis - they account for just 1-2% of total twitter data. Thus, a higher resolution spatial analysis was out of the scope of this project Middleton et al. (2018), Priedhorsky et al. (2014), Tasse et al. (2017). Tweets using the English language to communicate in other languages are misclassified because words present in them are not a part of the NLTK library that TextBlob is built on Figure 6. In this tweet, a positive sentiment is expressed in Arabic. *'Alhamdulilah'* means *'praise be to God'*. However, the tweet is calculated to have negative polarity.

8 Figures

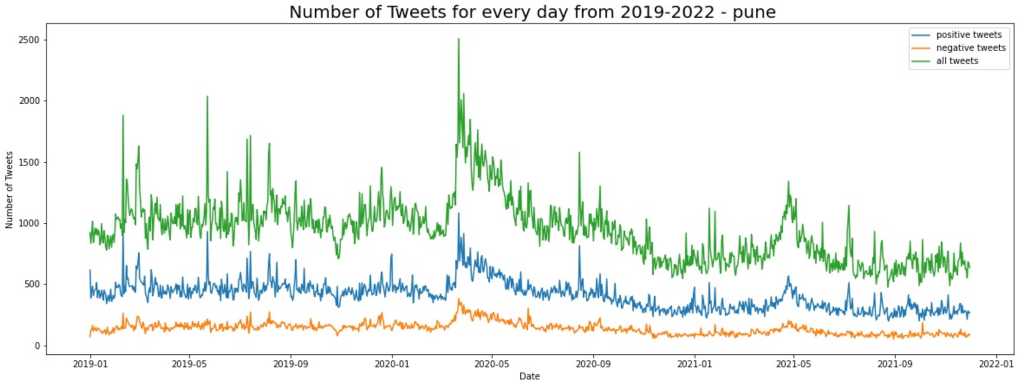


Figure 1: Tweet Frequency in Pune

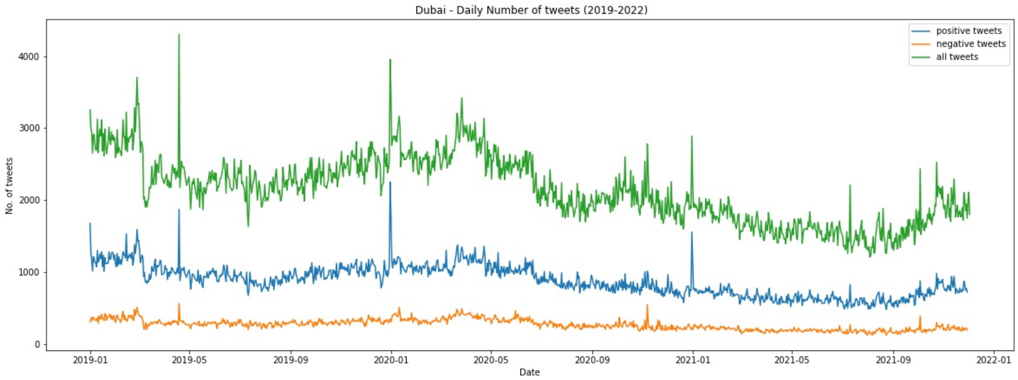
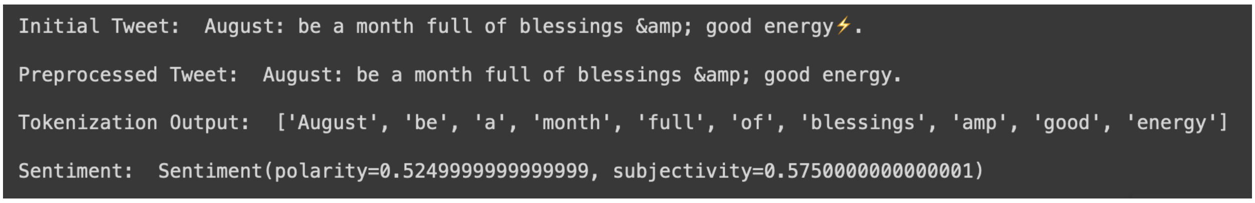


Figure 2: Tweet Frequency in Dubai



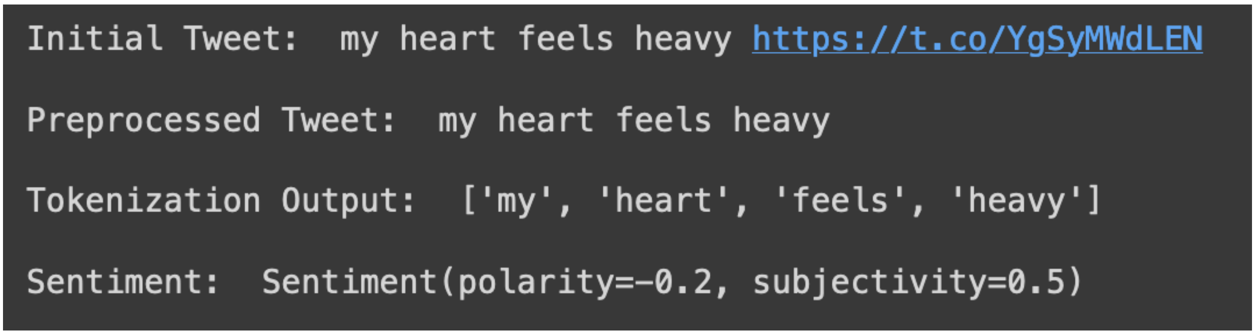
```
Initial Tweet:  August: be a month full of blessings &amp; good energy⚡.

Preprocessed Tweet:  August: be a month full of blessings &amp; good energy.

Tokenization Output:  ['August', 'be', 'a', 'month', 'full', 'of', 'blessings', 'amp', 'good', 'energy']

Sentiment:  Sentiment(polarity=0.5249999999999999, subjectivity=0.5750000000000001)
```

Figure 3: Pre-processing, tokenization, and sentiment analysis of a positive tweet.



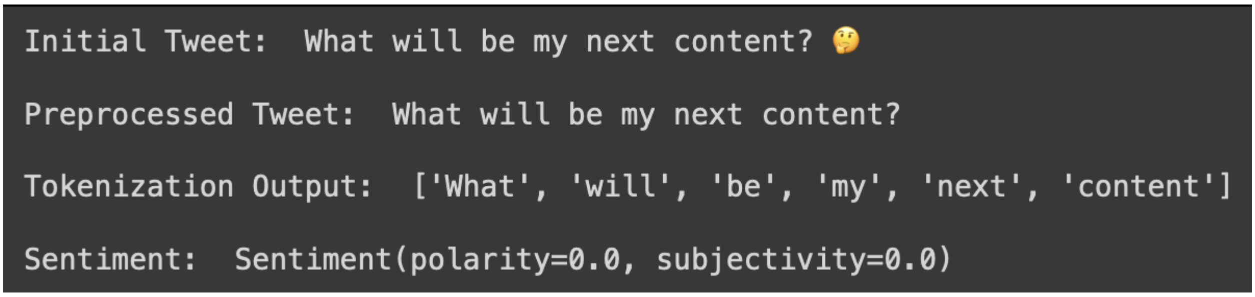
```
Initial Tweet:  my heart feels heavy https://t.co/YgSyMWdLEN

Preprocessed Tweet:  my heart feels heavy

Tokenization Output:  ['my', 'heart', 'feels', 'heavy']

Sentiment:  Sentiment(polarity=-0.2, subjectivity=0.5)
```

Figure 4: Pre-processing, tokenization, and sentiment analysis of a negative tweet.



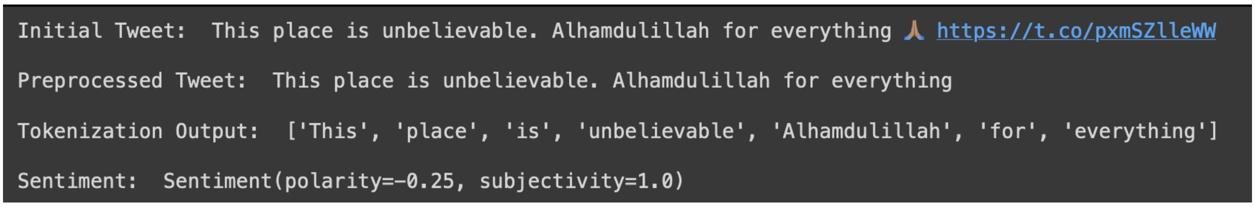
```
Initial Tweet:  What will be my next content? 🤔

Preprocessed Tweet:  What will be my next content?

Tokenization Output:  ['What', 'will', 'be', 'my', 'next', 'content']

Sentiment:  Sentiment(polarity=0.0, subjectivity=0.0)
```

Figure 5: Pre-processing, tokenization, and sentiment analysis of a neutral tweet.



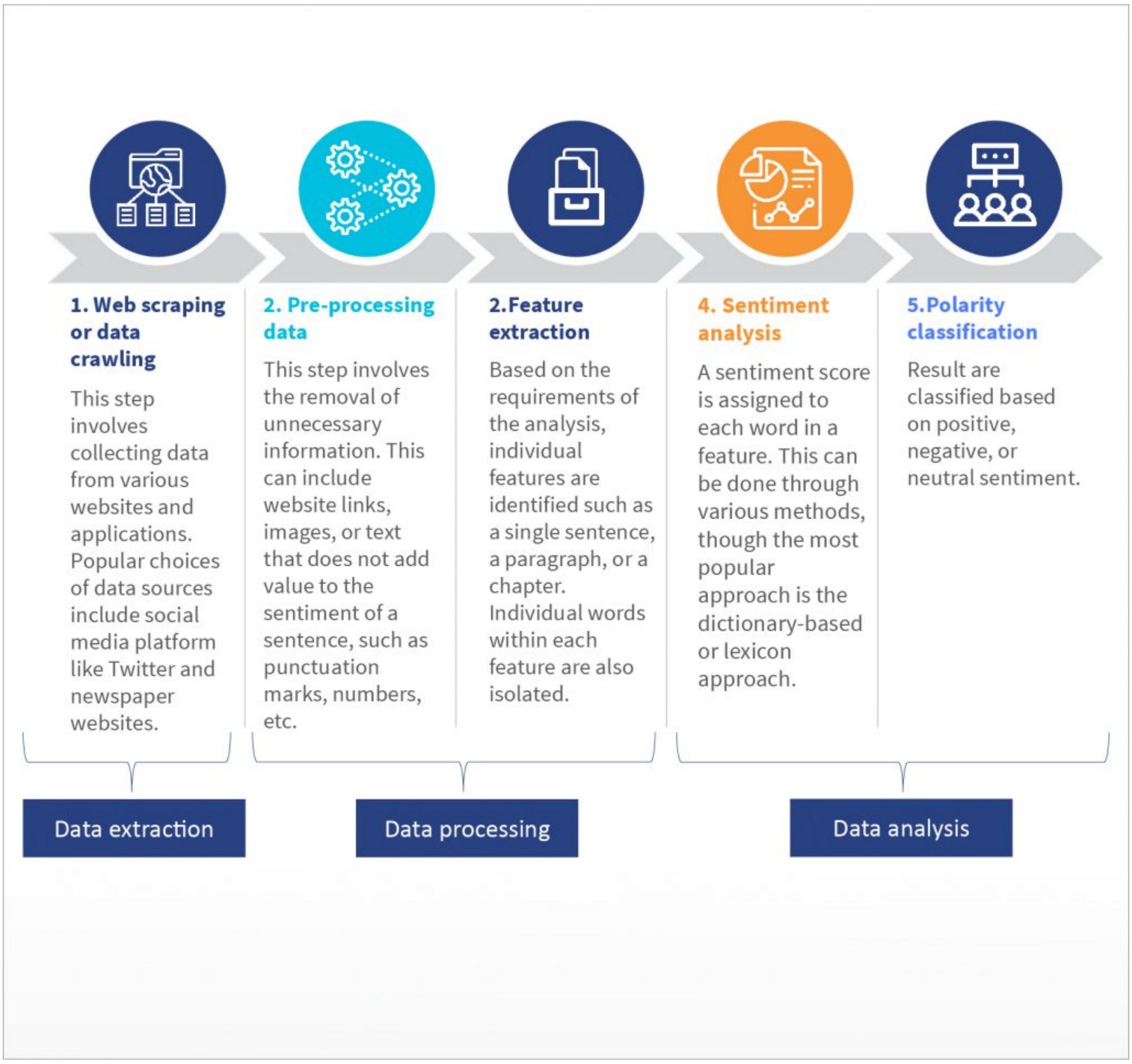
```
Initial Tweet:  This place is unbelievable. Alhamdulillah for everything 🙏 https://t.co/pxmSZlleWW

Preprocessed Tweet:  This place is unbelievable. Alhamdulillah for everything

Tokenization Output:  ['This', 'place', 'is', 'unbelievable', 'Alhamdulillah', 'for', 'everything']

Sentiment:  Sentiment(polarity=-0.25, subjectivity=1.0)
```

Figure 6: Skewed calculation of sentiment due to transliteration



1. Web scraping or data crawling

This step involves collecting data from various websites and applications. Popular choices of data sources include social media platform like Twitter and newspaper websites.

2. Pre-processing data

This step involves the removal of unnecessary information. This can include website links, images, or text that does not add value to the sentiment of a sentence, such as punctuation marks, numbers, etc.

2.Feature extraction

Based on the requirements of the analysis, individual features are identified such as a single sentence, a paragraph, or a chapter. Individual words within each feature are also isolated.

4. Sentiment analysis

A sentiment score is assigned to each word in a feature. This can be done through various methods, though the most popular approach is the dictionary-based or lexicon approach.

5.Polarity classification

Result are classified based on positive, negative, or neutral sentiment.

Figure 7: Image Sourced From: *Sentiment Analysis Workflow* (n.d.)

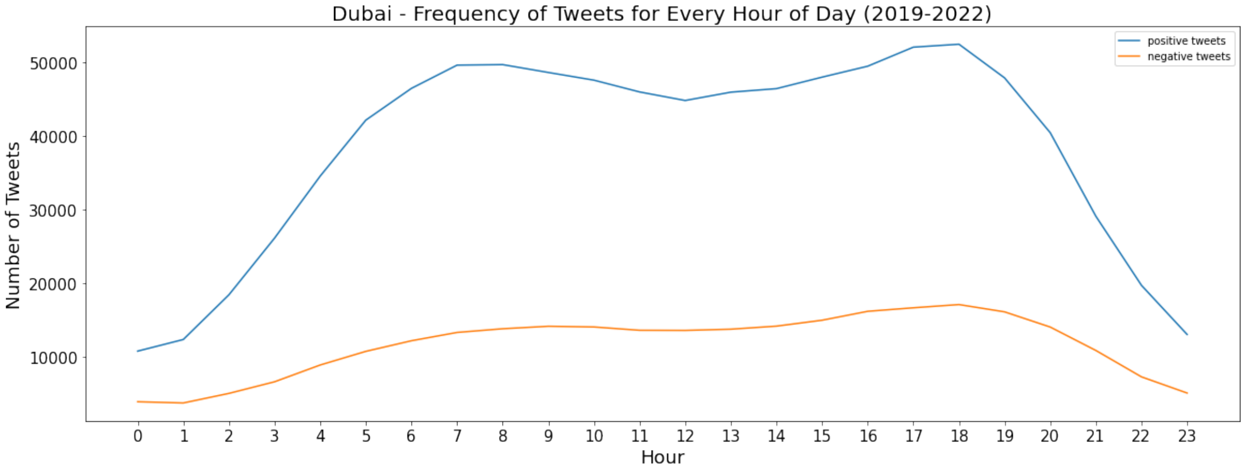


Figure 8: Dubai - Frequency of Tweets per Hour of Day

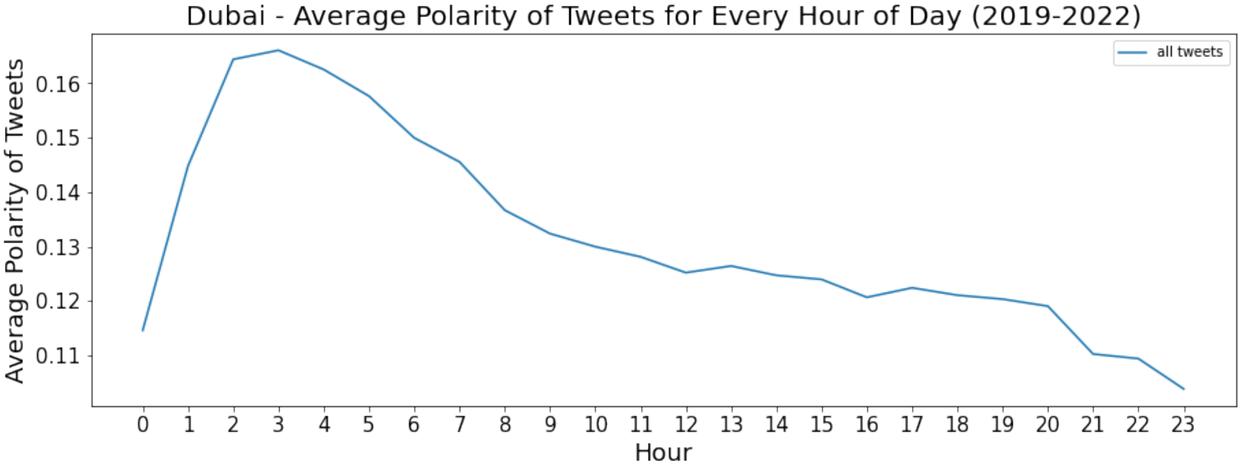


Figure 9: Dubai - Average Polarity of Tweets per Hour of Day

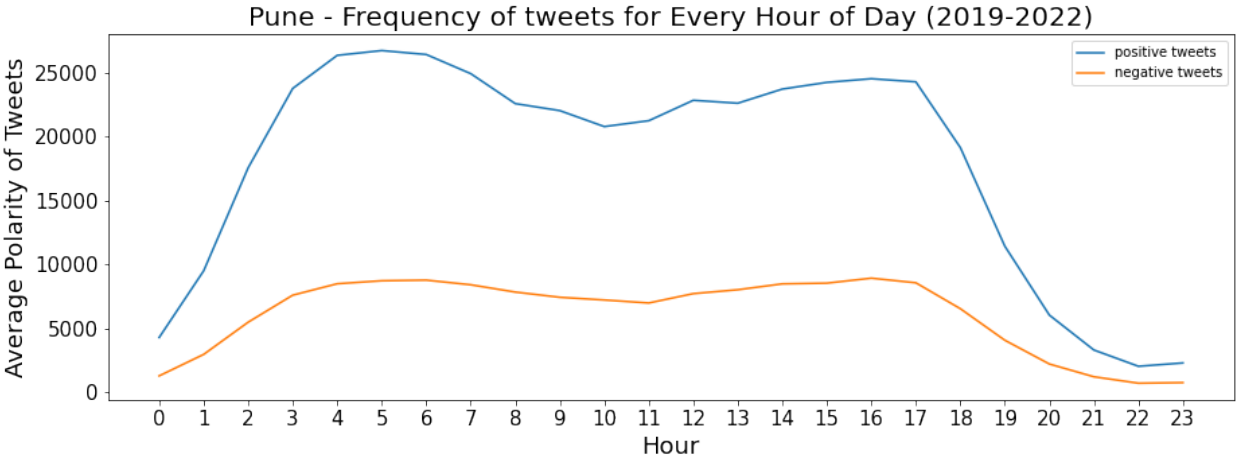


Figure 10: Pune - Frequency of Tweets per Hour of Day

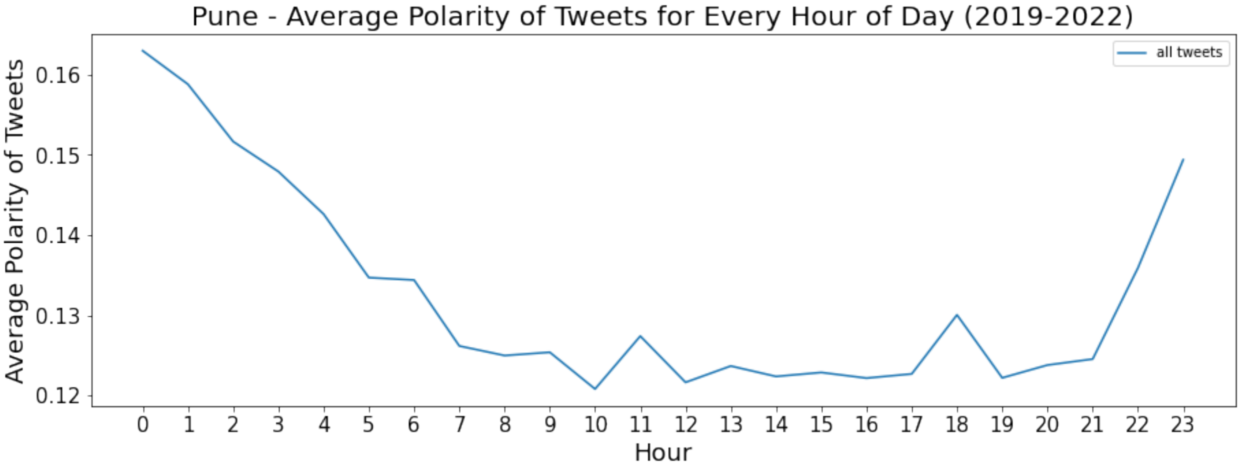


Figure 11: Pune - Average Polarity of Tweets per Hour of Day

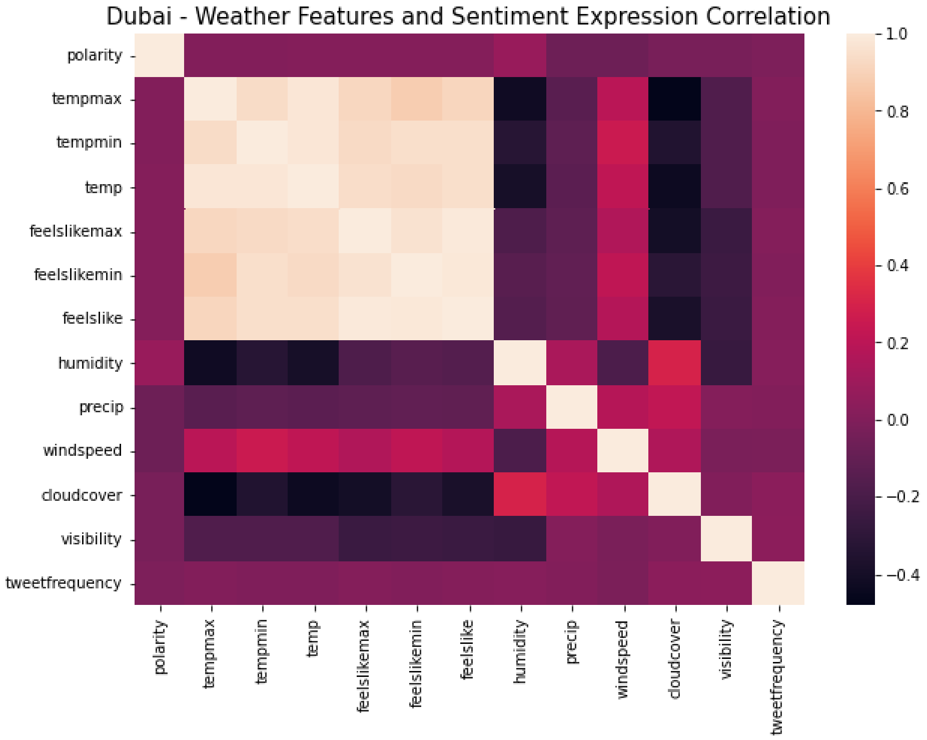


Figure 12: Dubai - Correlation of Weather Features and Sentiment Expression

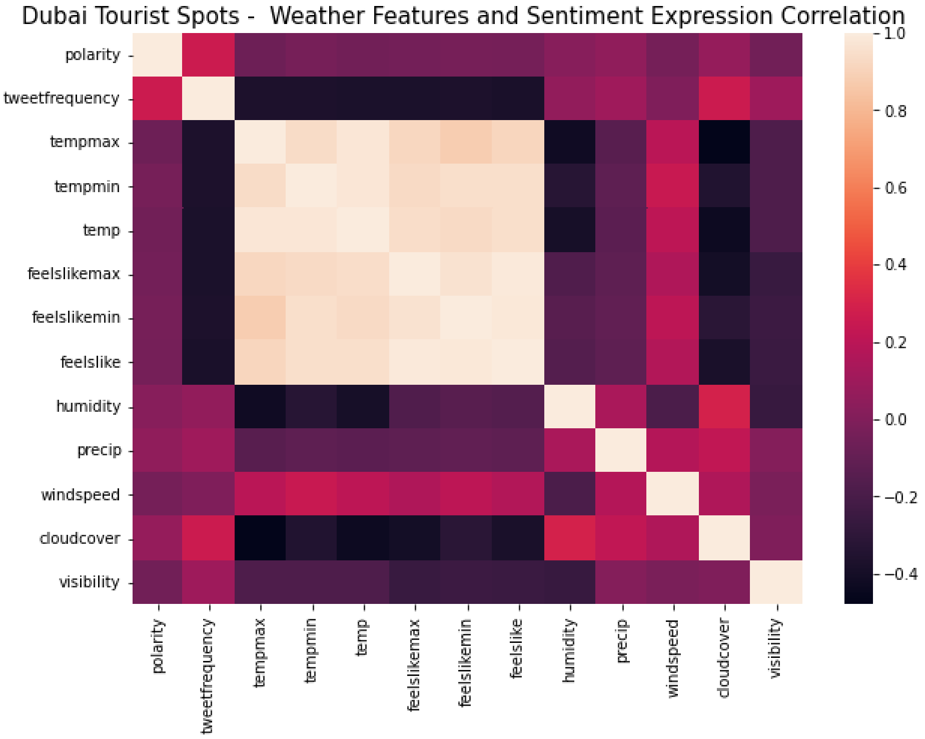


Figure 13: Dubai Tourist Spots - Correlation of Weather Features and Sentiment Expression

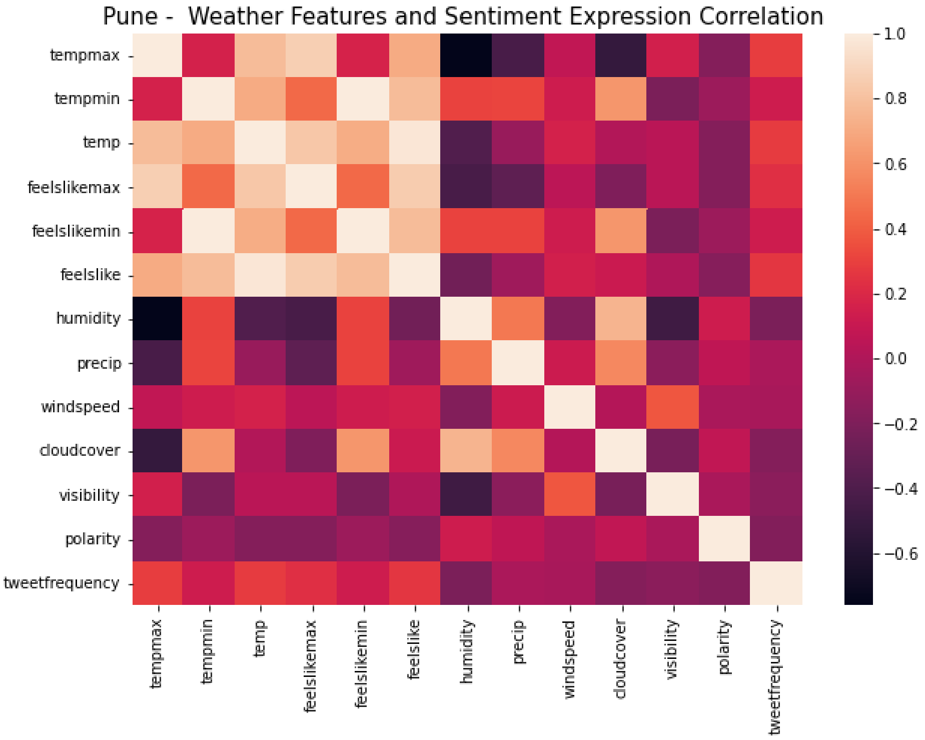


Figure 14: Pune -Correlation of Weather Features and Sentiment Expression

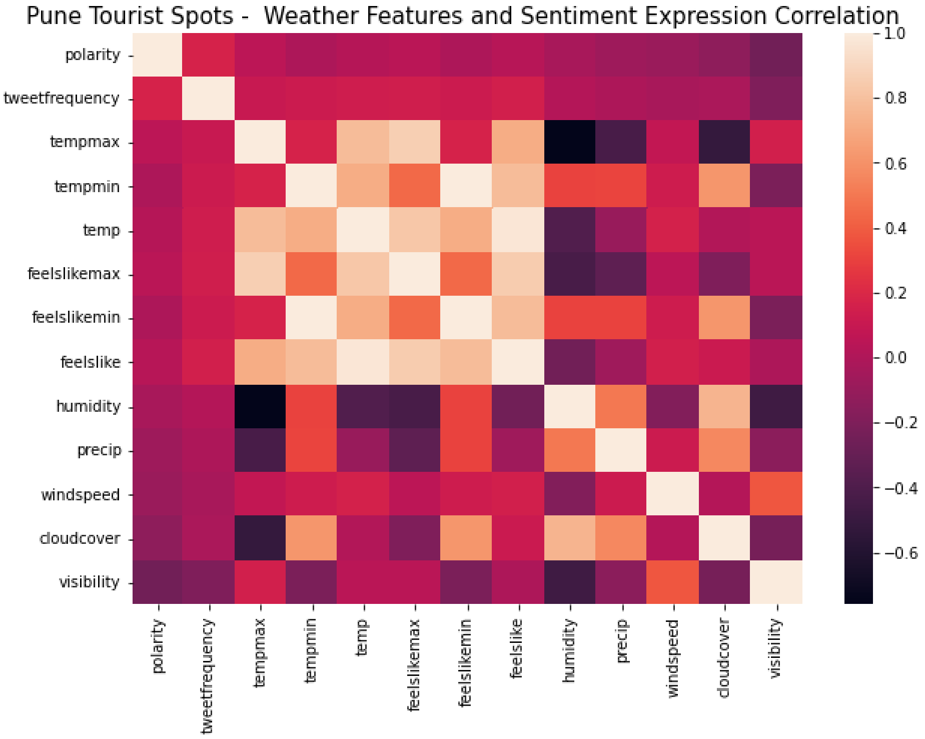
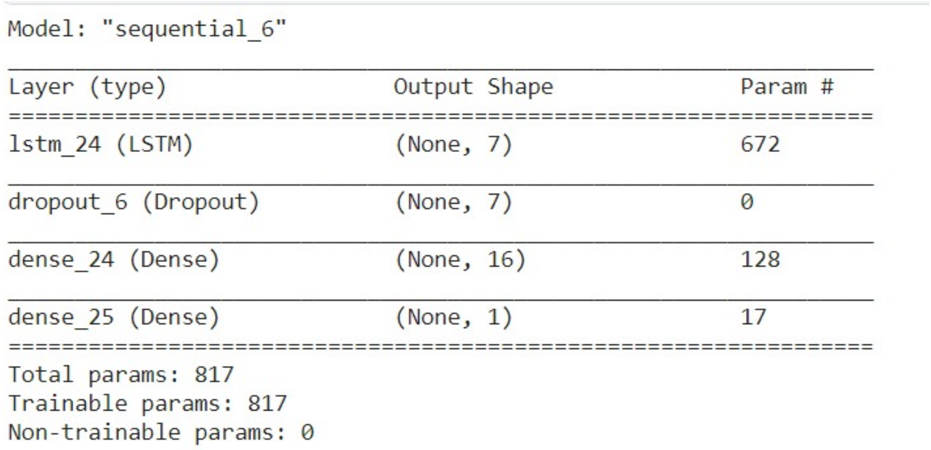


Figure 15: Pune Tourist Spots - Correlation of Weather Features and Sentiment Expression



```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_24 (LSTM)               (None, 7)                 672
_________________________________________________________________
dropout_6 (Dropout)          (None, 7)                 0
_________________________________________________________________
dense_24 (Dense)             (None, 16)                128
_________________________________________________________________
dense_25 (Dense)             (None, 1)                 17
=================================================================
Total params: 817
Trainable params: 817
Non-trainable params: 0
```

Figure 16: LSTM Model Architecture

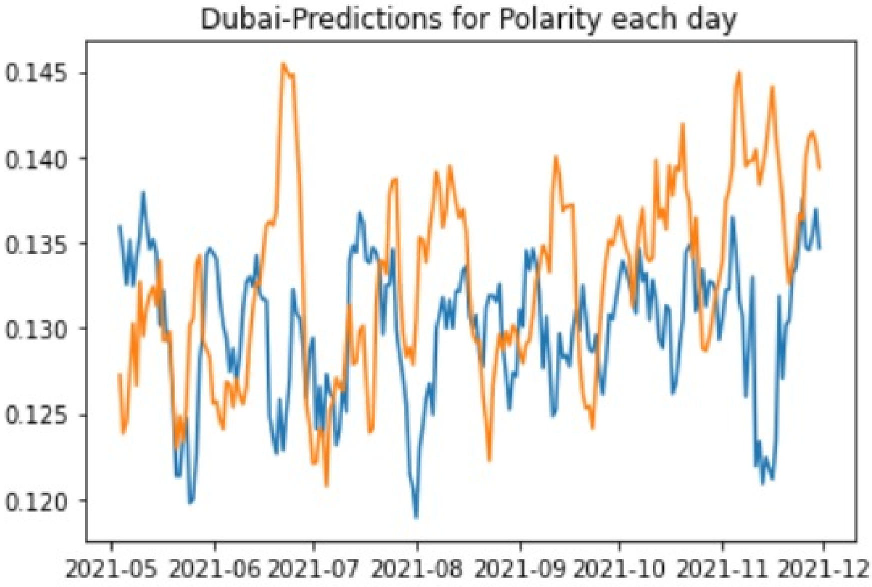


Figure 17: Dubai Predictions for Expressed Sentiment Polarity

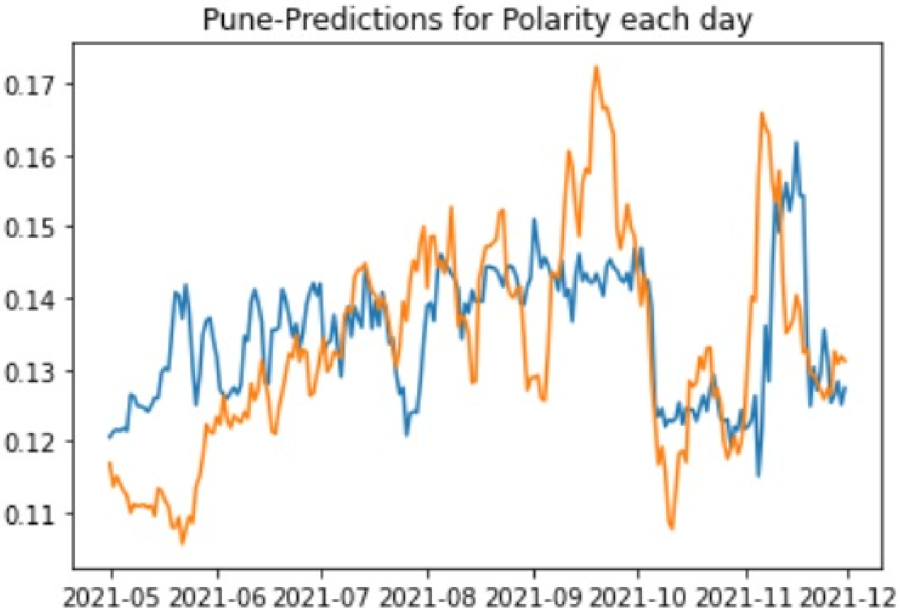


Figure 18: Pune Predictions for Expressed Sentiment Polarity

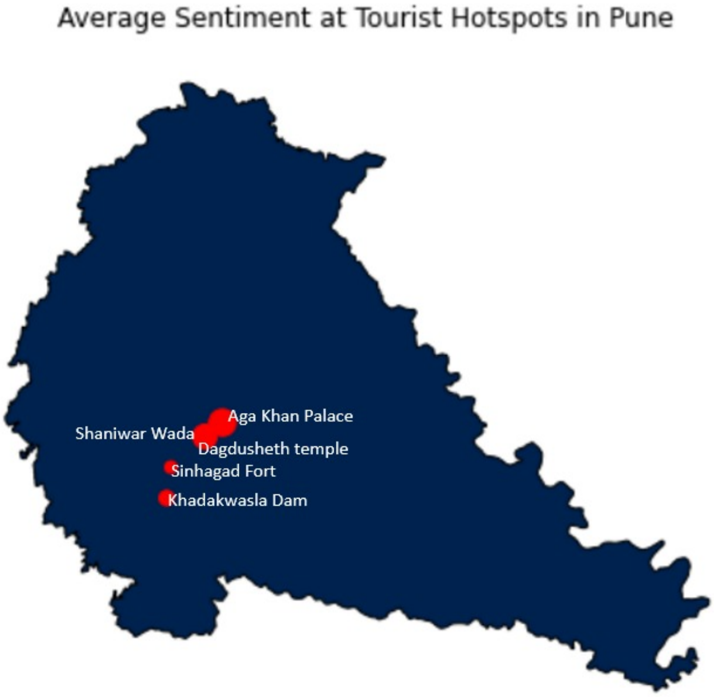


Figure 19:

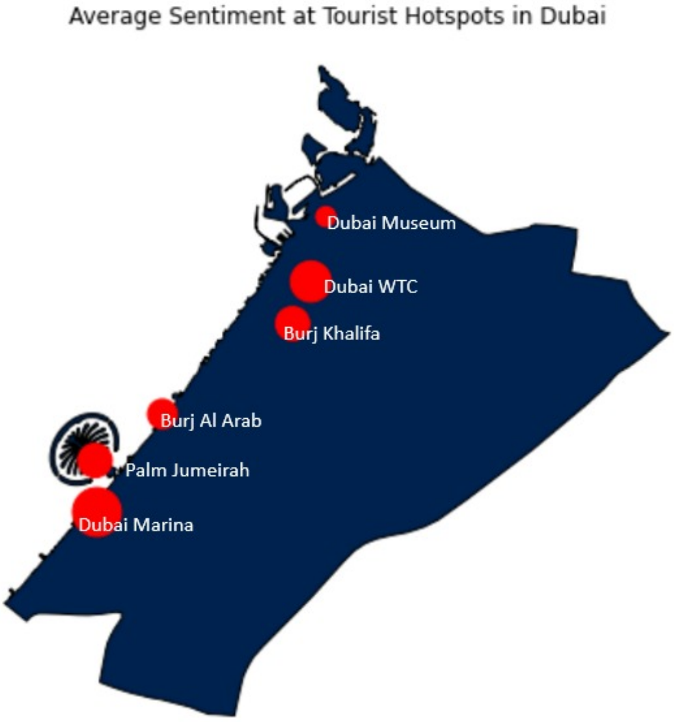


Figure 20:

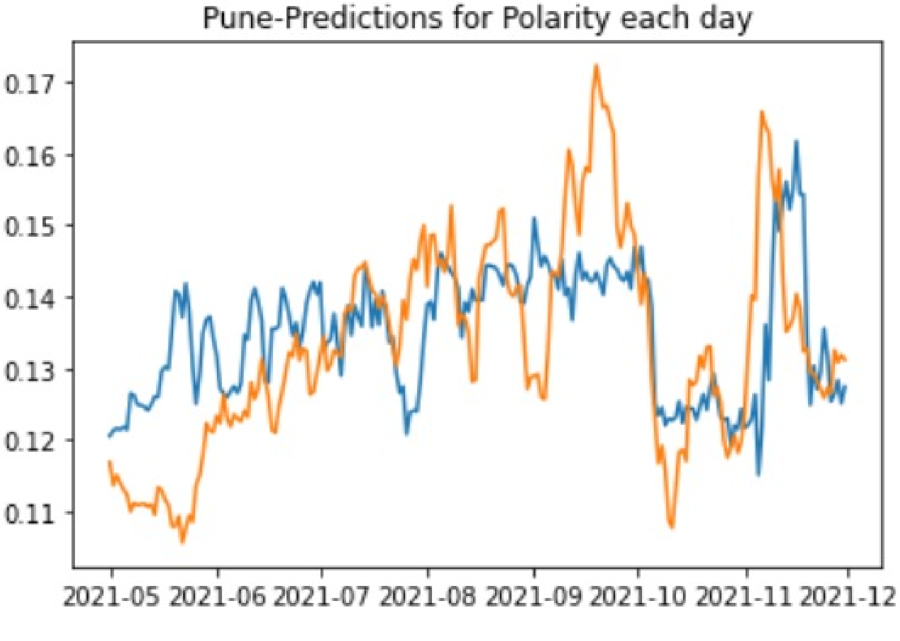


Figure 21:

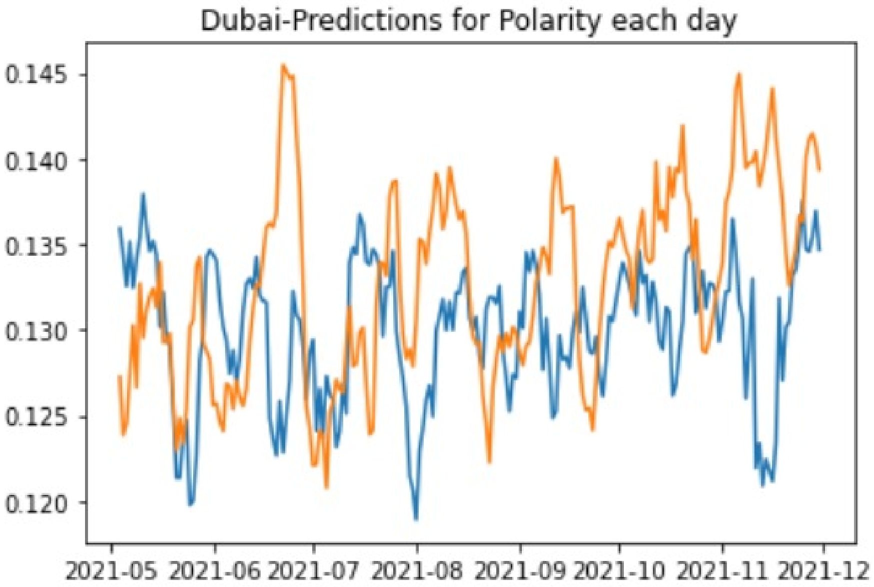


Figure 22:

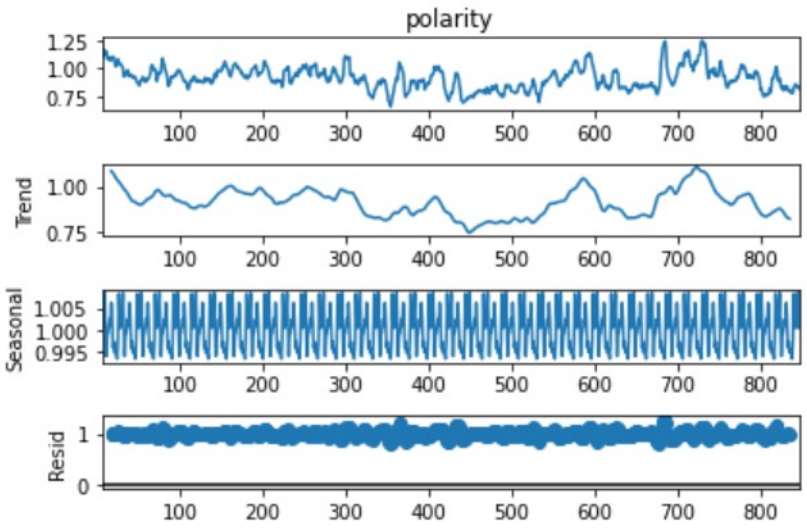


Figure 23:

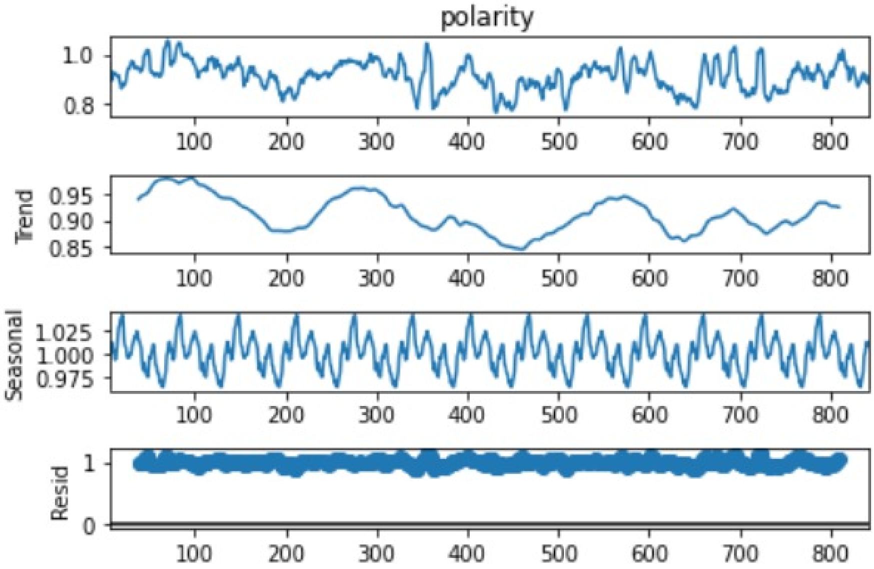


Figure 24:

References

Ahmed, W. (2017), ‘Using twitter as a data source: an overview of social media research tools (updated for 2017)’, *Impact of Social Sciences Blog* .

Deshpande, B. & Deshpande, R. (2016), ‘A study on development of tourism i n maharashtra’, *International Journal of Scientific and Research Publications* **6**(7), 175–181.

Gasmi, H., Bouras, A. & Laval, J. (2018), ‘Lstm recurrent neural networks for cybersecurity named entity recognition’, *ICSEA* **11**, 2018.

Loria, S. (2018), ‘textblob documentation’, *Release 0.15* **2**, 269.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S. & Kompatsiaris, Y. (2018), ‘Location extraction from social media: Geoparsing, location disambiguation, and geotagging’, *ACM Transactions on Information Systems (TOIS)* **36**(4), 1–27.

Padilla, J. J., Kavak, H., Lynch, C. J., Gore, R. J. & Diallo, S. Y. (2018), ‘Temporal and spatiotemporal investigation of tourist attraction visit sentiment on twitter’, *PloS one* **13**(6), e0198857.

Priedhorsky, R., Culotta, A. & Del Valle, S. Y. (2014), Inferring the origin locations of tweets with quantitative confidence, *in* ‘Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing’, pp. 1523–1536.

Scott, D. & Lemieux, C. (2010), ‘Weather and climate information for tourism’, *Procedia Environmental Sciences* **1**, 146–183.

Sentiment Analysis Workflow (n.d.), https://www.microsave.net/2020/12/21/respect-the-sentiment-using-sentiment-analysis-to-inform-policy-and-regulation/. Accessed: 2021-12-12.

Stephens, M., bin Khalifa, S. S. b. S., Nahyan, A. & Schroeder, C. M. (2019), Perspective—future disruptive governments: Catching up with technological advancements and new horizons, *in* ‘Future Governments’, Emerald Publishing Limited.

Tasse, D., Liu, Z., Sciuto, A. & Hong, J. (2017), State of the geotags: Motivations and recent changes, *in* ‘Proceedings of the International AAAI Conference on Web and Social Media’, Vol. 11.

tourism (n.d.), https://www.dsc.gov.ae/en-us/Themes/Pages/Tourism.aspx?Theme=30. Accessed: 2021-12-12.

Valdez, D., Ten Thij, M., Bathina, K., Rutter, L. A. & Bollen, J. (2020), 'Social media insights into us mental health during the covid-19 pandemic: longitudinal analysis of twitter data', *Journal of medical Internet research* **22**(12), e21418.

Zaidan, E. & Kovacs, J. F. (2017), 'Resident attitudes towards tourists and tourism growth: A case study from the middle east, dubai in united arab emirates', *European Journal of Sustainable Development* **6**(1), 291–291.