Project 3

Singapore Travel Fair Campaign

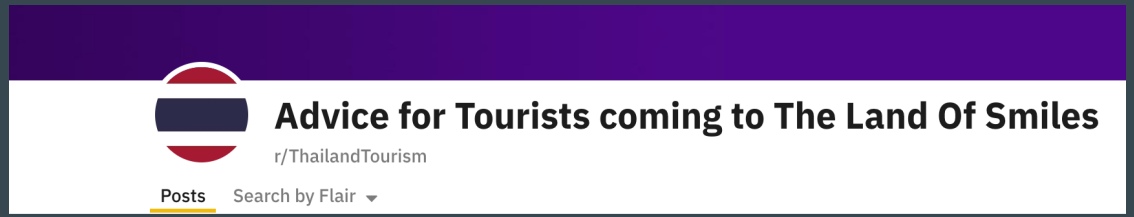
Kenneth Lim

Agenda

1. Problem Statement
2. Data Pre-processing + EDA
3. Data Modelling
4. Results
5. Conclusion and Recommendations

Problem Statement

- Singapore lifts remaining Covid-19 pre-departure testing measures in Feb 2023
- Surge in Travel bookings and holidays among Singaporeans
- Assist Marketing and Operations team in an online local travel technology agency to digitalize their travel promotion packages by training of ML model
- Collection of Subreddit travel posts for feedback analysis



Top searched destinations this winter season[3]

1. Singapore	6. Seoul
2. Bangkok	7. Taipei
3. Tokyo	8. Osaka
4. Bali	9. Johor Bahru
5. Kuala Lumpur	10. London

[3] Based on accommodation search data on Expedia.com.sg during the period of July to November 2022,for travel in December 2022 to February

Why Select Subreddit of these countries?

- “Recent study from the Expedia Team showed that 66% of Singaporeans prioritize experiences and travel over things in life Post Covid-19 Pandemic”

Source: https://www.expedia.com.sg/stories/press/expedia-66-of-singapore-travellers-prioritise-experiences-and-travel-over-things-in-life-post-pandemic/

Data Pre-processing

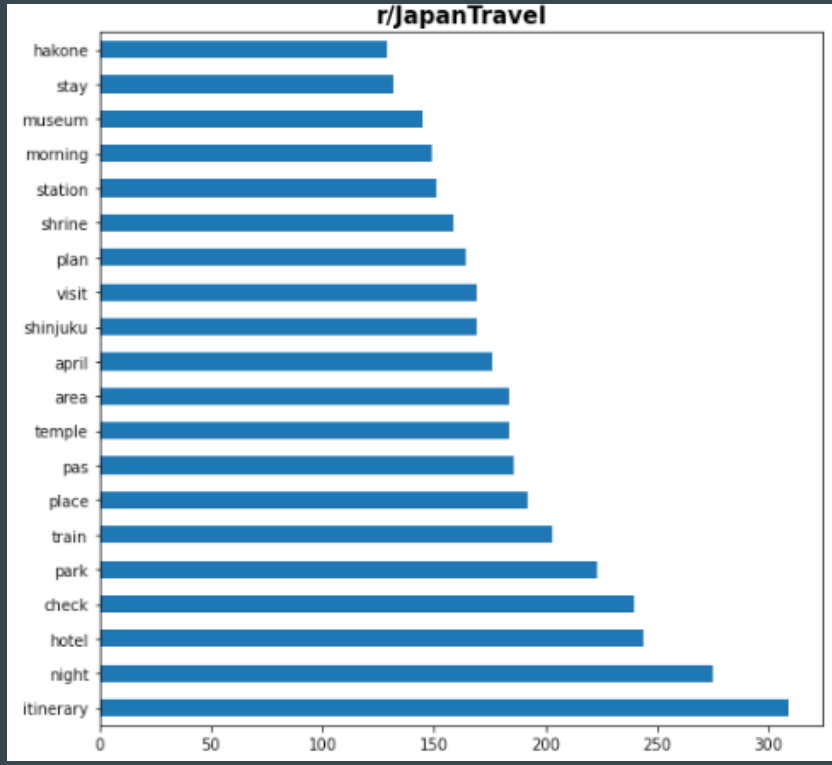
Remove web URLs, non-words and special characters

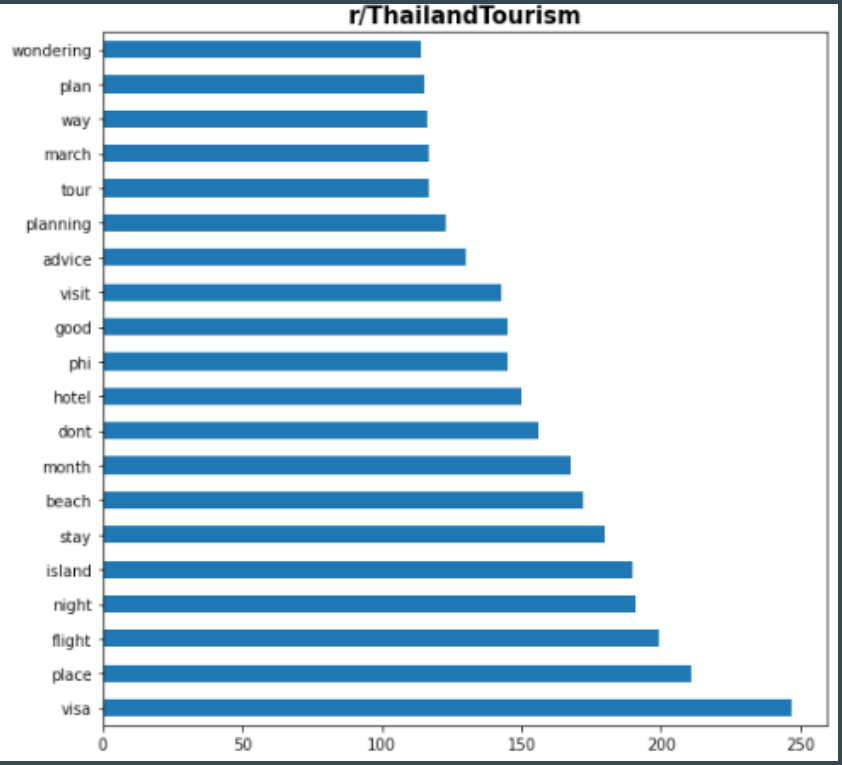
Tokenize text sentences into lower case words

Filter common stopwords and keywords relating to Japan and Thailand

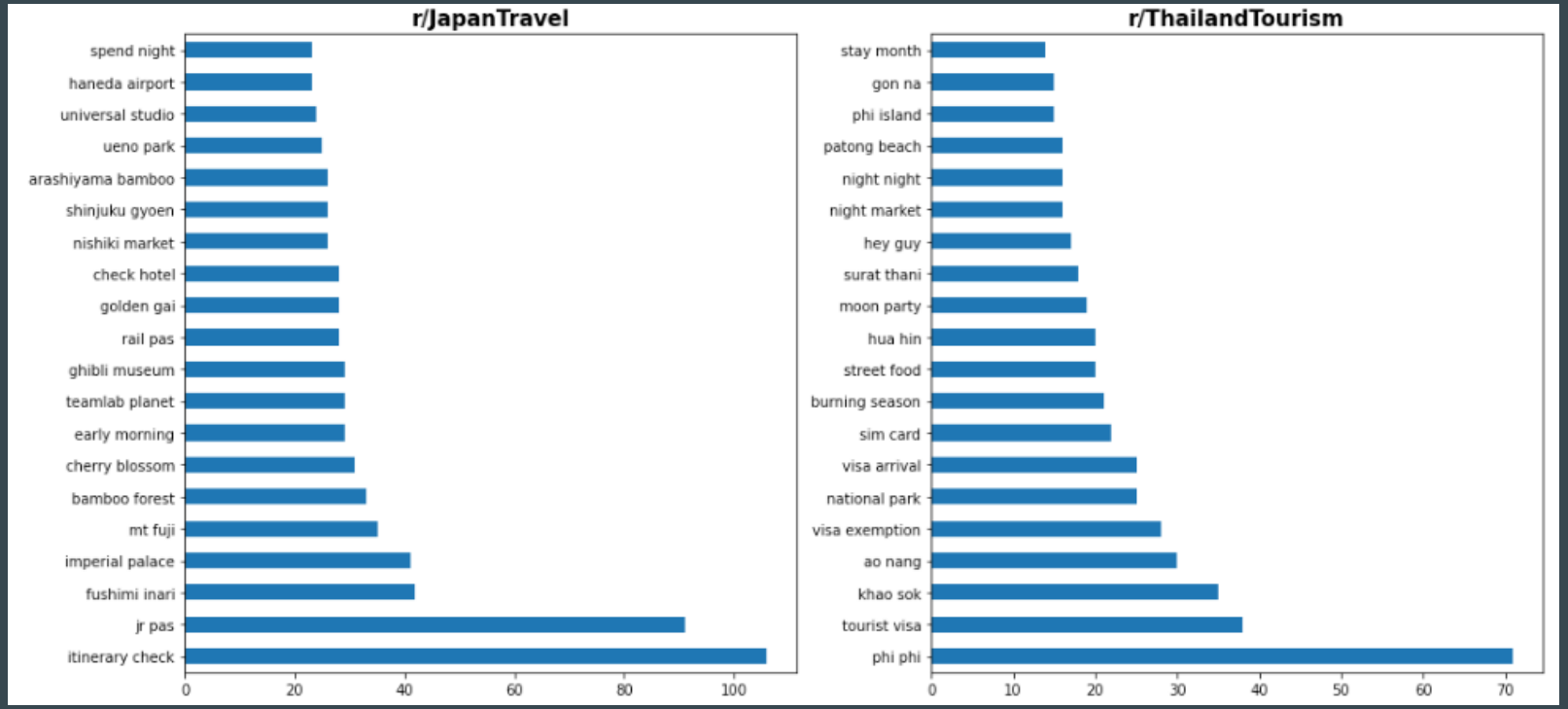
Pass into CountVectorizer/TFID Vectorizer for N-gram word feature analysis

Uni-gram

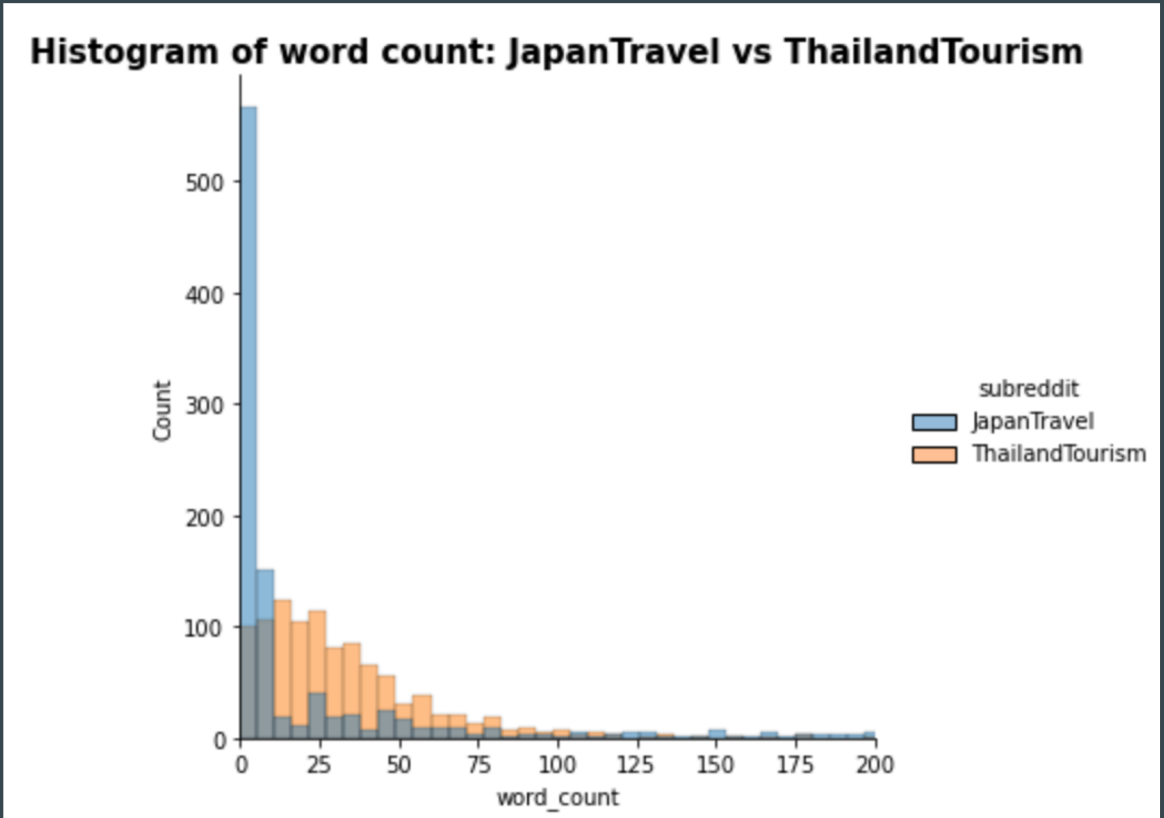

r/JapanTravel
hakone
stay
museum
morning
station
shrine
plan
visit
shinjuku
april
area
temple
pas
place
train
park
check
hotel
night
itinerary
0
50
100
150
200
250
300


r/ThailandTourism
wondering
plan
way
march
tour
planning
advice
visit
good
phi
hotel
dont
month
beach
stay
island
night
flight
place
visa
0
50
100
150
200
250

Bi-gram


r/JapanTravel
spend night
haneda airport
universal studio
ueno park
arashiyama bamboo
shinjuku gyoen
nishiki market
check hotel
golden gai
rail pas
ghibli museum
teamlab planet
early morning
cherry blossom
bamboo forest
mt fuji
imperial palace
fushimi inari
jr pas
itinerary check
0
20
40
60
80
100
r/ThailandTourism
stay month
gon na
phi island
patong beach
night night
night market
hey guy
surat thani
moon party
hua hin
street food
burning season
sim card
visa arrival
national park
visa exemption
ao nang
khao sok
tourist visa
phi phi
0
10
20
30
40
50
60
70

Word Count


Histogram of word count: JapanTravel vs ThailandTourism
Count
word_count
subreddit
JapanTravel
ThailandTourism
0
100
200
300
400
500
0
25
50
75
100
125
150
175
200

Data Modelling

Goal: To train a novel prototype NLP Classifier Model that can accurately classify under the correct subreddit categories in long-term for further fine-tuning and deployment

Bag-of-Words

- Overall individual word counts
- Word frequency count

Vectorization

- **Condense individual words into sparse matrix of array**

Evaluation Metric

- **Reduce Misclassification rate**
- **Optimize accuracy and ROC-AUC metrics**

Classification Algorithms

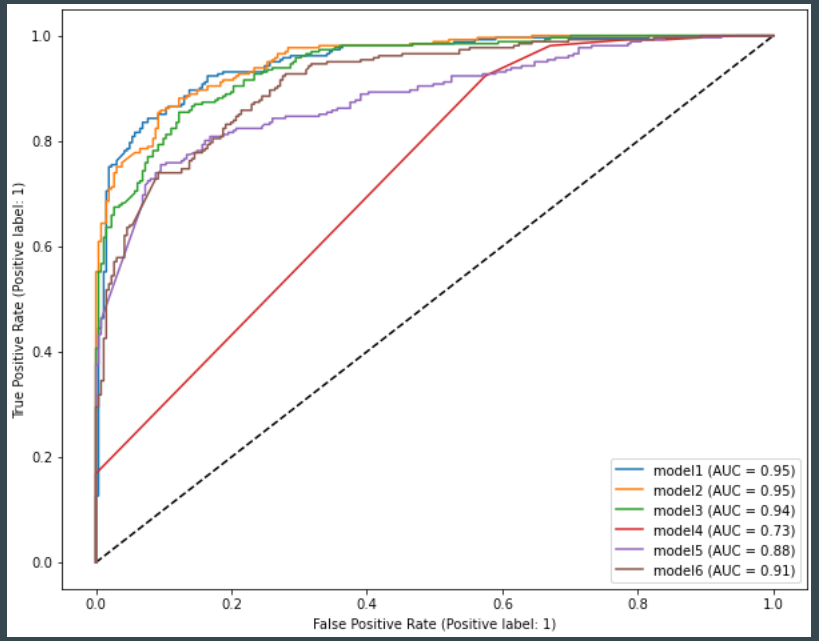
- **Naïve Bayes**
- **Logistic Regression**
- **Decision Tree**
- **RandomForest**
- **GradientBooting**

Results

Model No.	Model	Accuracy	Precision	Recall	F1	AUC
1	Multinomial Naive Bayes(CountVectorizer)	0.864	0.961	0.759	0.848	0.95
2	Multinomial Naive Bayes(TFIDF)	0.860	0.952	0.759	0.844	0.95
3	**Logistic Regression(TFIDF)**	**0.856**	0.881	0.824	0.851	0.94
4	Decision Tree(TFIDF)	0.674	0.616	0.923	0.739	0.73
5	Random Forest(TFIDF)	0.810	0.812	0.808	0.810	0.88
6	Gradient Boosting (TFIDF)	0.818	0.803	0.843	0.822	0.91

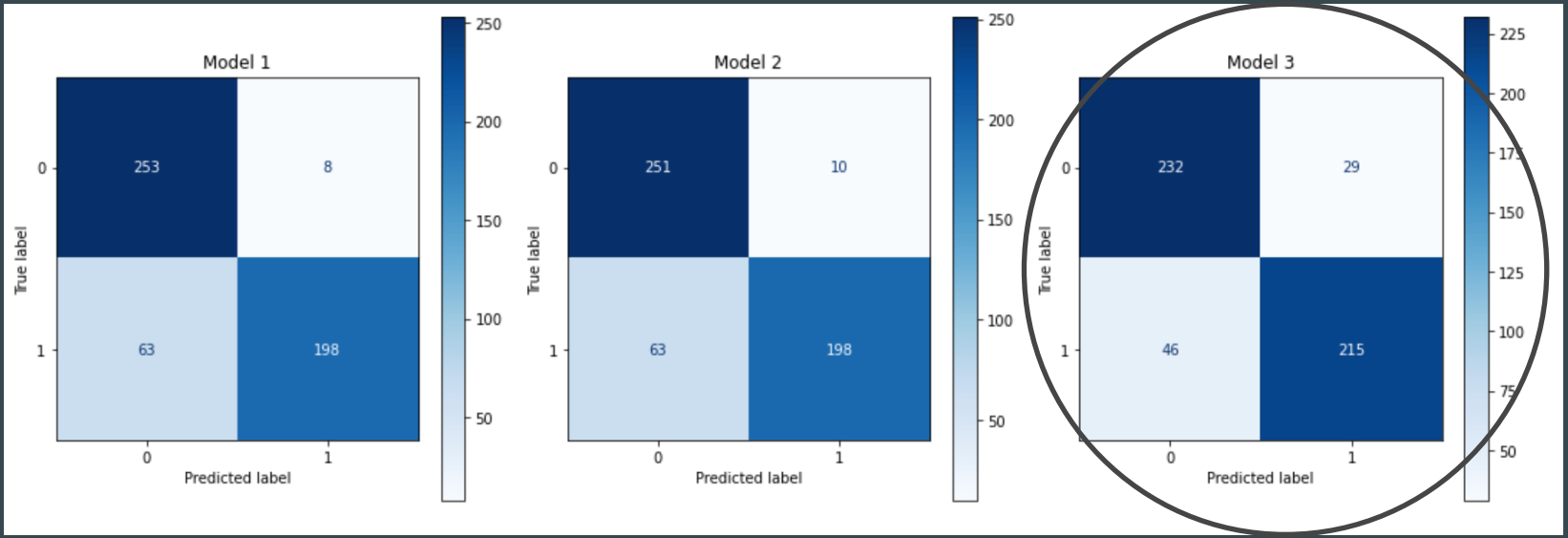
Model No.	Model	Train Accuracy	Test Accuracy
1	Multinomial Naive Bayes(CountVectorizer)	0.882	0.864
2	Multinomial Naive Bayes(TFIDF)	0.924	0.860
3	Logistic Regression(TFIDF)	0.955	0.856
4	Decision Tree(TFIDF)	0.675	0.674
5	Random Forest(TFIDF)	0.824	0.810
6	Gradient Boosting (TFIDF)	0.986	0.818

ROC-AUC Curve



Model No.	Model
1	Multinomial Naive Bayes(CountVectorizer)
2	Multinomial Naive Bayes(TFIDF)
3	Logistic Regression(TFIDF)
4	Decision Tree(TFIDF)
5	Random Forest(TFIDF)
6	Gradient Boosting (TFIDF)

Confusion Matrix

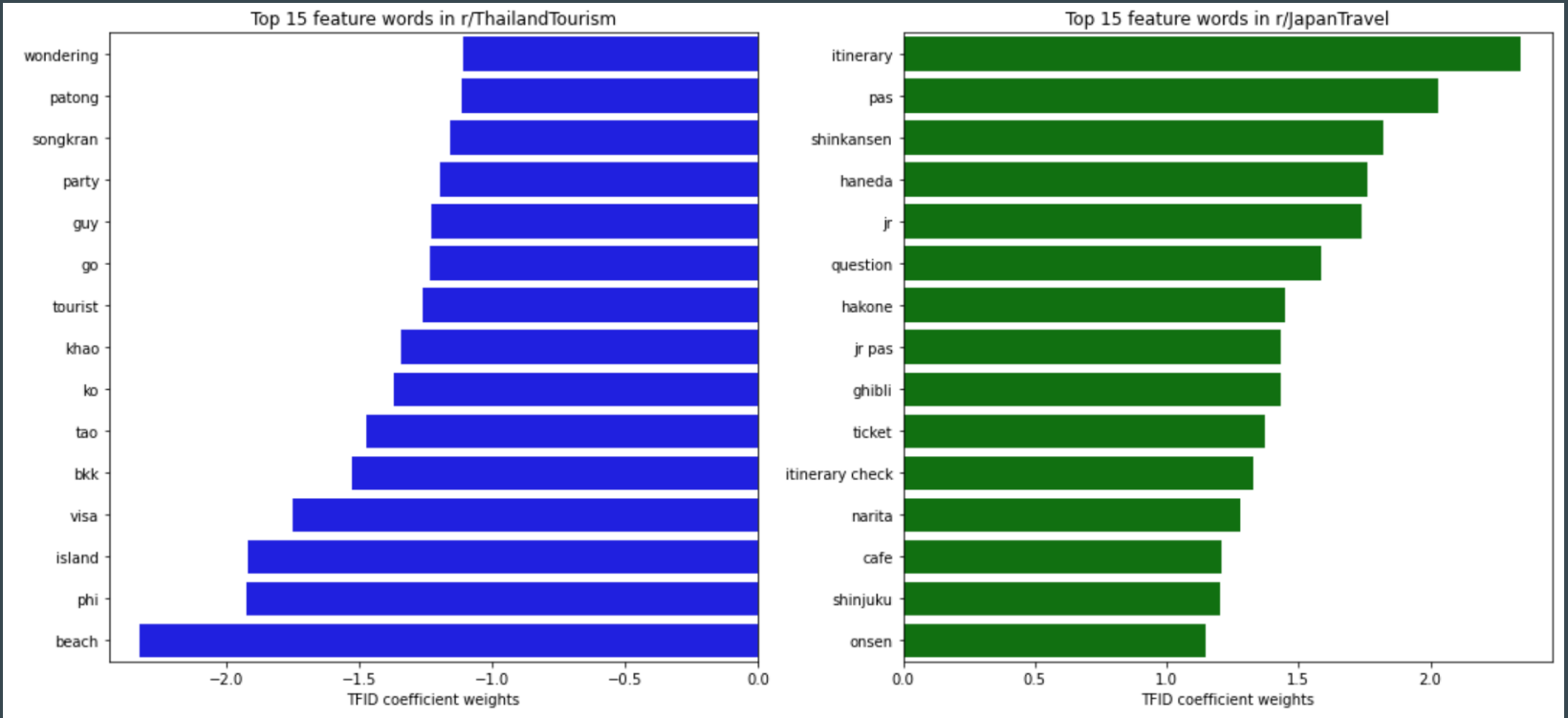


Model 1: Naïve Bayes (CountVectorizer)

Model 2: Naïve Bayes (TFIDVectorizer)

Model 3: Logistic Regression (TFIDVectorizer)

Word Feature Importance



Conclusion

Accuracy: 85.6%

- Selected Model for Production: **TFIDF, Logistic Regression**
- **ROC-AUC: 0.94**
- NLP Classifier is successful in gathering and segregating travel feedbacks from the web
- Enables Marketing and Operations team to leverage on this information to execute their campaigns for upcoming travel fairs/fairs

Recommendations

1. Train on larger validation dataset consisting keywords throughout all 4 seasons of the year

2. Deep Learning techniques such as Recurrent Neural Networks for sequential text data. Pre-trained models on existing larger dataset and predict its performance on r/JapanTravel and r/ThailandTourism

3. Employ the model to train on recommendations and feedback from Singapore reviews to validate model's performance

Thank You

Contact us:

TheSGTravelGuru
555 Siam-Tokyo Street
Singapore, SG 12345

the_sg_travel_guru@example.com
www.TheSGTravelGuru.com

Sources

1. https://www.straitstimes.com/life/travel/bangkok-tokyo-and-bali-are-top-year-end-destinations-among-singapore-travellers
2. https://www.travelandleisureasia.com/sg/news/year-end-travel-destinations-among-singapore-travellers/
3. https://www.straitstimes.com/singapore/singapore-will-lift-remaining-covid-19-border-restrictions-from-feb-13
4. https://www.expedia.com.sg/stories/press/expedia-66-of-singapore-travellers-prioritise-experiences-and-travel-over-things-in-life-post-pandemic/
5. https://www.v7labs.com/blog/transfer-learning-guide#h2