Subreddit Classification NLP: TravelHacks / AwardTravel

DestinationWeddings.com
Where Romance Meets Travel

Aichieh Lin
04/03/2023

Problem Statement

In order to design marketing strategies to meet customer demands, what features can we add to our website using information from both subreddit?

Text Cleaning

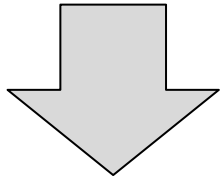
1. Expand contractions
2. Remove special characters and non-letters
3. Lowercase text
4. Remove whitespace
5. Remove special characters
6. Remove emails and weblinks
7. Remove stop words
8. Lemmatize

```
travelhacks.shape, awardtravel.shape
```

```
((3773, 8), (3847, 8))
```

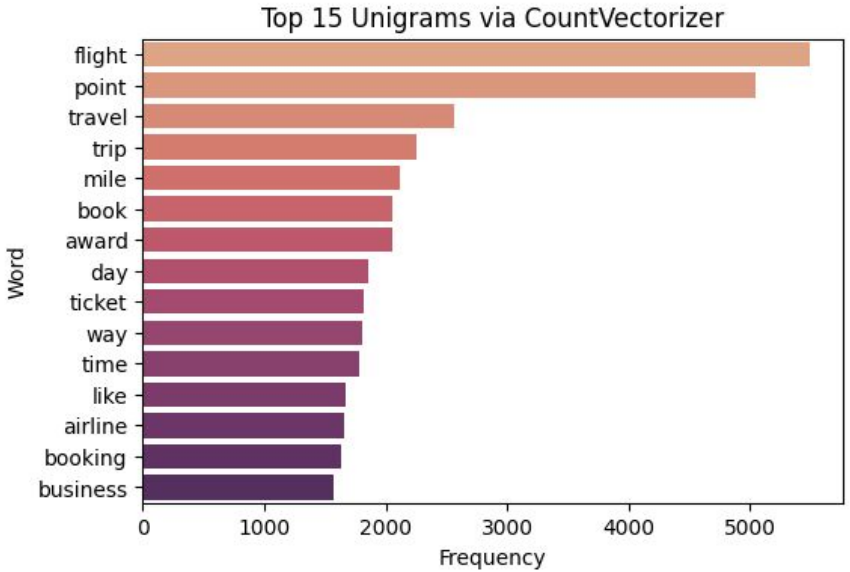
Before and After

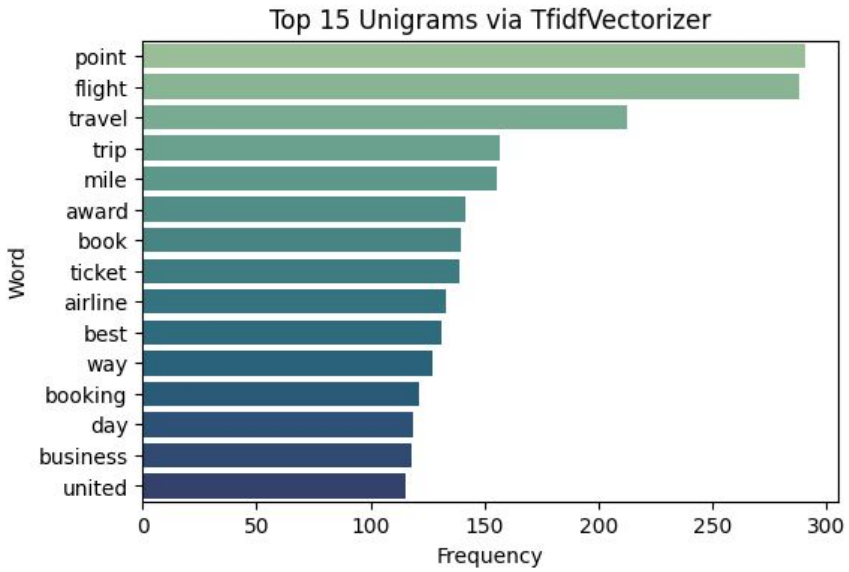
"Sfo to hkg january time frame Got about 400k MR. Would like business for 2 to go from SFO to HKG in mid to late January time frame... Willing to fly out of LAX. \n\nCan't seem to find any on cathay's site from SFO or LAX and united is like 200k each way per person from SFO.\n\nAnyone got any recommendations on what to look at?"



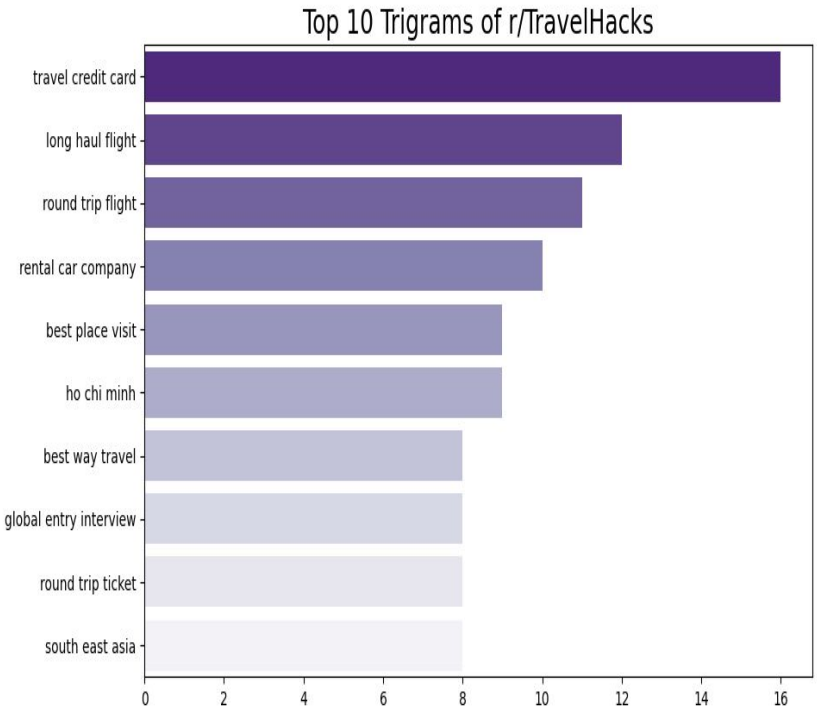
'sfo hkg january time frame got k mr like business sfo hkg mid late january time frame willing fly lax find cathay site sfo lax united like k way person sfo got recommendation look'

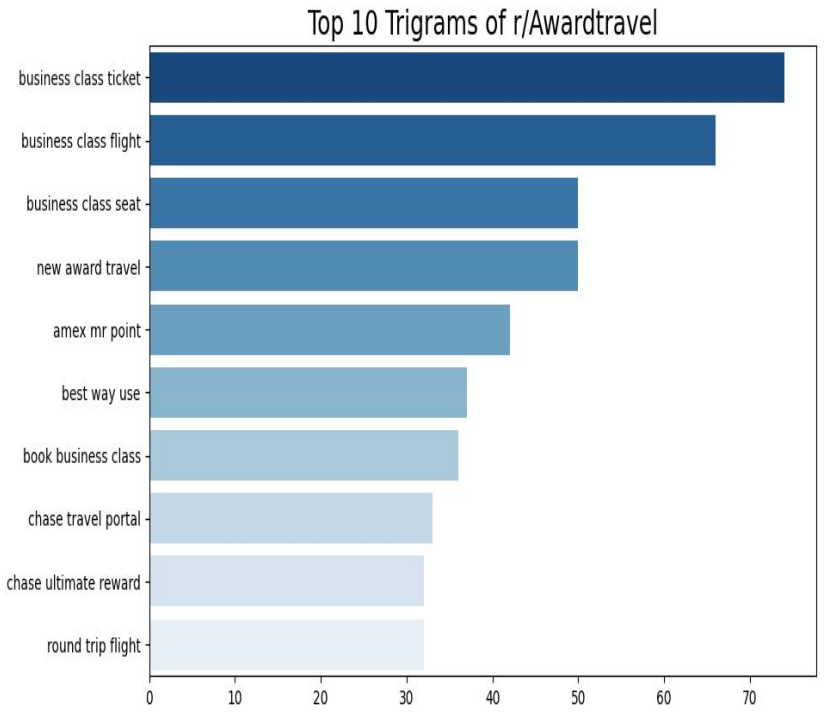
CountVectorizer vs TFIDFVectorizer



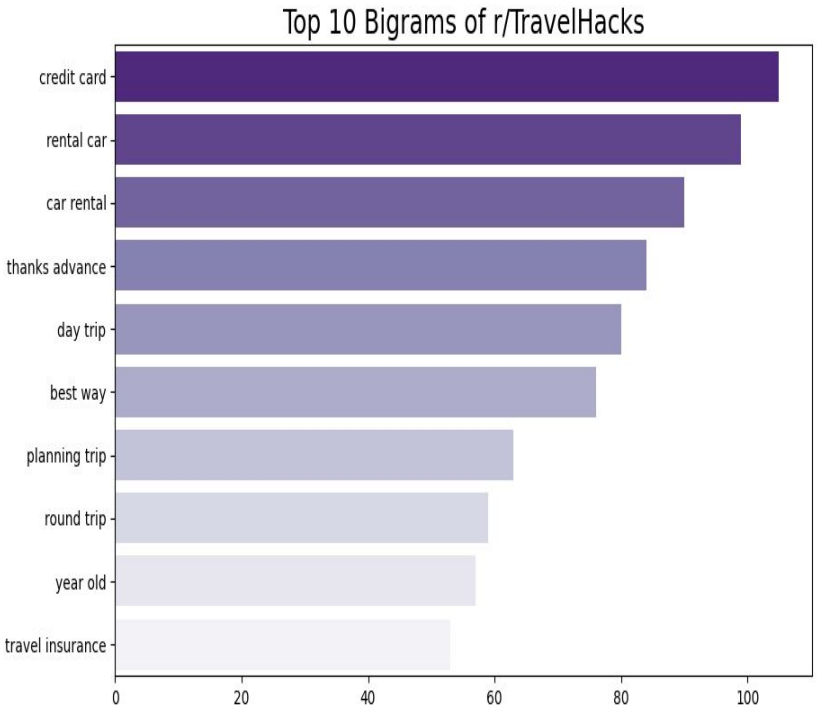


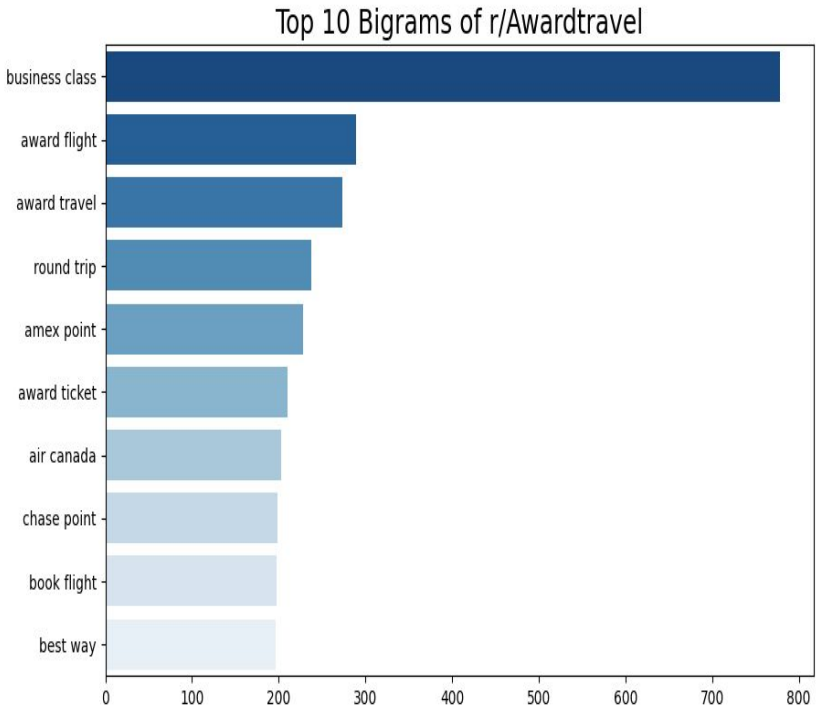
r/TravelHacks vs r/AwardTravel



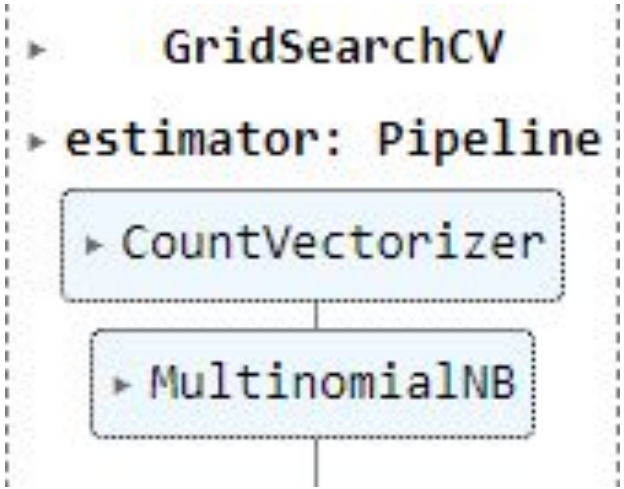


r/TravelHacks vs r/AwardTravel





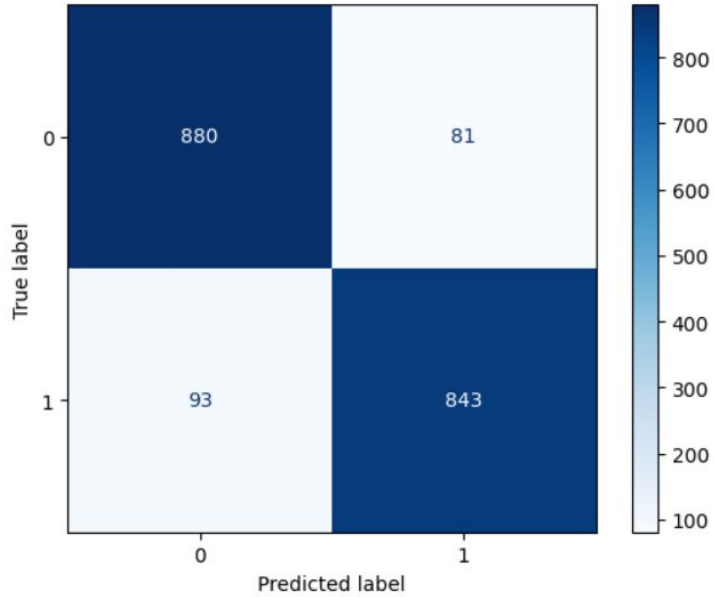
Best Model



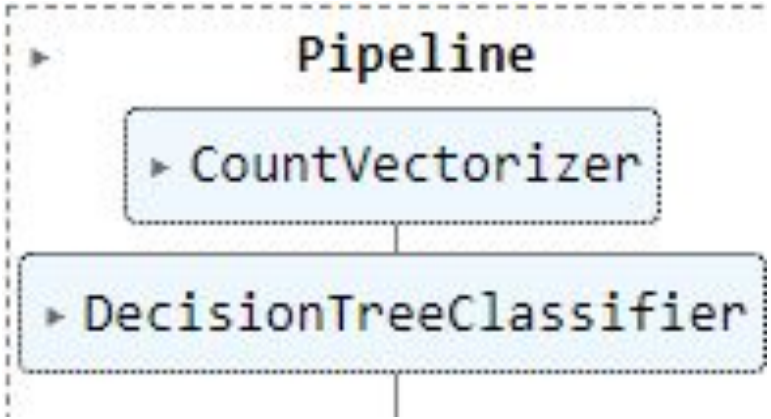
Baseline score: 49%
Testing accuracy: 90.4%
Training accuracy: 93%

Grid Search of CountVectorizer and MultinomialNB
Specificity: 0.9157127991675338
Sensitivity: 0.9006410256410257
Classification Report

	precision	recall	f1-score	support
0	0.90	0.92	0.91	961
1	0.91	0.90	0.91	936
accuracy			0.91	1897
macro avg	0.91	0.91	0.91	1897
weighted avg	0.91	0.91	0.91	1897

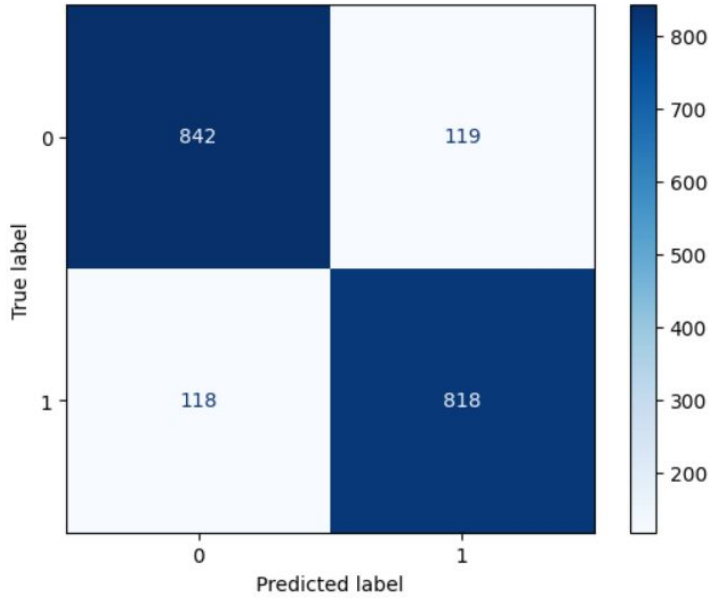


Worst Model



CountVectorizer and DecisionTree
Specificity: 0.8761706555671176
Sensitivity: 0.8739316239316239
Classification Report

	precision	recall	f1-score	support
0	0.88	0.88	0.88	961
1	0.87	0.87	0.87	936
accuracy			0.88	1897
macro avg	0.88	0.88	0.88	1897
weighted avg	0.88	0.88	0.88	1897



Testing accuracy: 88.3%
Training accuracy: 99.9%

Conclusion and Recommendation

- r/TravelHacks and r/AwardTravel is somewhat similar
- frequent bigrams and trigrams suggest:
  - r/awardtravel: points for flight tickets, hotels, credit cards
  - r/travelhacks: rental car, destination, and trip planning
- the best model: MultinomialNB model with CountVector and GridSearch
  - Predict 91% accurately
- the worst model: DecisionTree with CountVector
  - Predict 88% accurately
- major limitations: numbers were removed during text cleaning; therefore, it will be hard to provide any features related to numbers such as cost, distance, and number of points for award redemption.

Reference

1. photo : https://www.danahareldesign.com/our-pinterest-selection-honeymoon-locations
2. Destinationweddings trademark: https://www.destinationweddings.com/?cam=SSEM-Google-EN-US-Search-Destination-Wedding-Generic-EXA&adg=Destination-Wedding_Generic&gclid=CjoKCQjw8qmhBhClARIsANAtbofcpcrmBlL5iDKBrY3pGjebb8BBistOFnZ1QDYa_V7mk1hMKhrlzBUaAtB_EALw_wcB