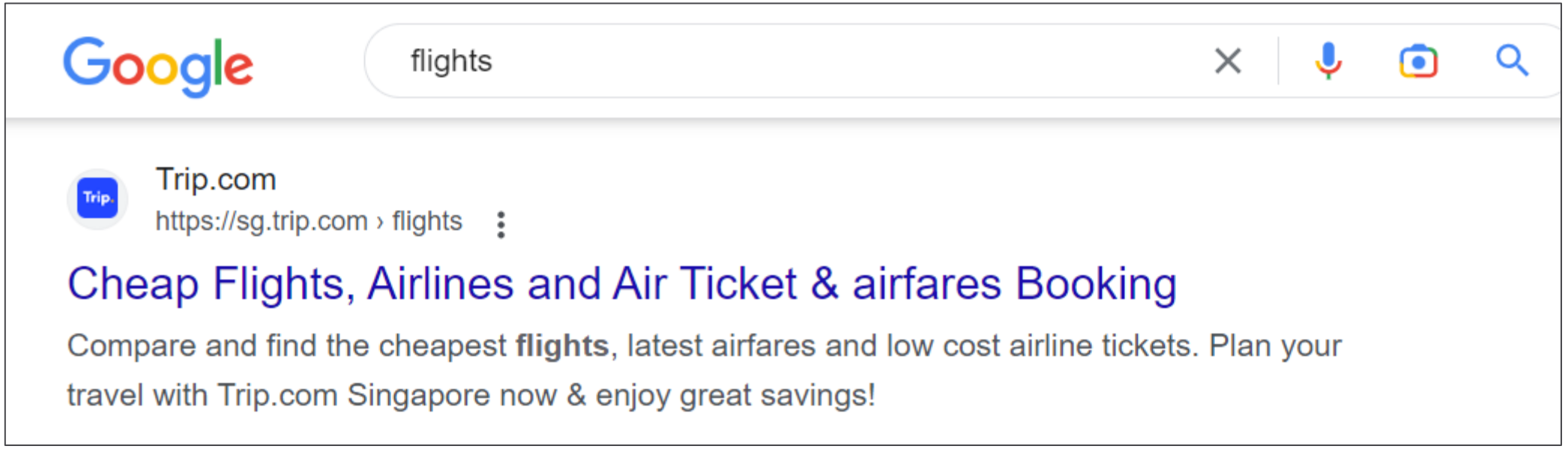
Using Text Classification as a Keyword Strategy for SEO

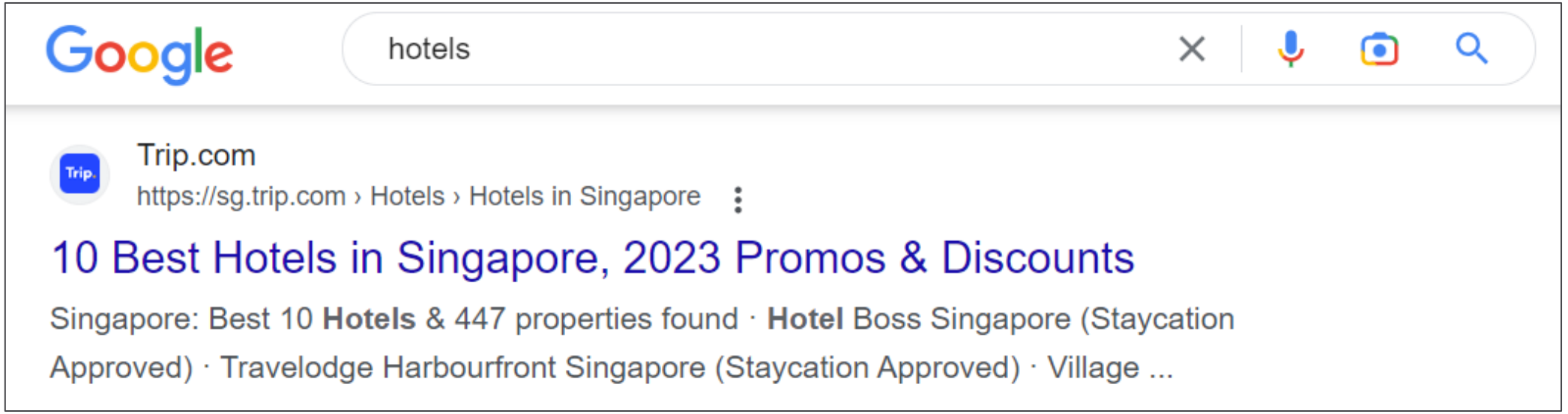
Ng Zeng Di

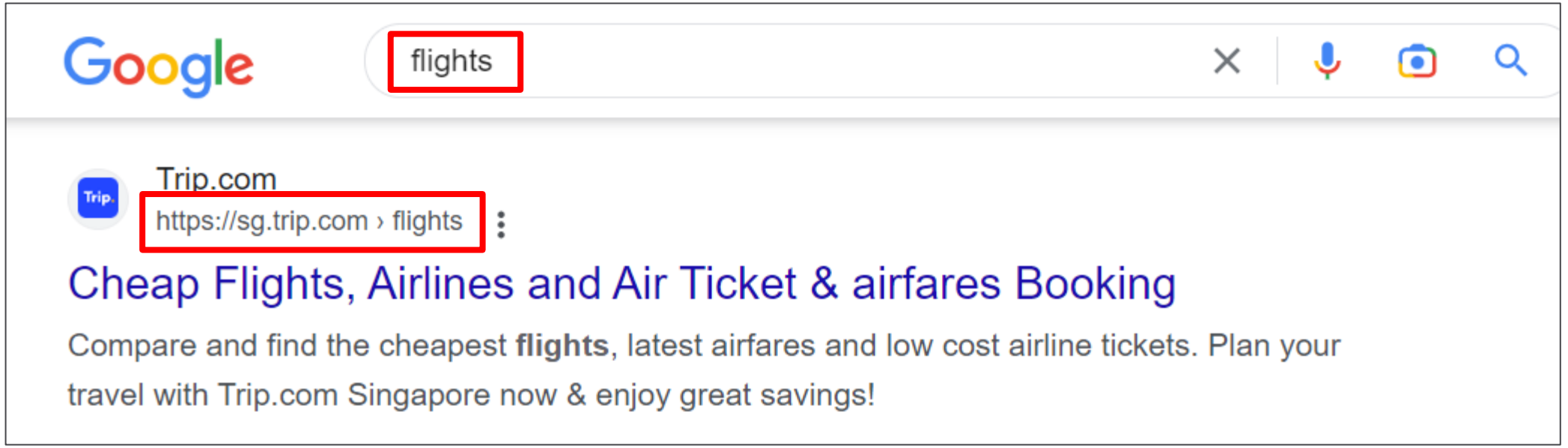
18 March 2023

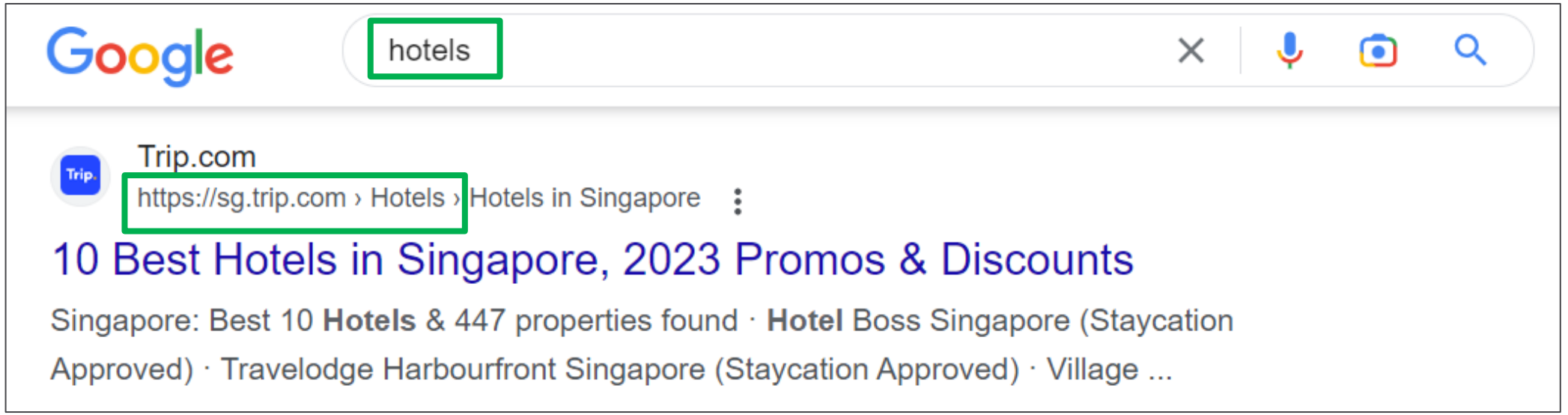
Background

- Trip.com Group wants to expand to the North America Market
- Research long-tail keyword opportunities, target low-competition keywords with high conversion rates


Google
flights
Trip.com
https://sg.trip.com › flights
Cheap Flights, Airlines and Air Ticket & airfares Booking
Compare and find the cheapest **flights**, latest airfares and low cost airline tickets. Plan your travel with Trip.com Singapore now & enjoy great savings!


Google
hotels
Trip.com
https://sg.trip.com › Hotels › Hotels in Singapore
10 Best Hotels in Singapore, 2023 Promos & Discounts
Singapore: Best 10 **Hotels** & 447 properties found · **Hotel** Boss Singapore (Staycation Approved) · Travelodge Harbourfront Singapore (Staycation Approved) · Village ...

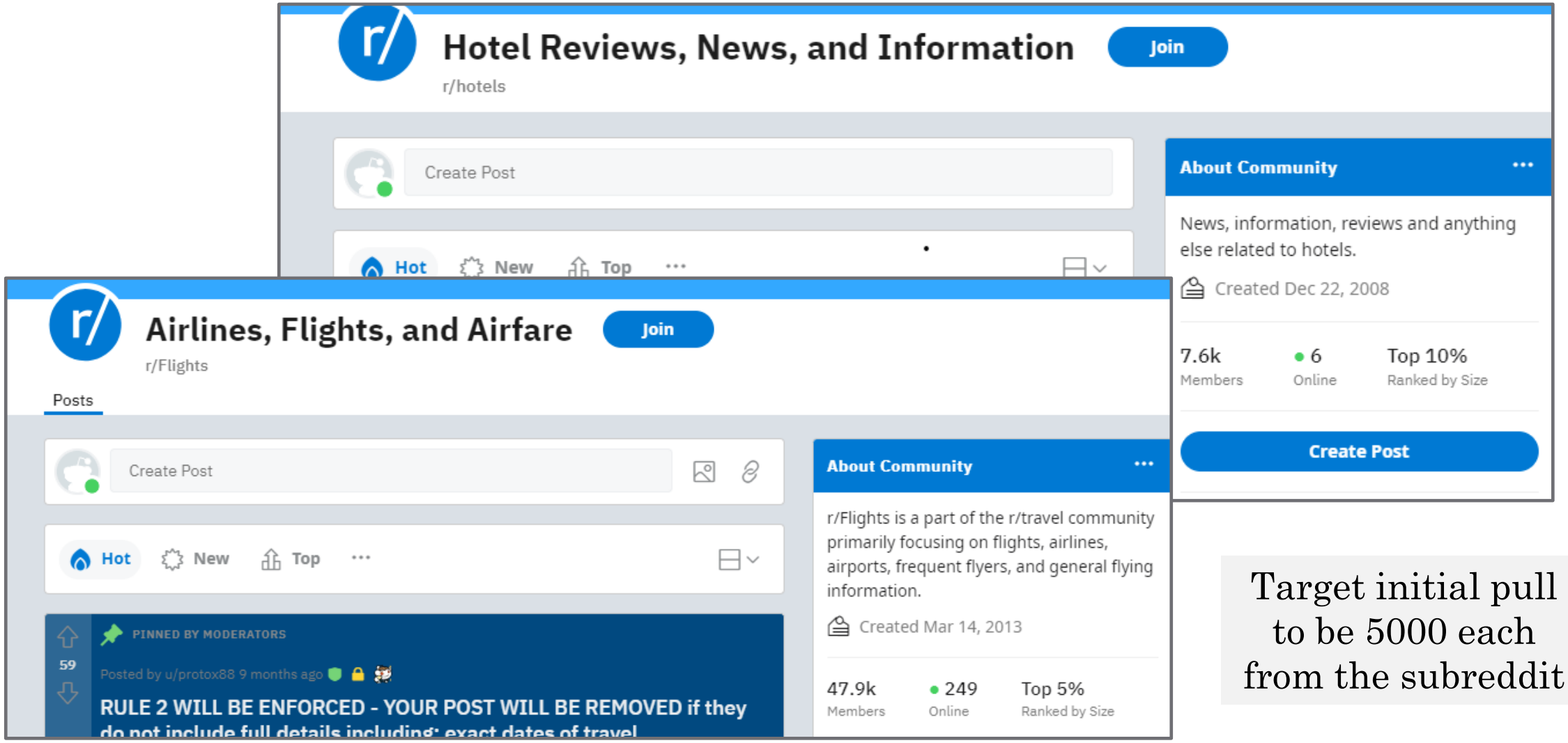

Google
flights
Trip.com
https://sg.trip.com › flights
Cheap Flights, Airlines and Air Ticket & airfares Booking
Compare and find the cheapest **flights**, latest airfares and low cost airline tickets. Plan your travel with Trip.com Singapore now & enjoy great savings!


Google
hotels
Trip.com
https://sg.trip.com › Hotels › Hotels in Singapore
10 Best Hotels in Singapore, 2023 Promos & Discounts
Singapore: Best 10 **Hotels** & 447 properties found · **Hotel** Boss Singapore (Staycation Approved) · Travelodge Harbourfront Singapore (Staycation Approved) · Village ...

Problem Statement

- Do a Keyword research to uncover queries to target and the popularity of these queries.
- We will be using supervised learning technique so we'll need some labeled data to train our model.

Collect Data from Reddit

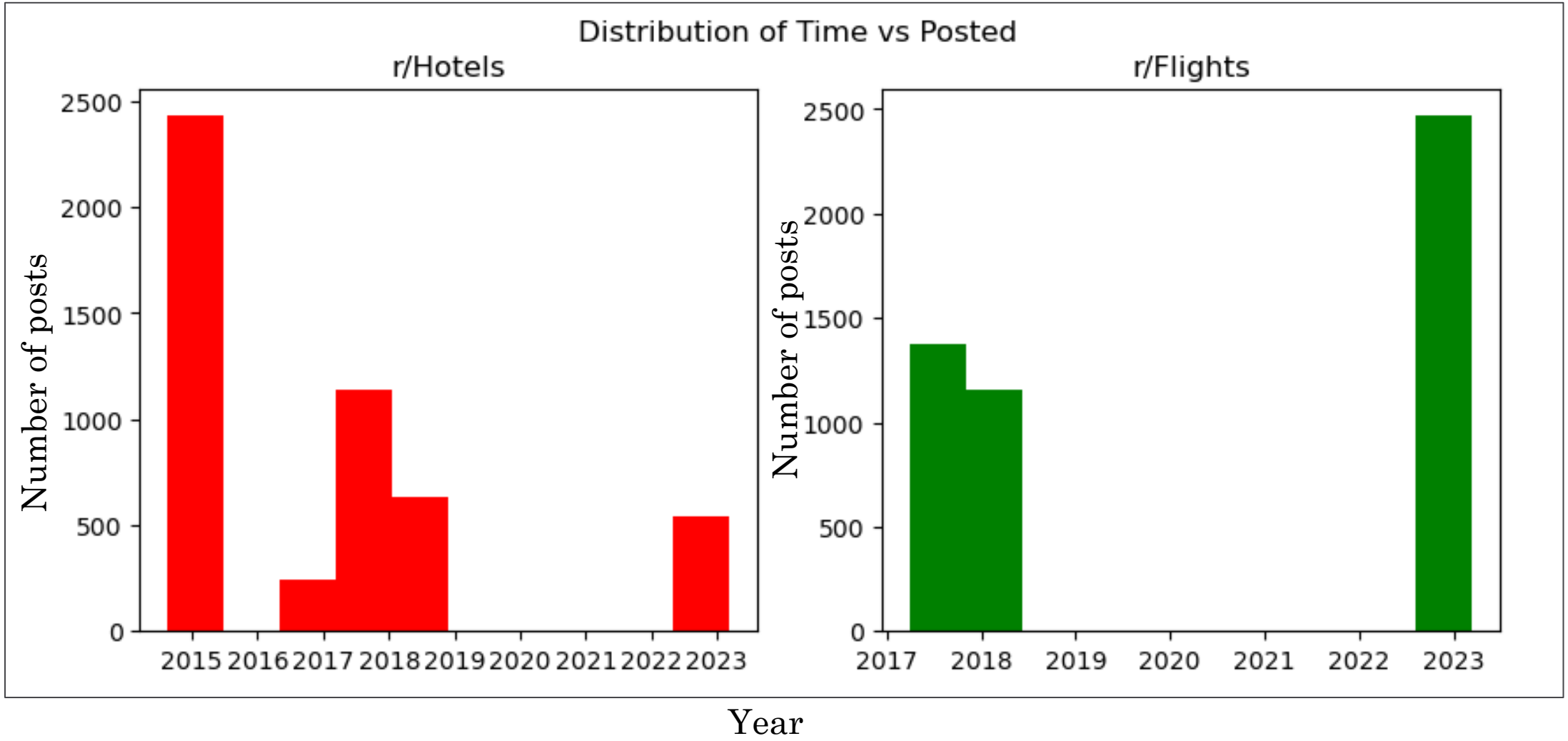


Target initial pull to be 5000 each from the subreddit

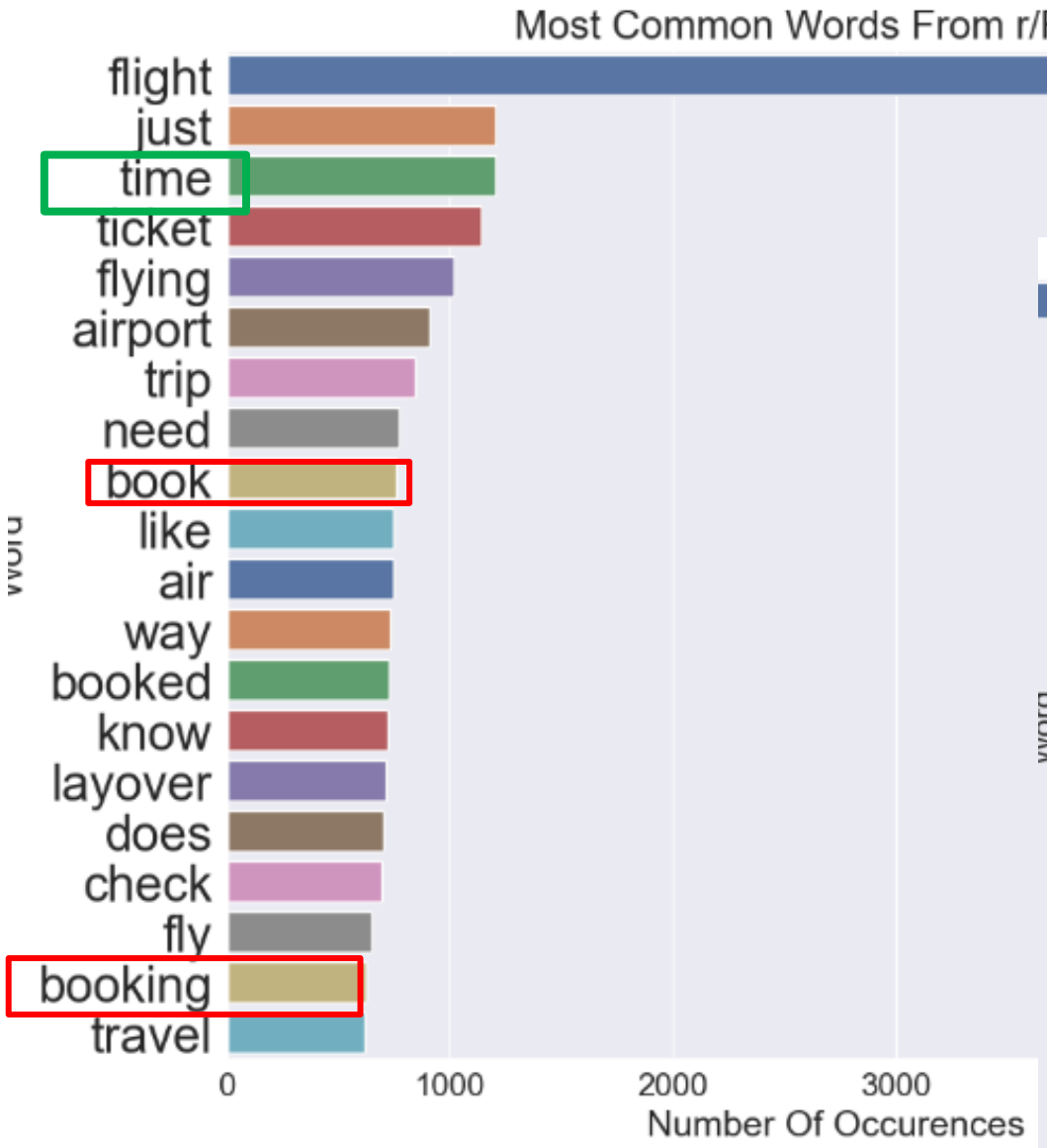
Data Cleaning and EDA

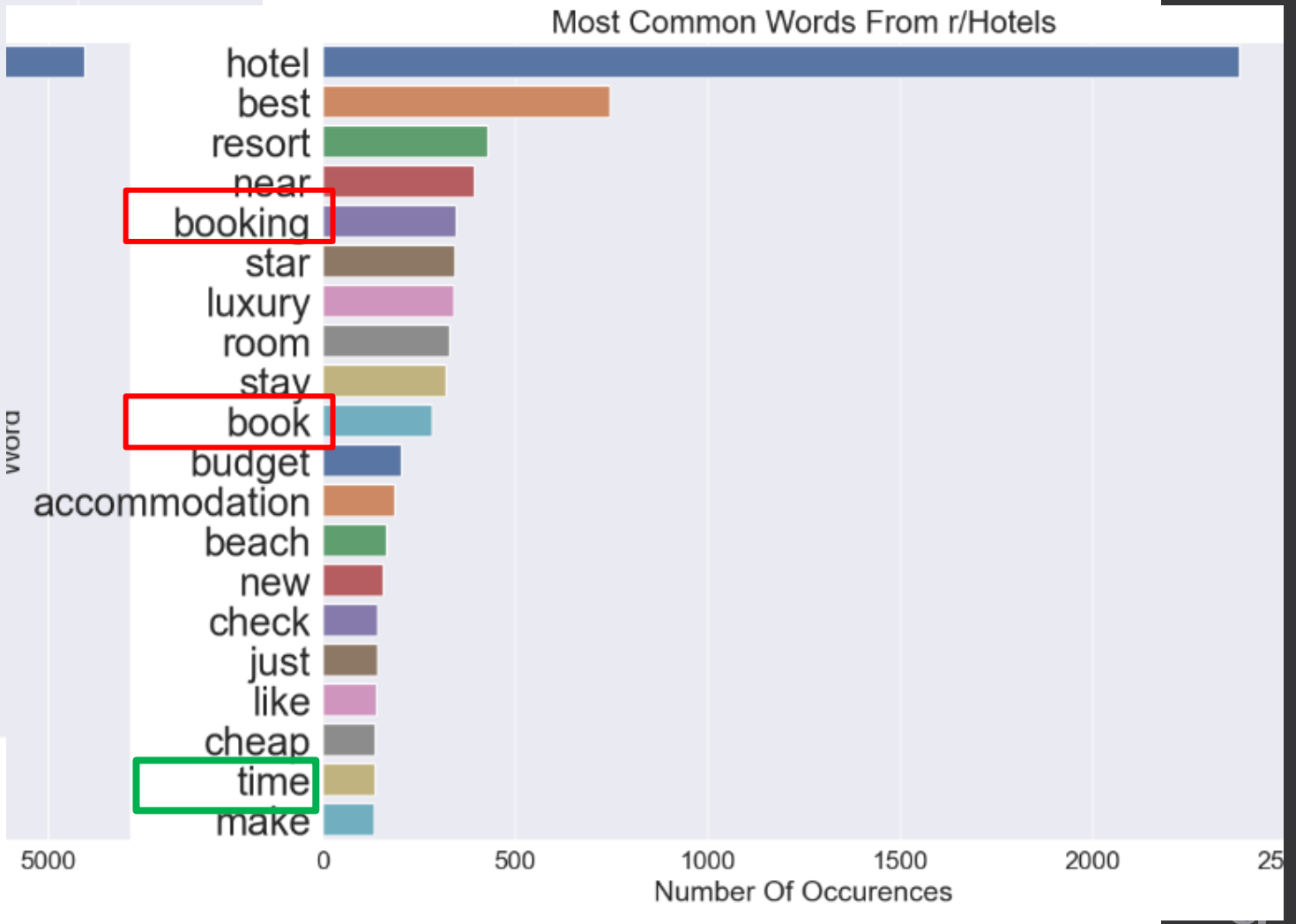
	subreddit	selftext	title	created_utc
0	Flights	To date, I just purchased my cheapest internat...	$9AUD Flight	1678772376
1	Flights	NaN	I'm an engineer with a design to reverse engin...	1678761426
2	Flights	[removed]	TAP AIR FLIGHT SONG	1678756695
3	Flights	Context: I got a flight centre gift card for $...	what would you do? awful rule regarding Porter...	1678753898
4	Flights	[removed]	EVA Air CC Verification	1678749479
5	Flights	So I learned, don't book through credit card t...	Booked through capital one	1678749255
6	Flights	Currently looking at flights from Europe to Ta...	EVA Air 777-300 or 787-9?	1678742265
7	Flights	[removed]	Charged twice for the same piece of checked lu...	1678739090
8	Flights	[removed]\n\n[View Poll](https://www.reddit.co...	Virgin Atlantic vs. Turkish from LHR to Texas?	1678735250
9	Flights	I'm flying from New Dehli to London with a lay...	Etihad Airways allow earlier flights?	1678732821

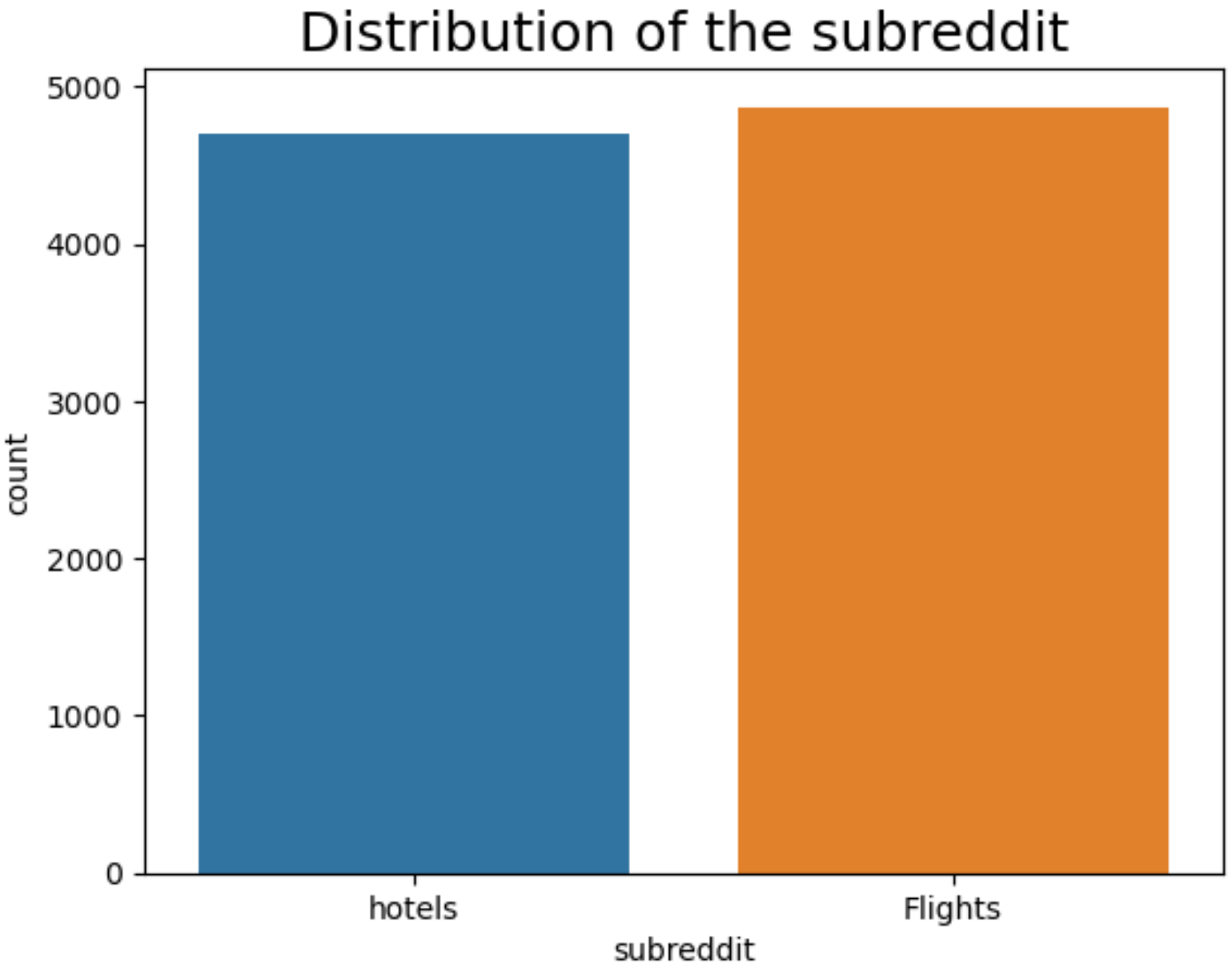
replace null values with '-'
noticed many '[removed]' in selftext
ensure no duplicate columns



- possibility as due to Covid generally no data in 2020-2022


Most Common Words From r/Flights
flight
just
time
ticket
flying
airport
trip
need
book
like
air
way
booked
know
layover
does
check
fly
booking
travel
0
1000
2000
3000
Number Of Occurences


Most Common Words From r/Hotels
hotel
best
resort
near
booking
star
luxury
room
stay
book
budget
accommodation
beach
new
check
just
like
cheap
time
make
5000
0
500
1000
1500
2000
Number Of Occurences

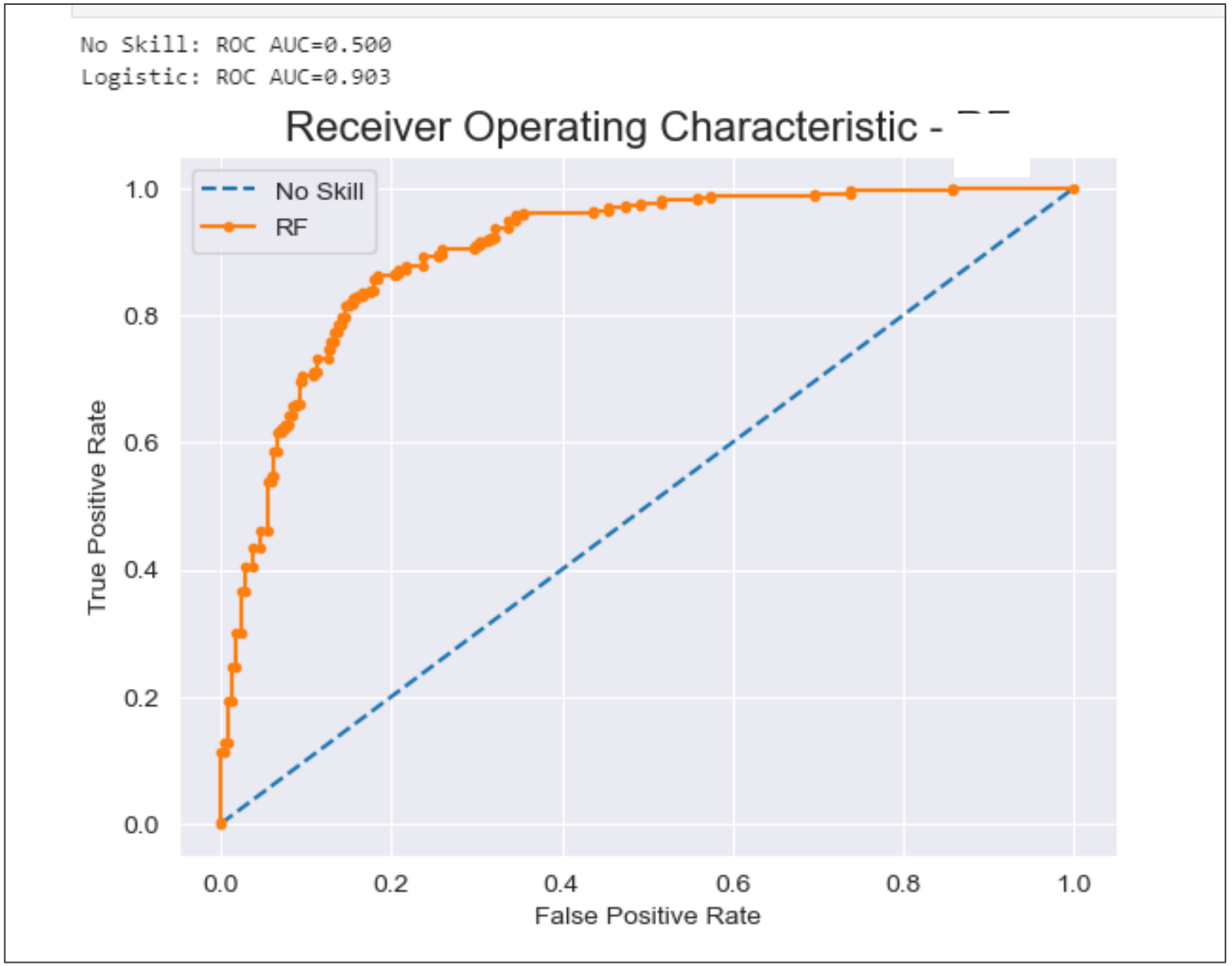

Distribution of the subreddit
5000
4000
3000
2000
1000
0
count
hotels
Flights
subreddit

How Do You Measure the Effectiveness of a Text Classification Model in SEO?

- This can be done through various metrics such as **precision, recall, and F1 score.**
- **A high precision indicates that the model is able to accurately identify the presence of the keyword in texts,**
- For example, in keyword classification, it may be **more important to prioritize precision** in order to **minimize false positives** and avoid targeting the wrong keywords.

	model_name	accuracy_score	precision_score	recall_score	f1_score
	LogisticRegression	0.946708	0.946665	0.946772	0.946699
	Random Forest	0.924765	0.924805	0.924679	0.924731
	Decsision Tree	0.914316	0.914289	0.914289	0.914289
	Multinomial Naive Bayes	0.914316	0.921528	0.913187	0.913767
	K Nearest Neighbor	0.523511	0.725225	0.531753	0.399425

0.94 Precision and Recall score indicates model is good at identifying true positives while minimizing false positives and false negatives.



- Fairly smooth curve, with an AUC of 0.903
- High degree of accuracy in distinguishing between positive and negative examples

Conclusion

- Logistic Regression was the best model, followed by random forest.
- However, the training score (0.99) is higher than the testing score (0.925), indicating overfitting.
- Ways to help:
  - Pull data from Quora as well to increase the training data
  - Use regularization like L2 or data augmentation techniques adding noise to the data.