Backpacking or Travel?

Using Naïve Bayes Model and Random Forest Classifier to classify subreddit posts

Agenda

PROBLEM STATEMENT

DATA BACKGROUND

EXPLORATORY DATA ANALYSIS

MODELING PROCESS AND RESULTS

TAKEAWAYS AND RECOMMENDATION

Problem Statement

- An unfortunate power outage on some Reddit servers has caused some posts (from r/backpacking and r/travel) to be stored incorrectly within the servers
- As an employee of Reddit, my supervisor has tasked me to **correctly reclassify these posts** by training classifier models to solve this issue
- We will be training the models based on about **2000 reddit posts** (about 1000 posts from each subreddit)

Data Background

- **Pushshift API**
  - 100 posts per requests
  - Removed any duplicated posts
- **Cleaning**
  - Dropped 6 rows containing missing values
  - Checked that there are also no mod bot messages
  - Removed posts containing ‘[removed]’ (2 rows)
  - Lowercased all words and removed hyperlinks, white spaces, numbers
- **Preprocessing**
  - Lemmatize words (days -> day, nights -> night)
  - Added to stop words: ‘backpacking’, ‘travel’ plus other generic words

r/backpacking
(cleaned)
993 posts

r/travel
(cleaned)
1045 posts

Combined
2038 posts

Exploratory Data Analysis

r/backpacking

- **832** unique users
- **1.19** post per user
- Longest post by word count: **7,493** words (trip report)
- Shortest post by word count: **6** words (title of an image)

- Most common Bigram and trigrams

r/travel

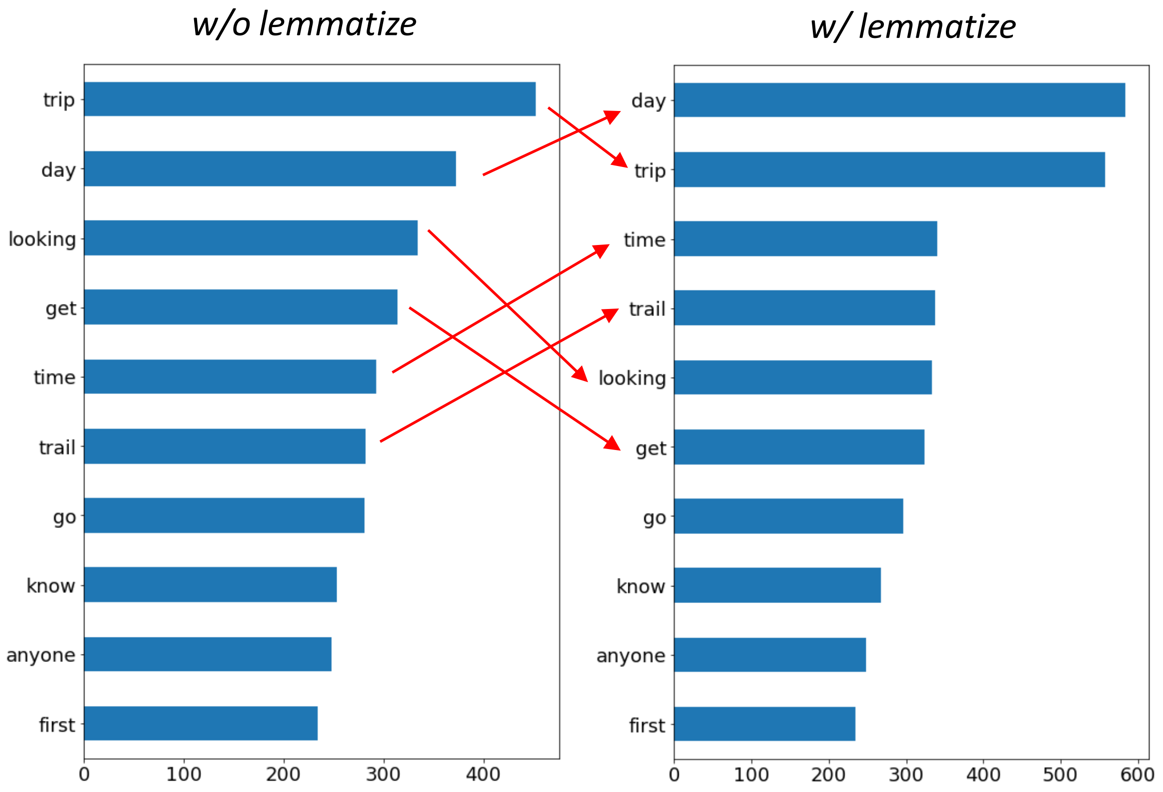
- **969** unique users
- **1.08** post per user
- Longest post by word count: **2,535** words (covid restriction discussion while traveling)
- Shortest post by word count: **3** words (user replying via post to thank someone)

- Most common Bigram and trigrams

Exploratory Data Analysis

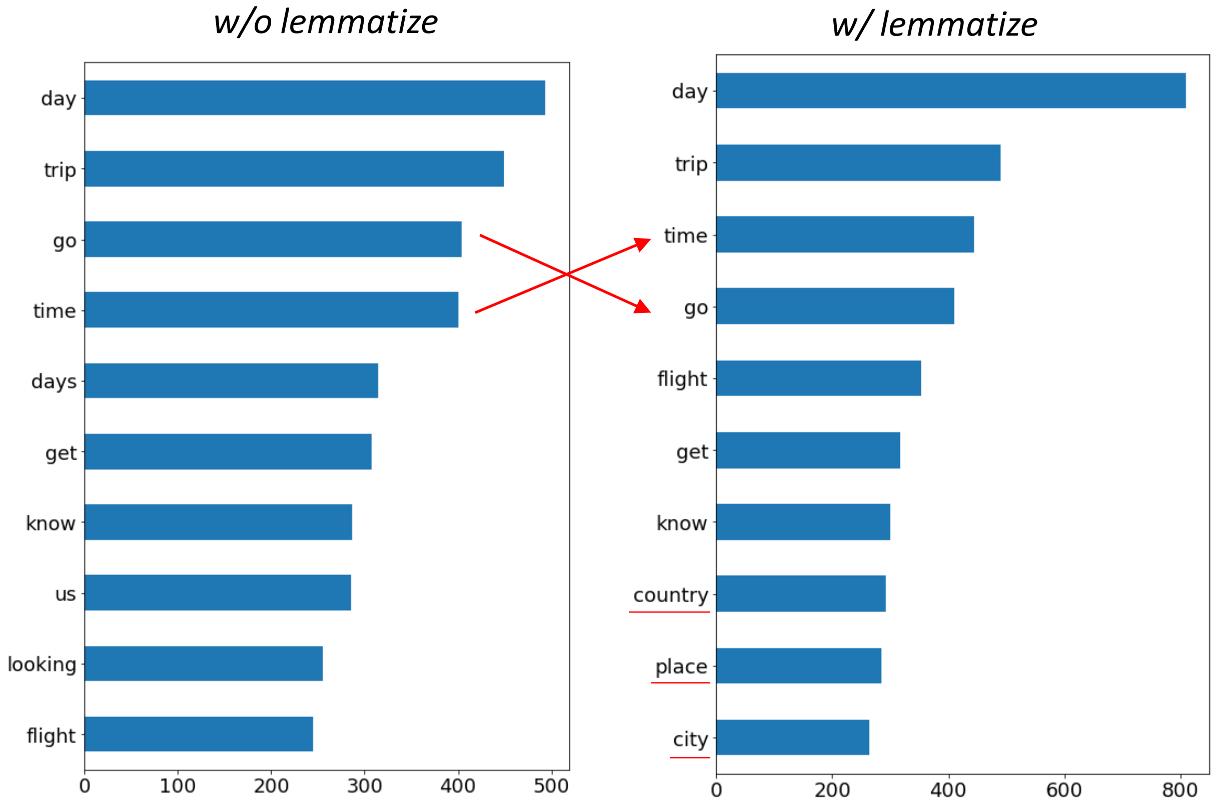
- 10 most frequent words using CountVectorizer

r/backpacking


w/o lemmatize

w/ lemmatize

r/travel


w/o lemmatize

w/ lemmatize

Modeling Process and Results

1. Train test split: stratify y, setting a random state to rerun models
2. Fit and run models using Pipeline and GridSearchCV:
    1. 2 models: Naïve Bayes and Random Forest
3. Baseline for each model is the default hyperparameters using CountVectorizer

Naïve Bayes

Tf – IDF Vectorizer

GridSearch best hyperparameters:

- 'nb__alpha’: 0.5
- 'tvec__max_features’: 7000
- 'tvec__ngram_range': (1, 2)

Train score: 0.8494

Test score: 0.8216

Random Forest

Tf – IDF Vectorizer

GridSearch best params:

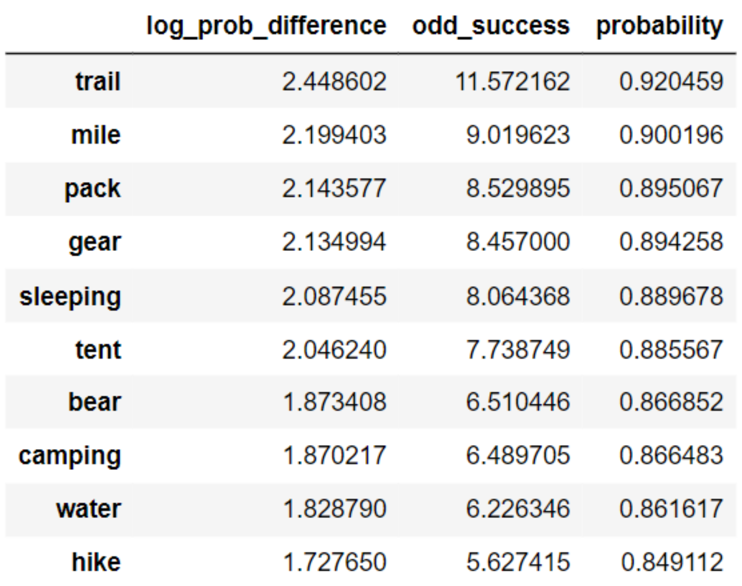
- 'rf__max_depth': None
- 'rf__n_estimators': 200
- 'tvec__max_features': 10000
- 'tvec__ngram_range': (1, 3)

Train score: 0.8220

Test score: 0.8098

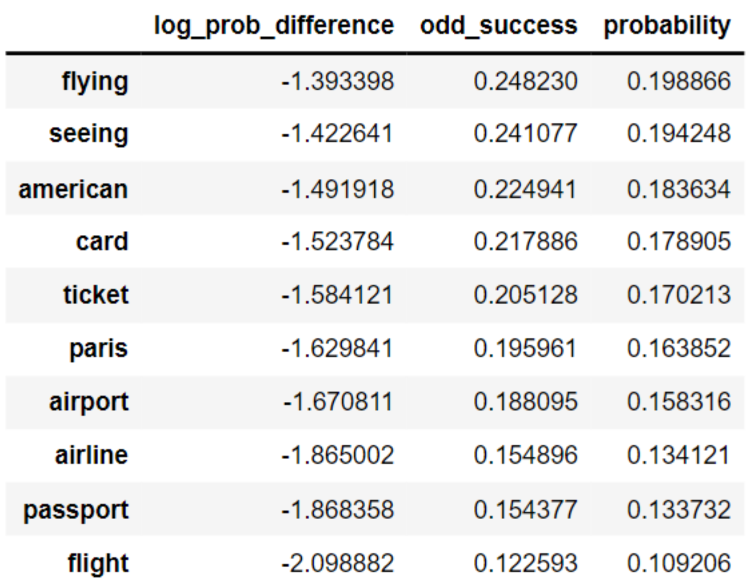
Modeling Process and Results
Feature importance of Naïve Bayes

Top 10 word contributors to differentiate backpacking post from travel post



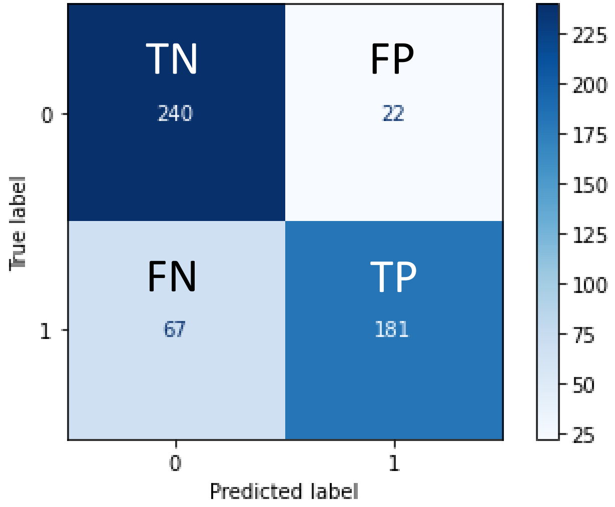
	log_prob_difference	odd_success	probability
trail	2.448602	11.572162	0.920459
mile	2.199403	9.019623	0.900196
pack	2.143577	8.529895	0.895067
gear	2.134994	8.457000	0.894258
sleeping	2.087455	8.064368	0.889678
tent	2.046240	7.738749	0.885567
bear	1.873408	6.510446	0.866852
camping	1.870217	6.489705	0.866483
water	1.828790	6.226346	0.861617
hike	1.727650	5.627415	0.849112

Bottom 10 word contributors to differentiate backpacking post from travel post



	log_prob_difference	odd_success	probability
flying	-1.393398	0.248230	0.198866
seeing	-1.422641	0.241077	0.194248
american	-1.491918	0.224941	0.183634
card	-1.523784	0.217886	0.178905
ticket	-1.584121	0.205128	0.170213
paris	-1.629841	0.195961	0.163852
airport	-1.670811	0.188095	0.158316
airline	-1.865002	0.154896	0.134121
passport	-1.868358	0.154377	0.133732
flight	-2.098882	0.122593	0.109206

Misclassification Analysis on Best Model: Naïve Bayes



- **Accuracy score:** 82.16%
- **Subreddit:**
  - 0: Backpacking
  - 1: Travel
- **False positives:** posts that incorrectly classified as backpacking
- **False positives:** posts that incorrectly classified as travel
- Most misclassified posts were **long posts**
  - Average word count: 97 words
  - The longest post being 721 words
- The most common words were: day, trip, time, go, get

Takeaways and Recommendations

- For the 2 subreddits: Naïve Bayes marginally performs better than Random Forest.
- Surprisingly, the concern for the naïve assumption that all features are independent has minimal impact to the model's capability to classify the reddit posts accurately

- Naïve Bayes or Random Forest?

Naïve Bayes	Random Forest
Good: • Easy to train and understand the results • It has different extensions for different needs • Computes faster	Good: • Random forest method works for all types of data {numeric, cardinal, ordinal}
Bad: • Assumes all variables are uncorrelated but generally not true	Bad: • Takes time to train and consumes more time to predict proportional to the number of trees (computationally more expensive)