Natural Language Processing

Classifying posts from subreddits

- **Who we are: Goals of the project**
- Text Data Exploration
- Results
- Analysis

We are

- Gareth
- Chris
- Nemo

We are **classifying posts** from the subreddits **r/JapanTravel** and **r/solotravel**

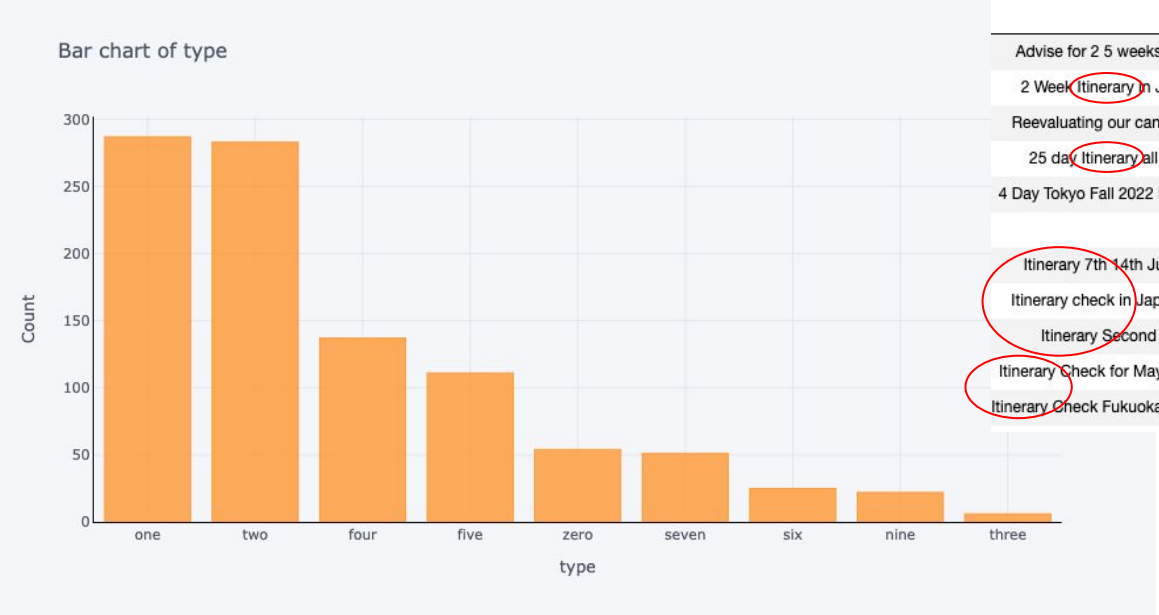
For easy tracking, allowing employees to focus their energies elsewhere

What was done?

- Scraped data from the subreddits using Reddit's API
- Saved the 'titles', 'body' (called 'selftext'), 'flair' and 'subreddit' into a dataframe
- Explored the texts and cleaned it
- Applied Naive Bayes and Logistic Regression, picking the best model

- Who we are: Goals of the project
- **Text Data Exploration**
- Results
- Analysis

Exploratory Data Analysis



Type: one

title_post
Advise for 2 5 weeks in JapanHi I am looking f...
2 Week Itinerary in Japan with a Father his 16...
Reevaluating our cancelled march 2020 trip to ...
25 day Itinerary all over Japan during July of...
4 Day Tokyo Fall 2022 Honeymoon Itinerary Any ...
...
Itinerary 7th 14th JuneHey all My wife and I h...
Itinerary check in Japan Tokyo 6 8 Hakone 8 9 ...
Itinerary Second time ideas First time itiner...
Itinerary Check for May Tokyo Kiso Fukushima K...
Itinerary Check Fukuoka Hiroshima Kyoto Kanaza...

Type: two

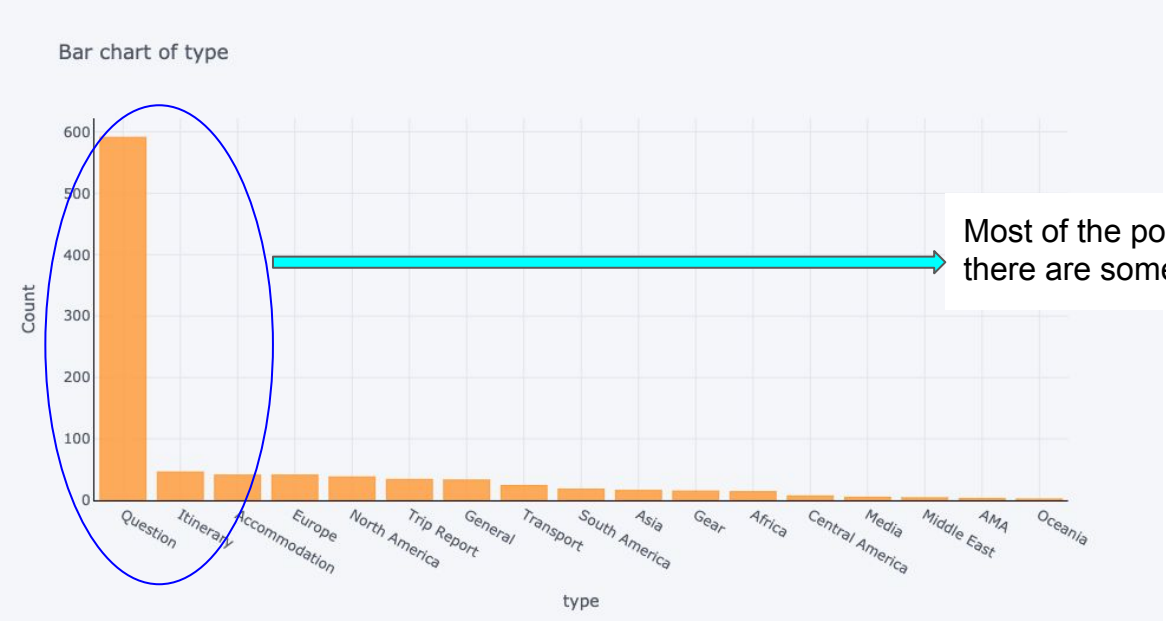
title_post
Question Car rental in Hokkaido for 6 peopleHi...
Google Reviews RestaurantsHello I have noticed...
AirBnB in JapanHow have your experiences been ...
Has anyone visited the grave of Jesus Christ i...
Recieved gifts after eating at a small restaur...
...
If I m wearing a yukata or kimono does it matt...
Does Nintendo Tokyo have long lines still I ll...
Osaka to Arima Onsenl have several logistical ...
Specific question about the Fuji Hakone Pass H...
How to procure Square Enix Cafe tickets Hi guy...

Type: four

title_post
5 Days in KYOTO What to skip Planning to visit...
Good Kabuki theatre to go to My friends and I ...
Planning a solo two week dream trip to Japan r...
Hey guys I have 10 hour layover at Narita Airp...
Romantic view In tokyoHey ors So I ve always w...
...
Ryokan recommendation in Nara My wife and daug...
Best place s to get sailor moon and other anim...
Trainspotting where to see JR Class 103 in Mar...
What s the most enjoyable Maid Cafe in Tokyo O...
Tokyo area animal tourism I want to surprise m...

Type one is mainly concerned with itinerary, while two and four are just general questions

Exploratory Data Analysis



Most of the posts in the subreddit are queries, while there are some posts on itineraries

Number of strings

Number of strings in

- JapanTravel subreddit: 406646
- solotravel subreddit: 159216

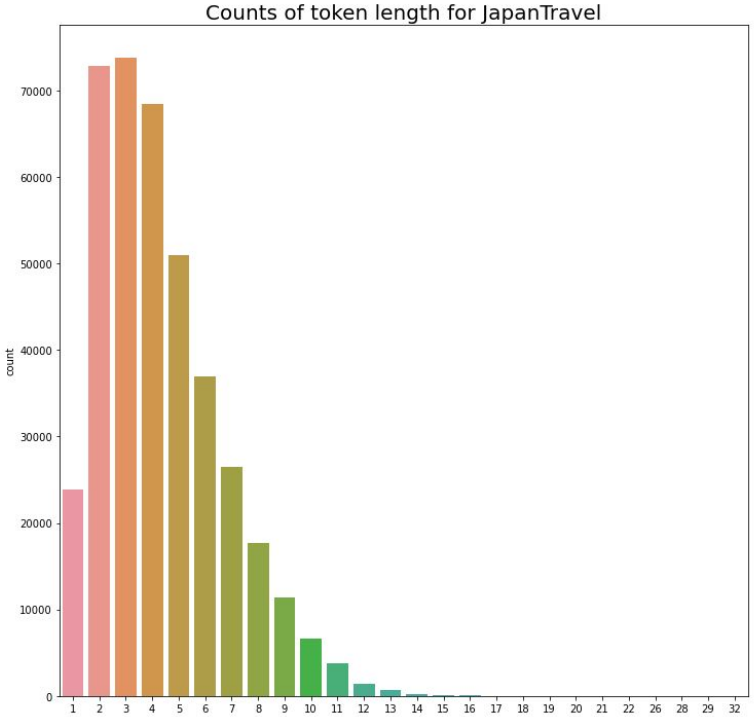
solotravel has **39%** as many strings as JapanTravel!

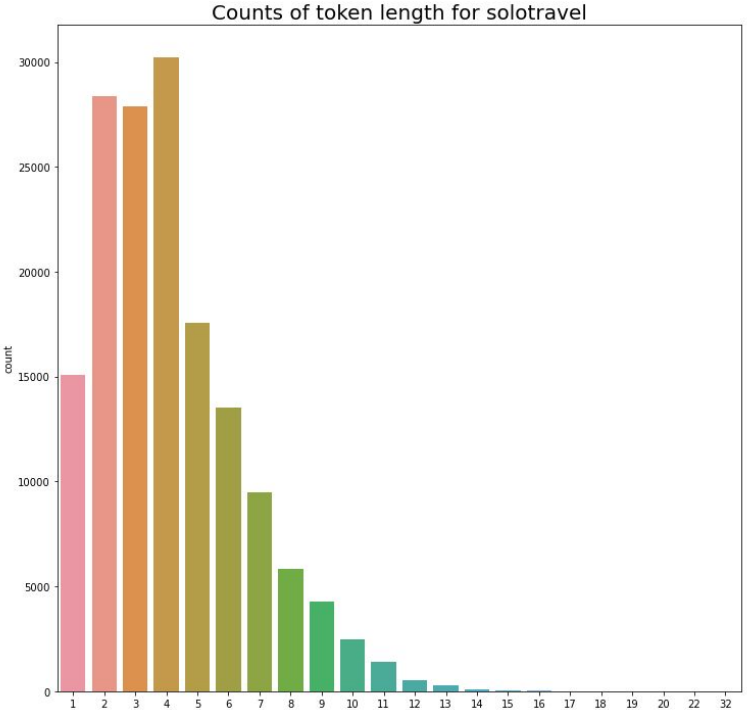
Number of unique strings in

- JapanTravel subreddit: 17366
- solotravel subreddit: 10411

solotravel has **59%** as many unique strings as JapanTravel!

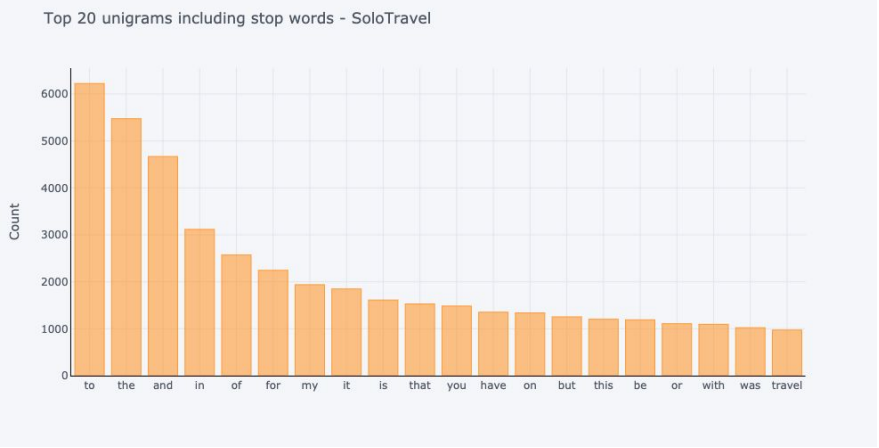
Token Length (no stop words)





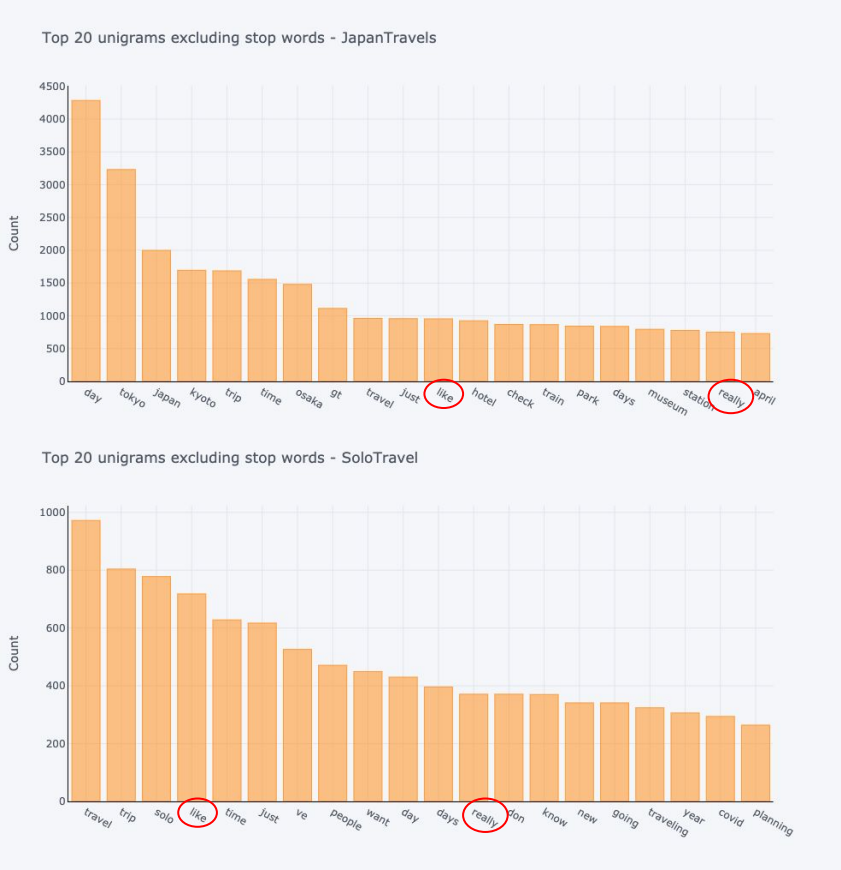
Average no. of characters in JapanTravel: **4.3**
Average no. of characters in solotravel: **4.1**

Exploratory Data Analysis



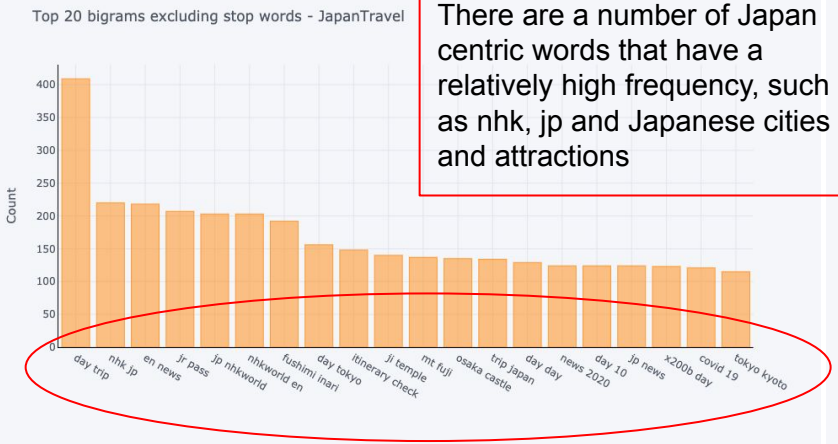
With the inclusion of stop words, JapanTravel subreddit shares 15 out of 20 highest frequency words

Exploratory Data Analysis

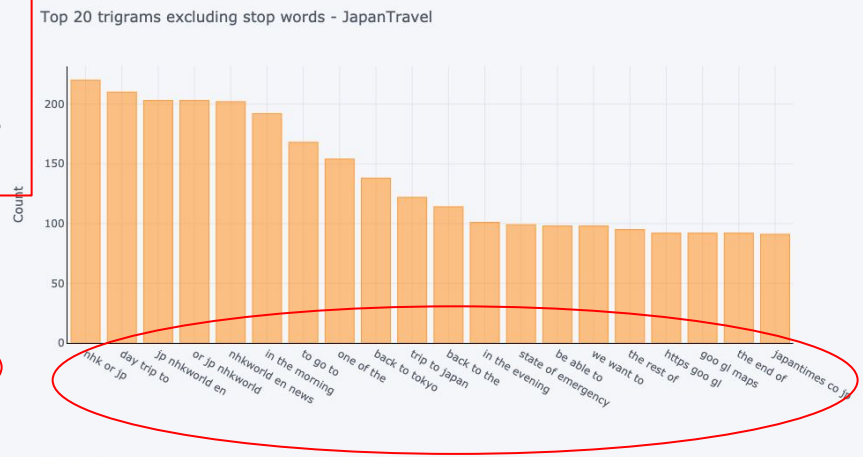


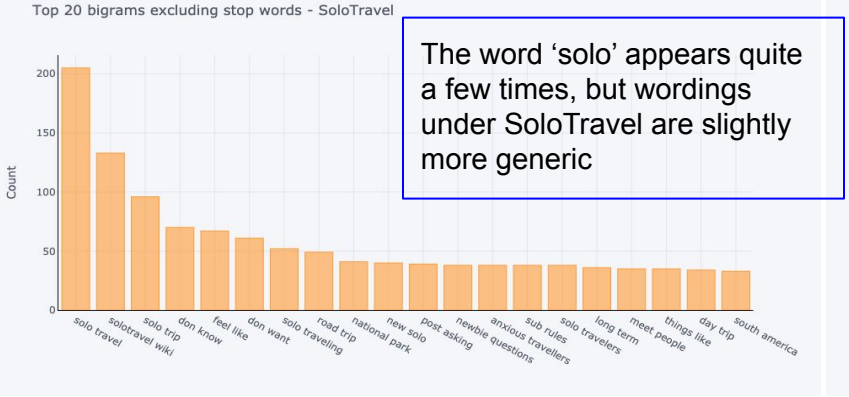
- With the exclusion of stop words, the words start to look a little more identifiable between the 2 subreddits
- They still share a couple of common words such as ‘like’ and ‘really’

Exploratory Data Analysis

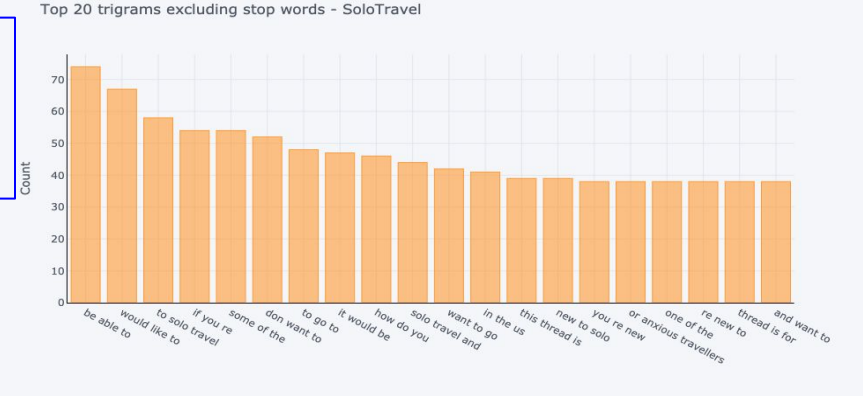


There are a number of Japan centric words that have a relatively high frequency, such as nhk, jp and Japanese cities and attractions





The word ‘solo’ appears quite a few times, but wordings under SoloTravel are slightly more generic

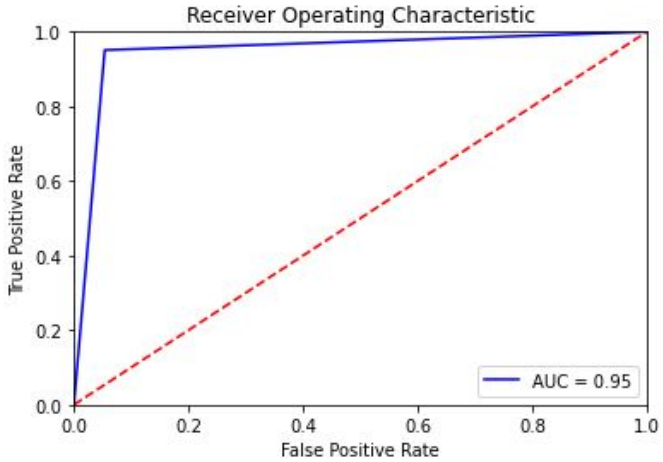


Model Selection

Model	Training Accuracy	Test Accuracy	**Difference**
Baseline (Select majority)	52.3%	52.3%	**0%**
Elasticnet (Countvec, **stop words removed**)	94.6%	60.3%	**34.3%**
Multinomial NB (Countvec, **stop words removed**)	87.4%	60.3%	**27%**
Elasticnet(TfidfVectorizer, **Lemmatization**, **w/ stop words**)	96.1%	94.9%	**1.2%**
Multinomial NB (TfidfVectorizer, **Lemmatization**, **w/ stop words**)	91.7%	92.2%	**0.5%**

Model Performance

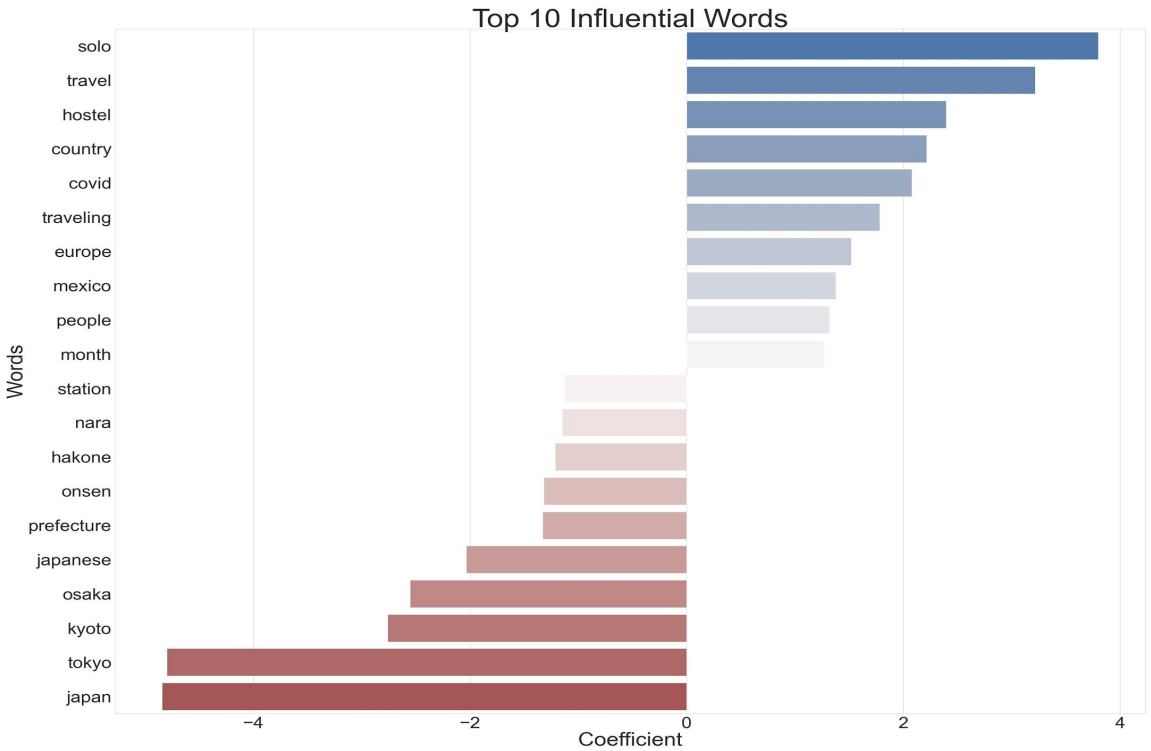
Logistic Regression with Elasticnet




```
Confusion matrix for Logistic Regression test results with tfidf and word lemmatization:

[[211  12]
 [ 12 233]]

True Negatives: 211
False Positives: 12
False Negatives: 12
True Positives: 233

Accuracy: 0.9487179487179487
Recall: 0.9510204081632653
Precision: 0.9510204081632653
```

	Predicted solotravel	Predicted JapanTravel
Actual solotravel	211	12
Actual JapanTravel	12	233

Analysis



- Calculates the coefficient for each word feature to determine influence of words
- An increase of word count by 1 increases odds of classification to SoloTravel
- Positive Coefficients → SoloTravel
- Negative Coefficients → Japan Travel

How did we get the production model?

- Adding **'title' to ‘main post’** gave the model more information to work with
- **Lemmatizing** words increased token frequency
- **Tf-idf vectorizer is better** than Count vectorizer

- **Logistic Regression**: Using ElasticNet prevented overfitting
- **Naive Bayes:** Captures initial beliefs, which helps in giving more accurate predictions

Conclusions

- Model Accuracy of 95%
- JapanTravel → Provide itineraries
- SoloTravel → Share experiences from fellow solo travellers

Further Exploration

- Include comments section
- Increase number of posts
- Implement notification system for new posts that alerts when new posts are classified