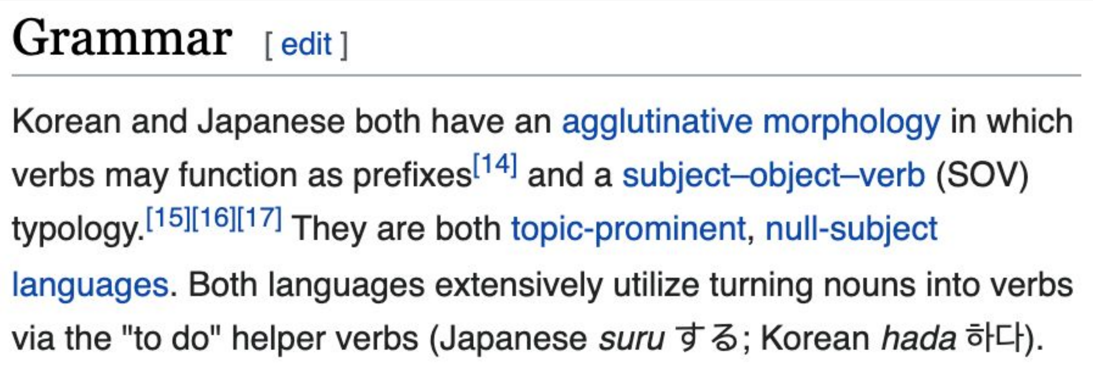
Japanese or Korean?

Using subreddit posts to help choose which language to learn

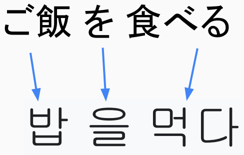
Problem Statement

This project aims to help potential learners who wish to pick up a foreign languages decide on which language is more favourable. The scope is narrowed down to Japanese and Korean, languages that are similar on many fronts.



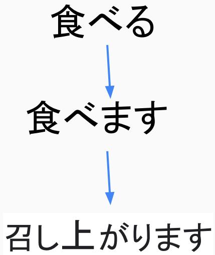
Grammar [edit]

Korean and Japanese both have an agglutinative morphology in which verbs may function as prefixes[14] and a subject–object–verb (SOV) typology.[15][16][17] They are both topic-prominent, null-subject languages. Both languages extensively utilize turning nouns into verbs via the "to do" helper verbs (Japanese *suru* する; Korean *hada* 하다).



ご飯 を 食べる

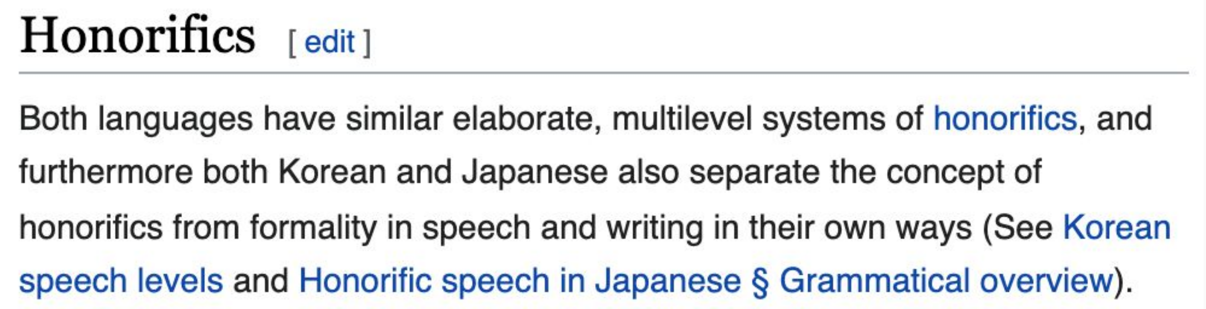
밥 을 먹다



食べる

食べます

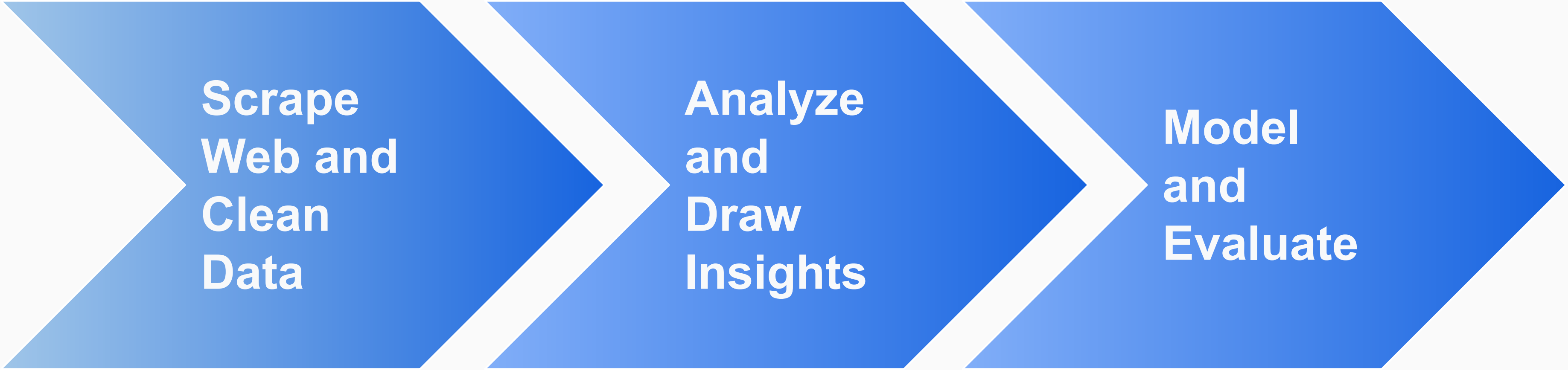
召し上がります



Honorifics [edit]

Both languages have similar elaborate, multilevel systems of honorifics, and furthermore both Korean and Japanese also separate the concept of honorifics from formality in speech and writing in their own ways (See Korean speech levels and Honorific speech in Japanese § Grammatical overview).

Process



Scrape Web and Clean Data

Analyze and Draw Insights

Model and Evaluate

Web Scraping

1. 1000 posts from
   - LearnJapanese (https://www.reddit.com/r/LearnJapanese/new)
   - korean (https://www.reddit.com/r/korean/new)

2. From start date 1 Jan 2021 GMT
   - Last post in Japanese subreddit on 24 Jan 2021 GMT
   - Last post in Korean subreddit on 2 Feb 2021 GMT

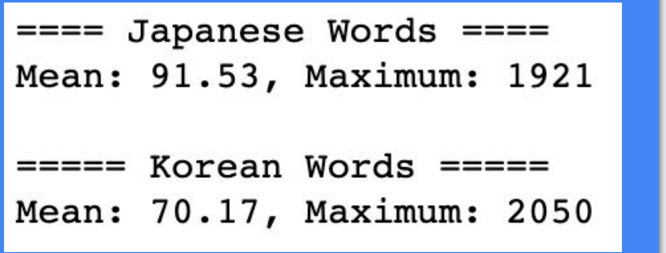
Data Cleaning

1. ‘Title’ and ‘selftext’ merged into single column ‘post’
2. Unwanted information removed
   - urls
   - ‘[removed]’
   - numbers and punctuations
3. Changed to lowercase
4. Dropped duplicates

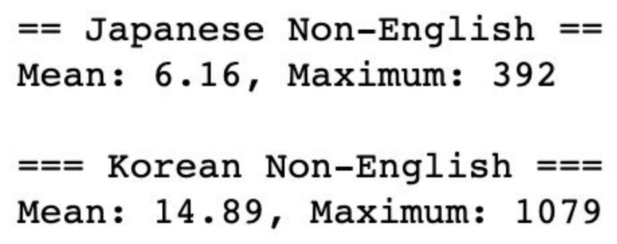
995 Japanese posts and 976 Korean posts available for analysis after cleaning.

Data Analysis

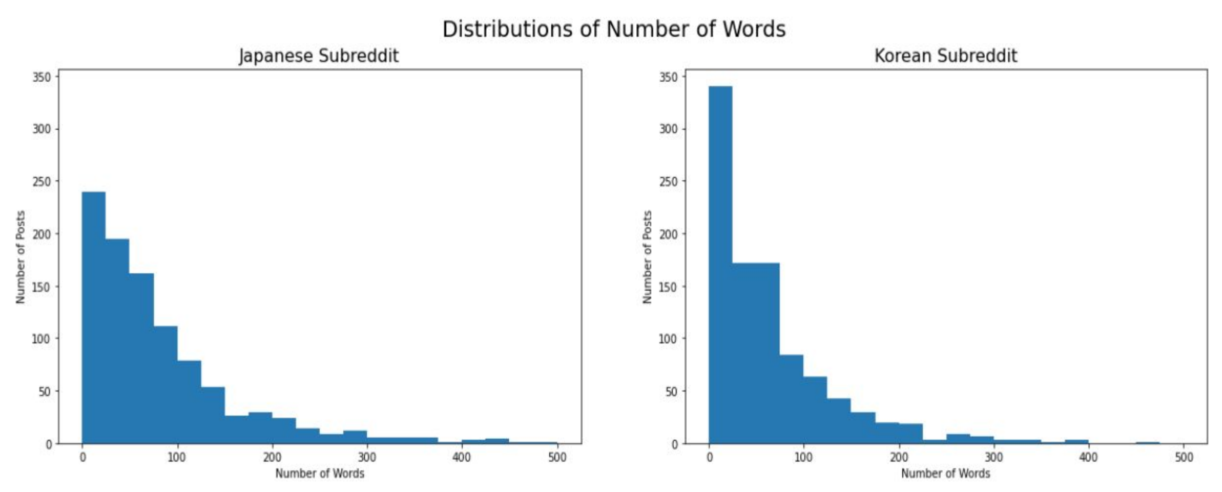
Posts on the Korean subreddit are generally shorter but have more non-English characters compared to Japanese subreddit .



```
==== Japanese Words ====
Mean: 91.53, Maximum: 1921

===== Korean Words =====
Mean: 70.17, Maximum: 2050
```



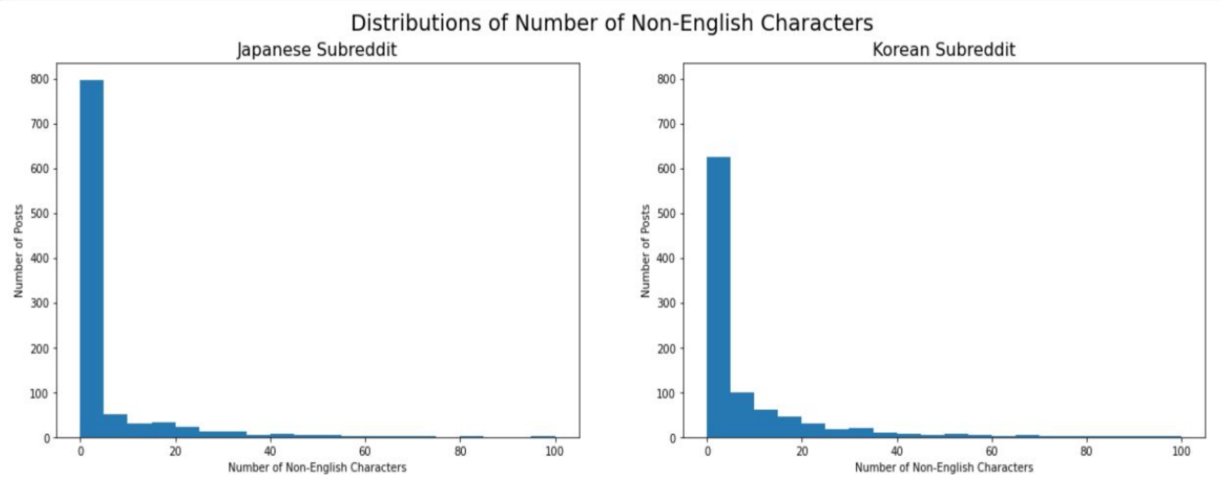
```
== Japanese Non-English ==
Mean: 6.16, Maximum: 392

=== Korean Non-English ===
Mean: 14.89, Maximum: 1079
```



Distributions of Number of Words

Japanese Subreddit

Korean Subreddit



Distributions of Number of Non-English Characters

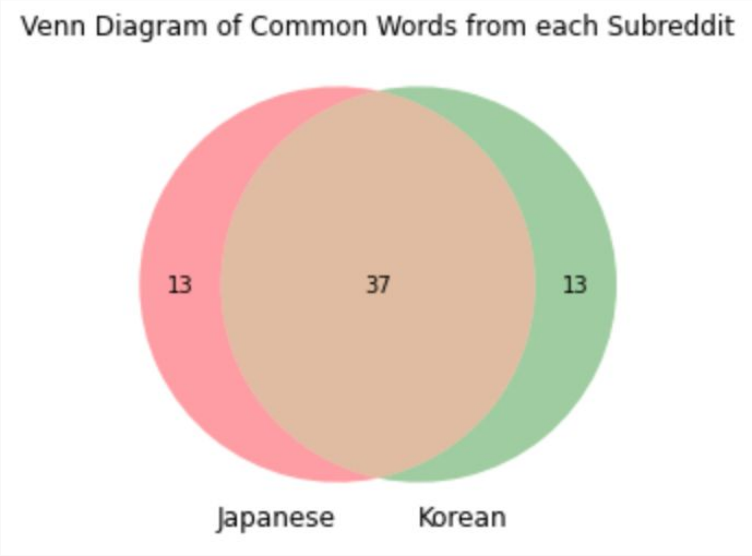
Japanese Subreddit

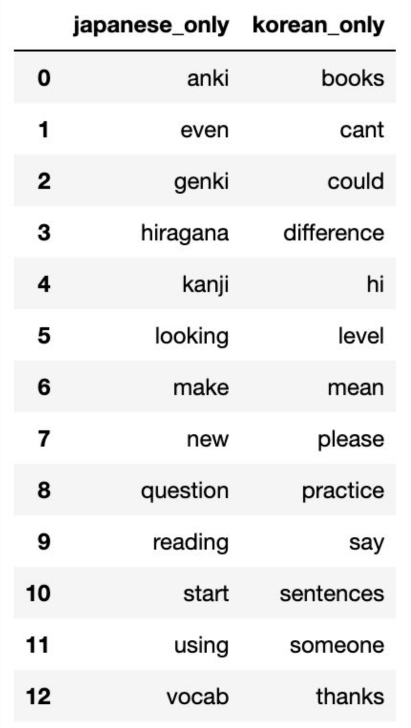
Korean Subreddit

Common Words (1-gram)

After removing English Stopwords, 37 of the top 50 common words from each subreddit are the same.

Chinese characters is 'kanji' in Japanese and 'hanja' in Korean. We see that 'kanji' is a common word in the japanese subreddit but its equivalent 'hanja' is not a common word in the korean subreddit.





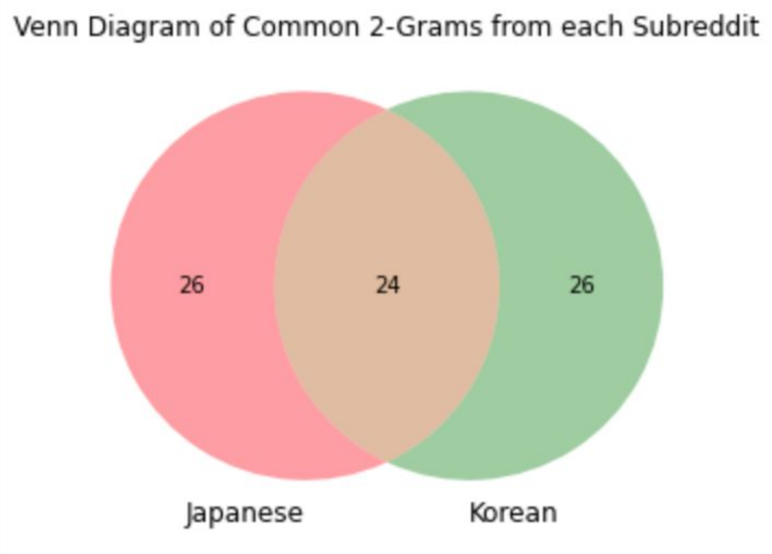
	japanese_only	korean_only
0	anki	books
1	even	cant
2	genki	could
3	hiragana	difference
4	kanji	hi
5	looking	level
6	make	mean
7	new	please
8	question	practice
9	reading	say
10	start	sentences
11	using	someone
12	vocab	thanks

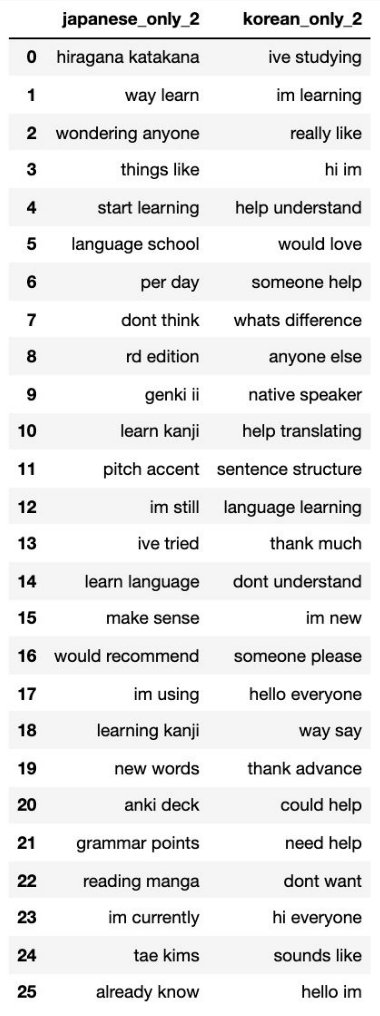
Common Words (2-grams)

The 2-grams are differ more between the subreddits, with less than half in common.

Insights

1. ‘Pitch accent’ in Japanese
   - あめ is rain(雨) or candy(飴)
2. ‘Hiragana’, ‘katakana’ and ‘kanji’ in Japanese, but no ‘hangul’ or ‘hanji’ in Korean
3. ‘Help’, ‘thank’, ‘translate’, ‘hi’ in Korean vs ‘already know’, ‘ive tried’ in Japanese

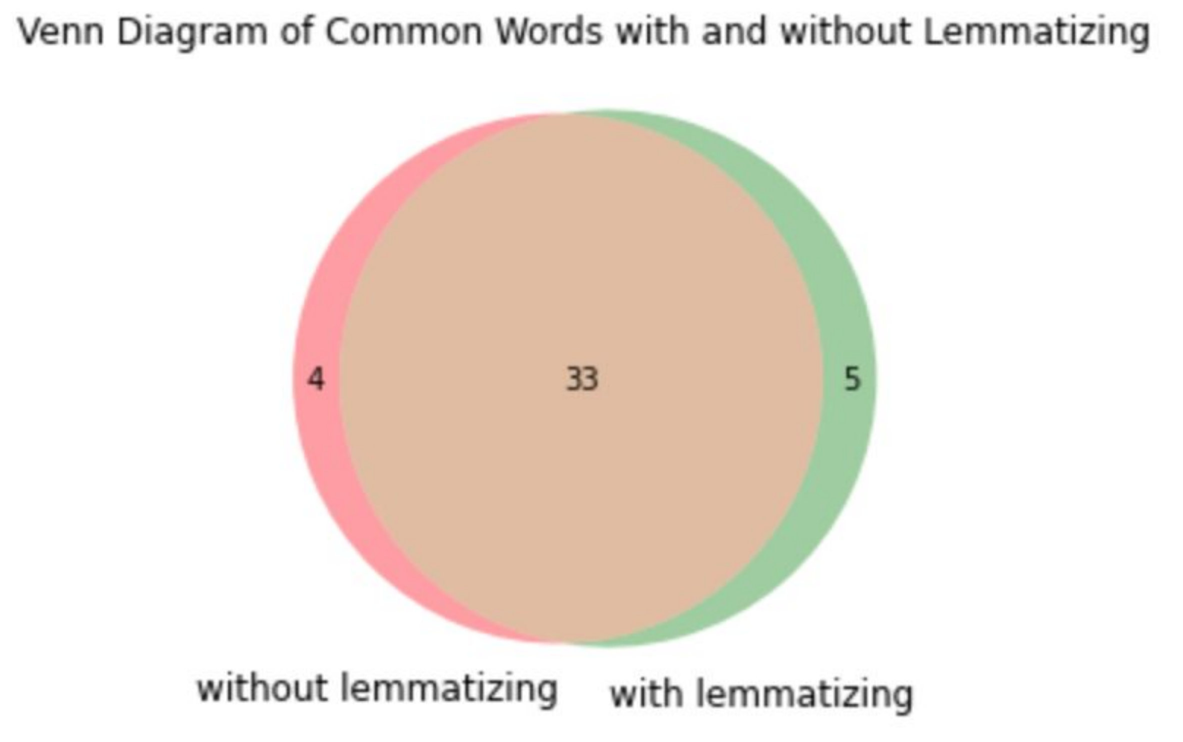




	japanese_only_2	korean_only_2
0	hiragana katakana	ive studying
1	way learn	im learning
2	wondering anyone	really like
3	things like	hi im
4	start learning	help understand
5	language school	would love
6	per day	someone help
7	dont think	whats difference
8	rd edition	anyone else
9	genki ii	native speaker
10	learn kanji	help translating
11	pitch accent	sentence structure
12	im still	language learning
13	ive tried	thank much
14	learn language	dont understand
15	make sense	im new
16	would recommend	someone please
17	im using	hello everyone
18	learning kanji	way say
19	new words	thank advance
20	anki deck	could help
21	grammar points	need help
22	reading manga	dont want
23	im currently	hi everyone
24	tae kims	sounds like
25	already know	hello im

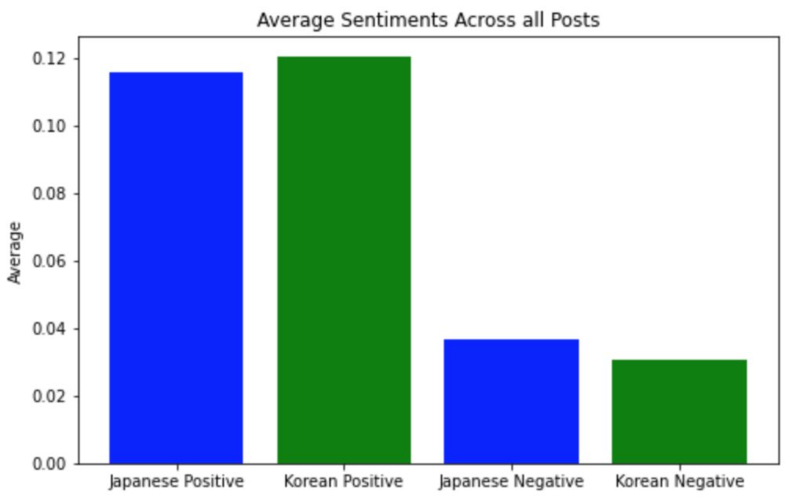
Common Words with Lemmatization

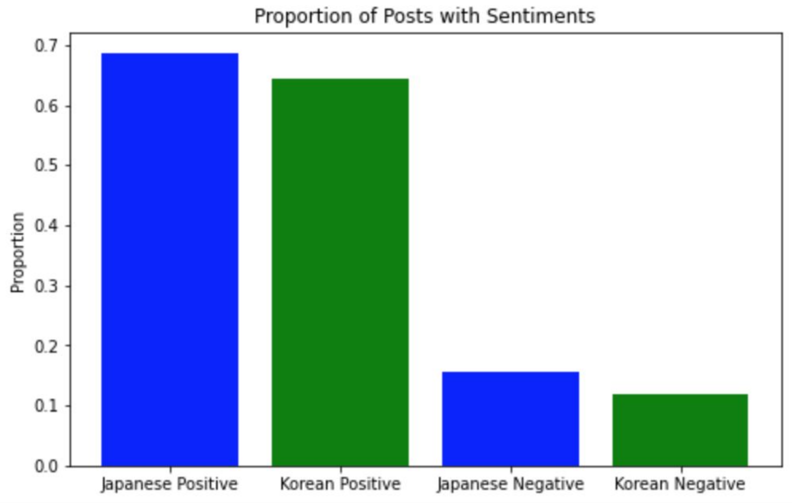
The common words with or without lemmatizing are quite similar, with more than 85% of the words in the intersection. There may not be much utility from lemmatizing the data.



Sentiment Analysis

On average, the Korean posts are slightly more positive and less negative than the Japanese posts, though the Japanese subreddit has a higher proportion of both positive and negative posts (fewer neutral ones).



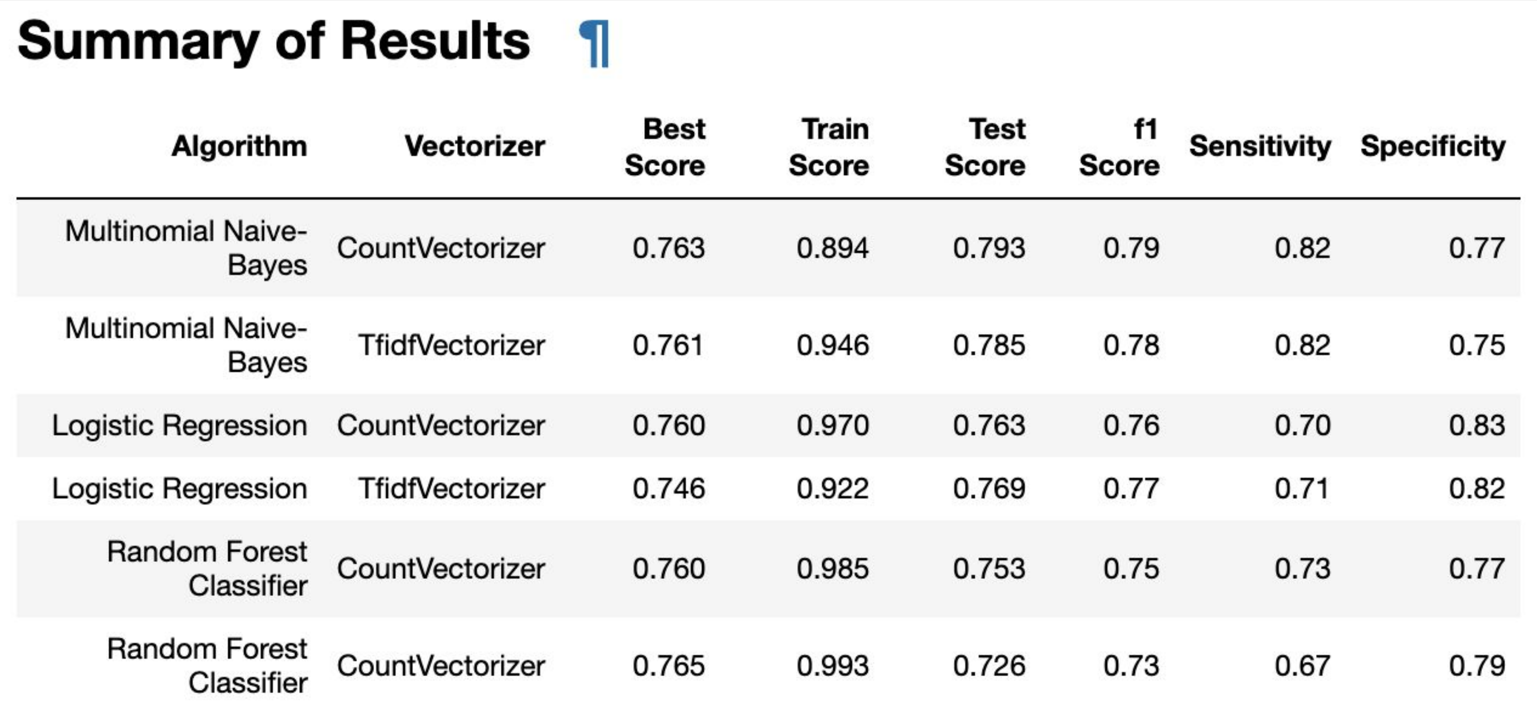


Recommendations to Potential Learners

1. Japanese is more suitable for learners who
    a. already can speak a language that has pitch accent or tones, such as Mandarin, Punjabi or Swedish;
    b. have experience with or are interested in picking up logographic characters (Kanji) and not limit themselves to a phonologic writing system with a fixed set of alphabet.

2. Usage of the Korean subreddit is recommended for beginners, as the posts have fewer negative sentiments (thus may be more encouraging) and the words frequently used indicate a helpful environment, whereas usage of the Japanese subreddit may be more suitable for intermediate learners.

Modelling



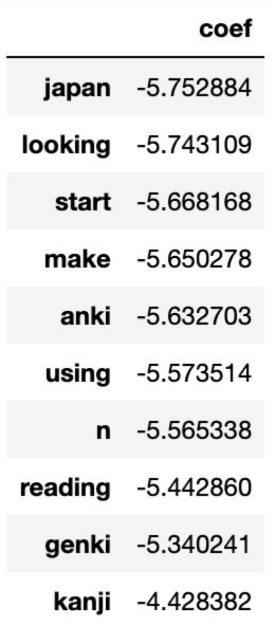
Summary of Results ¶

Algorithm	Vectorizer	Best Score	Train Score	Test Score	f1 Score	Sensitivity	Specificity
Multinomial Naive-Bayes	CountVectorizer	0.763	0.894	0.793	0.79	0.82	0.77
Multinomial Naive-Bayes	TfidfVectorizer	0.761	0.946	0.785	0.78	0.82	0.75
Logistic Regression	CountVectorizer	0.760	0.970	0.763	0.76	0.70	0.83
Logistic Regression	TfidfVectorizer	0.746	0.922	0.769	0.77	0.71	0.82
Random Forest Classifier	CountVectorizer	0.760	0.985	0.753	0.75	0.73	0.77
Random Forest Classifier	CountVectorizer	0.765	0.993	0.726	0.73	0.67	0.79

Important Features
(Multinomial Naive-Bayes)

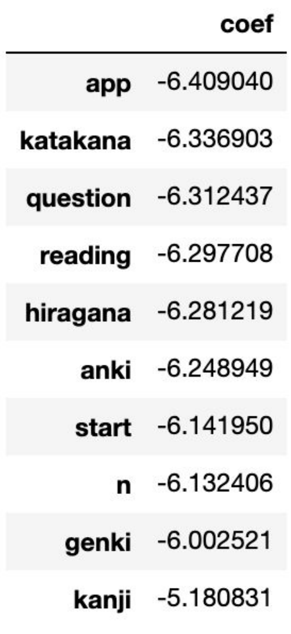
Only the important features for classifying a post as Japanese can be picked out from the words with less negative coefficients

- writing systems 'katakana', 'hiragana' and 'kanji'
- commonly used flashcard app 'anki'
- commonly used textbook 'genki'



	coef
japan	-5.752884
looking	-5.743109
start	-5.668168
make	-5.650278
anki	-5.632703
using	-5.573514
n	-5.565338
reading	-5.442860
genki	-5.340241
kanji	-4.428382

CountVectorizer



	coef
app	-6.409040
katakana	-6.336903
question	-6.312437
reading	-6.297708
hiragana	-6.281219
anki	-6.248949
start	-6.141950
n	-6.132406
genki	-6.002521
kanji	-5.180831

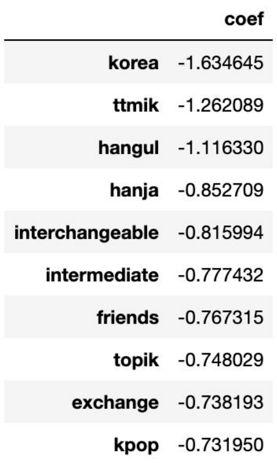
TfidfVectorizer

Important Features (Logistic Regression)

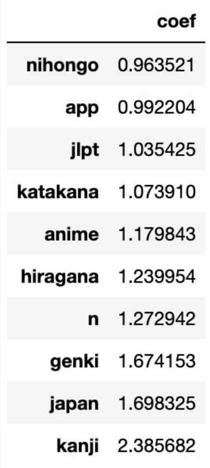
The more negative coefficients are words linked to Korean while the more positive coefficients are words linked to Japanese

- writing systems ('hangul' and 'hanja' for Korean and 'katakana', 'hiragana' and 'kanji' for Japanese)
- popular cultures ('kpop' for Korean and 'anime' for Japanese)
- proficiency tests ('topik' for Korean and 'jlpt' for Japanese)

CountVectorizer

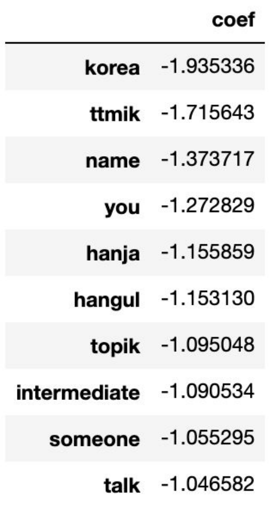


	coef
korea	-1.634645
ttmik	-1.262089
hangul	-1.116330
hanja	-0.852709
interchangeable	-0.815994
intermediate	-0.777432
friends	-0.767315
topik	-0.748029
exchange	-0.738193
kpop	-0.731950

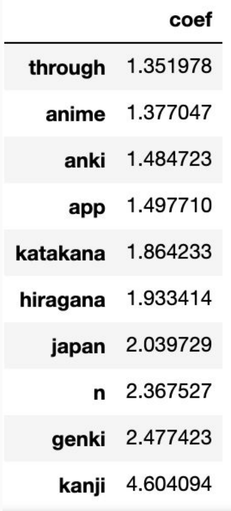


	coef
nihongo	0.963521
app	0.992204
jlpt	1.035425
katakana	1.073910
anime	1.179843
hiragana	1.239954
n	1.272942
genki	1.674153
japan	1.698325
kanji	2.385682

TfidfVectorizer



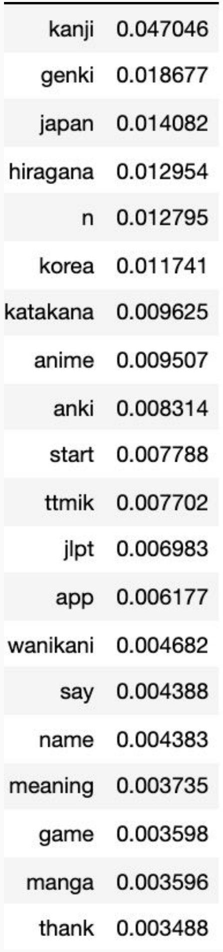
	coef
korea	-1.935336
ttmik	-1.715643
name	-1.373717
you	-1.272829
hanja	-1.155859
hangul	-1.153130
topik	-1.095048
intermediate	-1.090534
someone	-1.055295
talk	-1.046582



	coef
through	1.351978
anime	1.377047
anki	1.484723
app	1.497710
katakana	1.864233
hiragana	1.933414
japan	2.039729
n	2.367527
genki	2.477423
kanji	4.604094

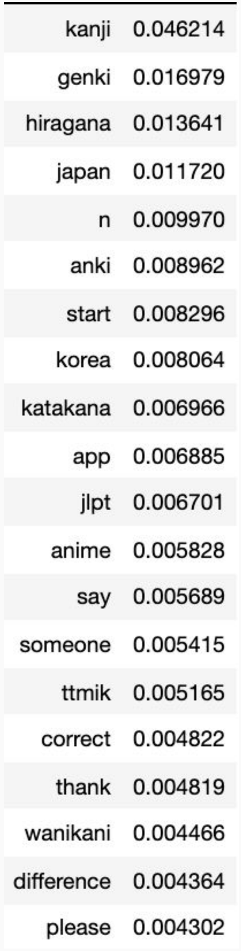
Important Features
(Random Forest Classifier)

the important features are similar to the ones picked out by Logistic Regression, but it cannot be determined which subreddit the word is linked to



kanji	0.047046
genki	0.018677
japan	0.014082
hiragana	0.012954
n	0.012795
korea	0.011741
katakana	0.009625
anime	0.009507
anki	0.008314
start	0.007788
ttmik	0.007702
jlpt	0.006983
app	0.006177
wanikani	0.004682
say	0.004388
name	0.004383
meaning	0.003735
game	0.003598
manga	0.003596
thank	0.003488

CountVectorizer



kanji	0.046214
genki	0.016979
hiragana	0.013641
japan	0.011720
n	0.009970
anki	0.008962
start	0.008296
korea	0.008064
katakana	0.006966
app	0.006885
jlpt	0.006701
anime	0.005828
say	0.005689
someone	0.005415
ttmik	0.005165
correct	0.004822
thank	0.004819
wanikani	0.004466
difference	0.004364
please	0.004302

TfidfVectorizer

Test Models on New Data

To check for future compatibility, one model for each of the 3 algorithms is selected and trained using the original full set of data (from 1 Jan 2021 to 2 Feb 2021) for use on future posts, to ensure that the models are not adversely affected by recent trends.

They are tested on 100 newest posts (retrieved on 20 Jan 2022) from each subreddit, and their performances are evaluated.

Results of Test on New Data

Multinomial Naive-Bayes:	Sensitivity 0.794, Specificity 0.796
Logistic Regression:	Sensitivity 0.691, Specificity 0.786
Random Forest Classifier:	Sensitivity 0.763, Specificity 0.776

There are sufficient differences between the 2 subreddits as 2 of the 3 trained models (Multinomial Naive-Bayes and Random Forest Classifier) accurately predict which subreddit the posts come from for more than 75% of the time.

The Random Forest Classifier is chosen to be the production model as it did not assign very high probabilities to the misclassified posts.

Possible Improvements in Future

1. This project used only 1000 posts from each subreddit. To improve the model, more posts can be scraped from the web so that more data could be used for training.
2. More hyperparameter tuning could be attempted to improve the models if there was more time.
3. The results of the 3 models could be combined using Voting Classifier.