Project 3:

Subreddit Classification using NLP models

__Problem Statement__

There is a large degree of overlap between r/Singapore and r/askSingapore. Questions which should technically be asked in r/askSingapore get asked in r/Singapore as well due to the latter's wider community. Consequently, it can be difficult to correctly predict which subreddit a particular question post belonged to.

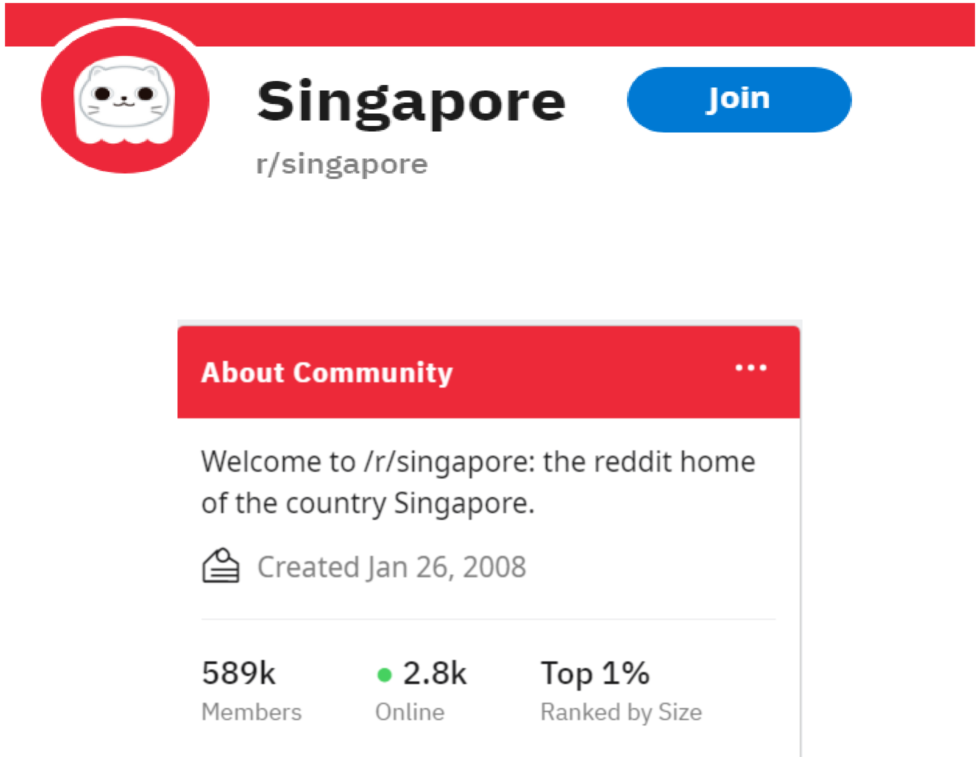
By: Tan Jun Jie

Date: 17 March 2023

Agenda

- Selection of Subreddits
- Sentiment Analysis
- Models Testing
- Best Model Evaluation
- Conclusion & Recommendations

r/Singapore VS r/askSingapore



About Community

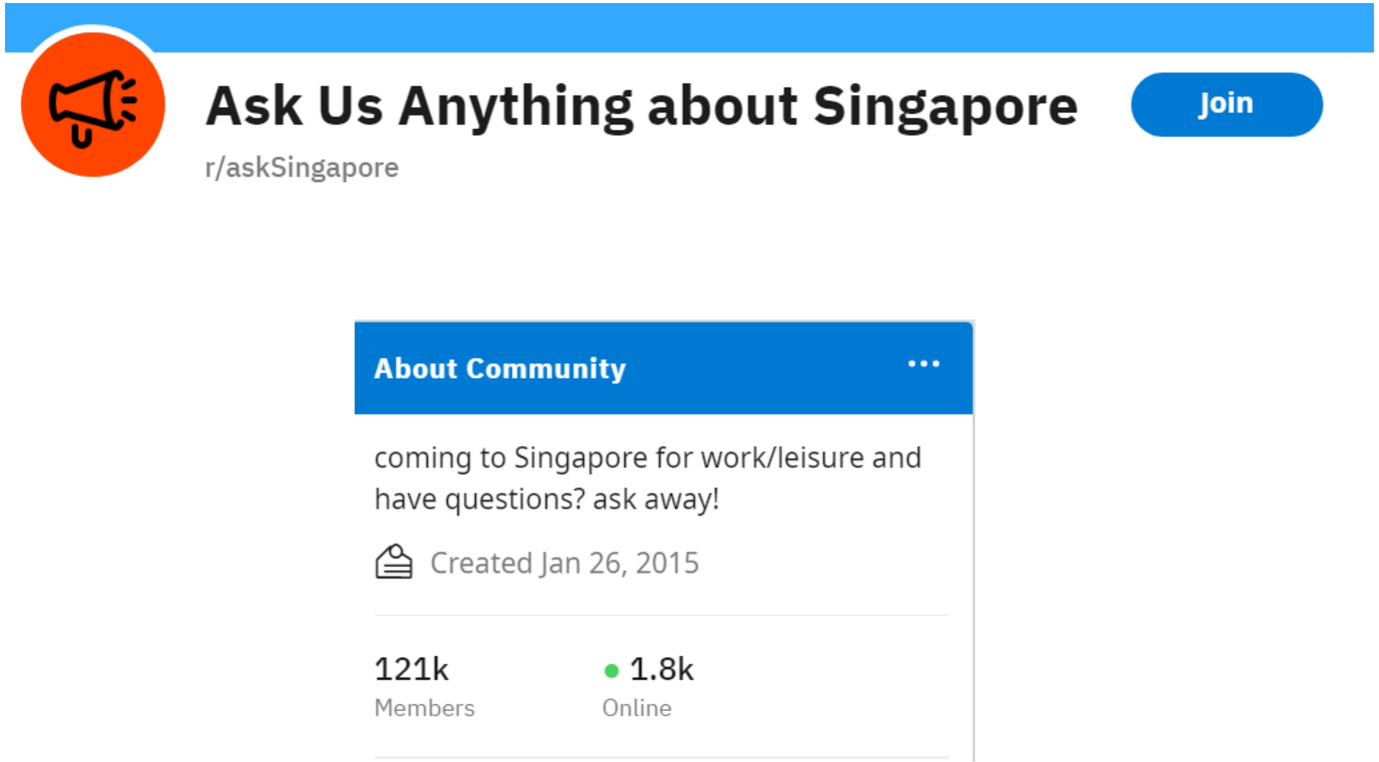
Welcome to /r/singapore: the reddit home of the country Singapore.

Created Jan 26, 2008

589k
Members

● 2.8k
Online

Top 1%
Ranked by Size



About Community

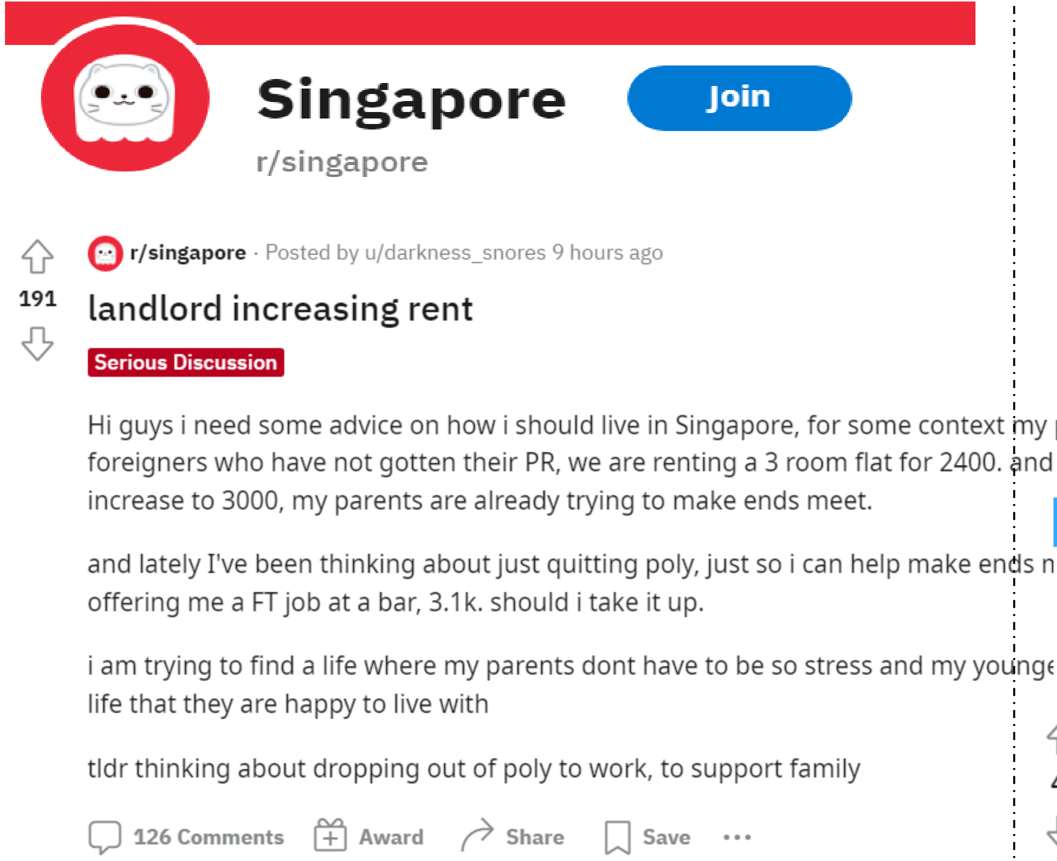
coming to Singapore for work/leisure and have questions? ask away!

Created Jan 26, 2015

121k
Members

● 1.8k
Online

r/Singapore VS r/askSingapore



r/singapore · Posted by u/darkness_snores 9 hours ago

191

landlord increasing rent

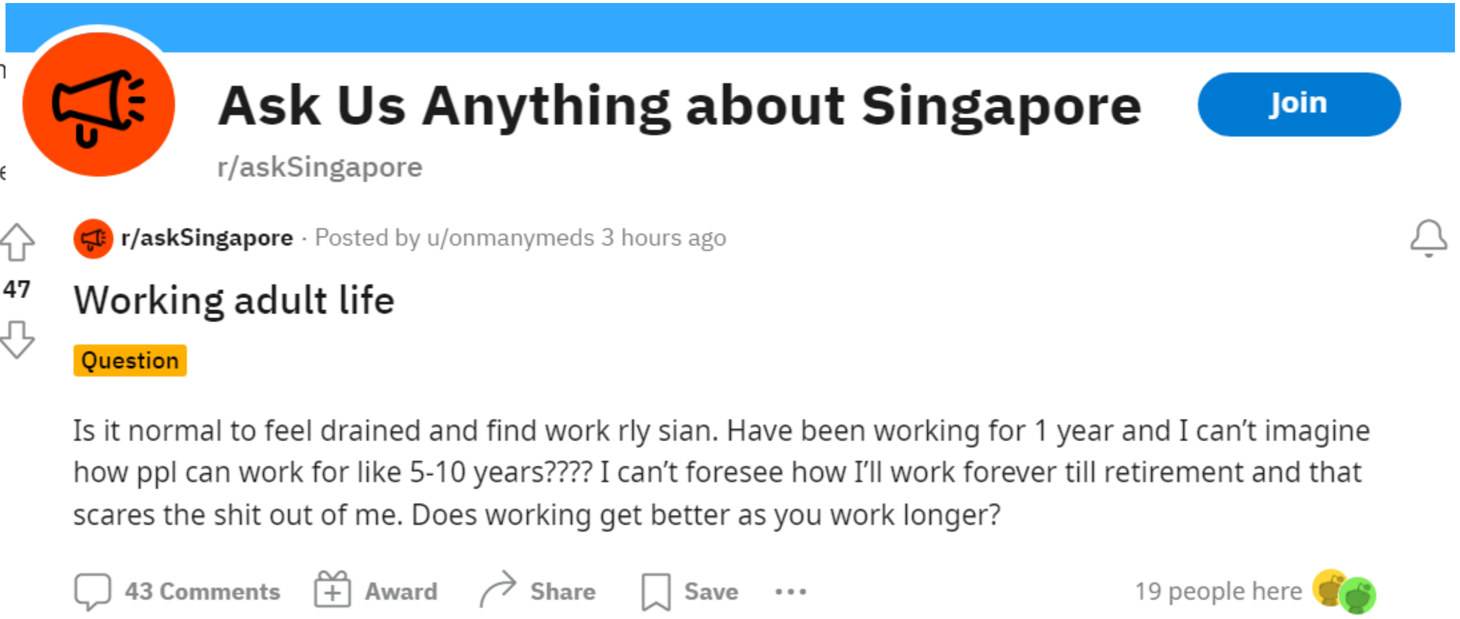
Serious Discussion

Hi guys i need some advice on how i should live in Singapore, for some context my parents are foreigners who have not gotten their PR, we are renting a 3 room flat for 2400. and landlord wants to increase to 3000, my parents are already trying to make ends meet.

and lately I've been thinking about just quitting poly, just so i can help make ends n offering me a FT job at a bar, 3.1k. should i take it up.

i am trying to find a life where my parents dont have to be so stress and my younge life that they are happy to live with

tldr thinking about dropping out of poly to work, to support family



r/askSingapore · Posted by u/onmanymeds 3 hours ago

47

Working adult life

Question

Is it normal to feel drained and find work rly sian. Have been working for 1 year and I can't imagine how ppl can work for like 5-10 years???? I can't foresee how I'll work forever till retirement and that scares the shit out of me. Does working get better as you work longer?

Sentiment Analysis

Sentiment Scores of Titles-only

	Negative	**Neutral**	**Positive**	**Compound**
r/singapore	0.07	0.83	0.10	0.03
r/askSingapore	0.06	0.82	0.12	**0.07**

Sentiment Scores of Titles + Selftext

	Negative	**Neutral**	**Positive**	**Compound**
r/singapore	0.07	0.82	0.11	0.07
r/askSingapore	0.06	0.78	0.16	**0.32**

Sentiments are scored using SentimentIntensityAnalyzer() from nltk library

Models Testing

Models Testing – Accuracy Scores

	NaïveBayes		LogisticRegression		kNearestNeighors	
	train	*test*	*train*	*test*	*train*	*test*
CountVectorizer	0.74	0.70	0.81	0.80	0.75	0.67
TfidfVectorizer	0.76	0.71	0.8	0.76	0.74	0.66

	SupportVectorMachine		RandomForestClassifier	
	train	*test*	*train*	*test*
CountVectorizer	0.77	0.74	0.79	0.76
TfidfVectorizer	0.81	0.77	0.80	0.77

**All models have been optimized using GridSearchCV*
**Balanced classes are used: 51% r/askSingapore and 48% r/Singapore*
**All models have been fitted on a combined dataset of 18,798 reddit posts from the two classes*
**All models have been scored using ‘Accuracy Score’ = Correct Predictions/Total Predictions*

Models Testing – Ranking by Accuracy Scores

Classification Model	Accuracy*	No. of Features
kNN_tvec^	70%	100
kNN_cvec	71%	100
NaiveBayes_cvec	72%	3,000
NaiveBayes_tvec	73%	3,000
SVM_cvec	75%	4,000
RandomForest_cvec	77%	100
LogReg_tvec	78%	3,000
SVM_tvec	79%	4,000
✓ RandomForest_tvec	79%	100
LogReg_cvec	80%	4,000

^cvec=CountVectorizer;   tvec=TfidfVectorizer

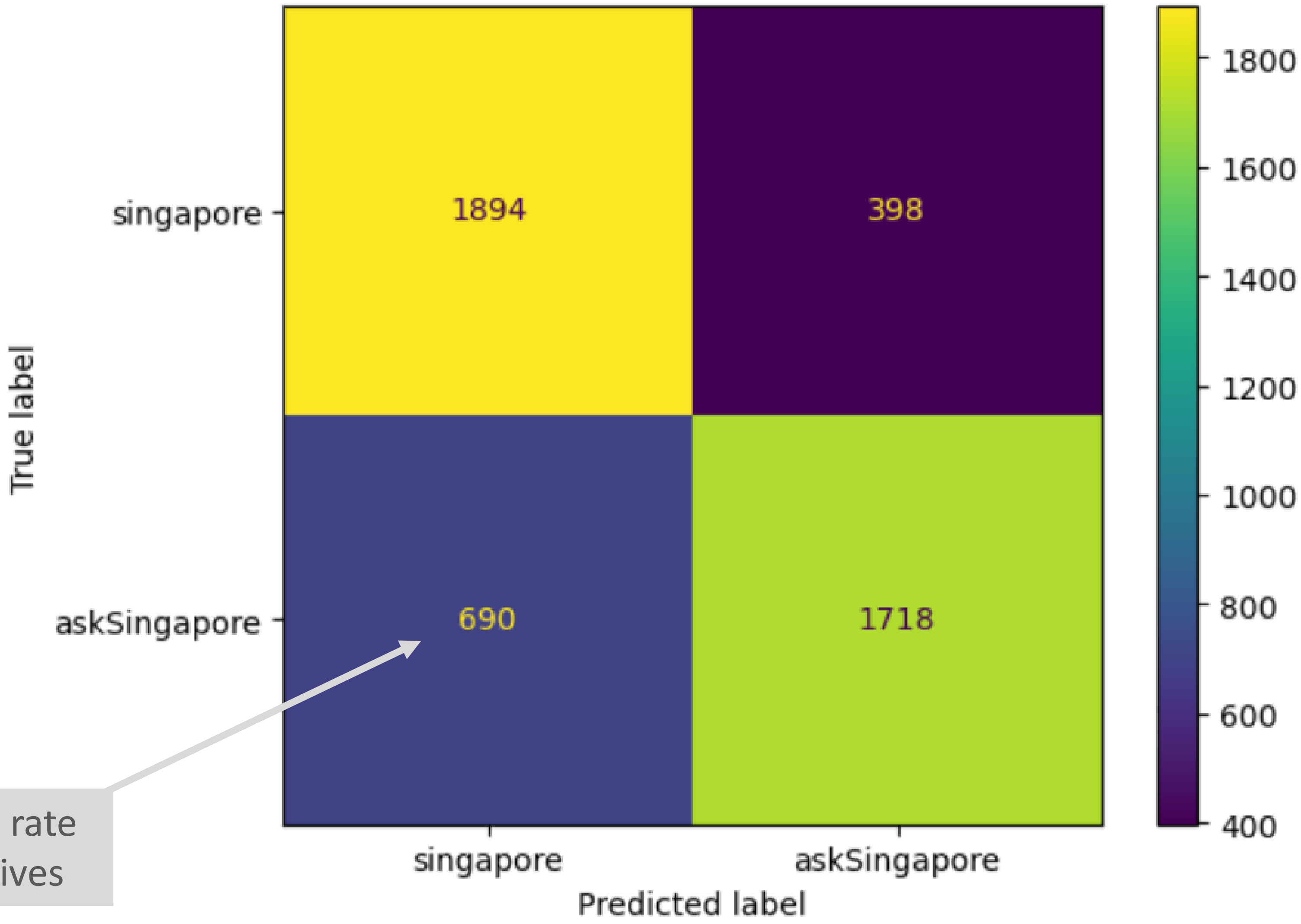
**Accuracy = average of Train+Test Accuracy scores*

Best Model Evaluation

TF-IDF Vectorizer with RandomForestClassifier

TF-IDF_RandomForest Scorecard

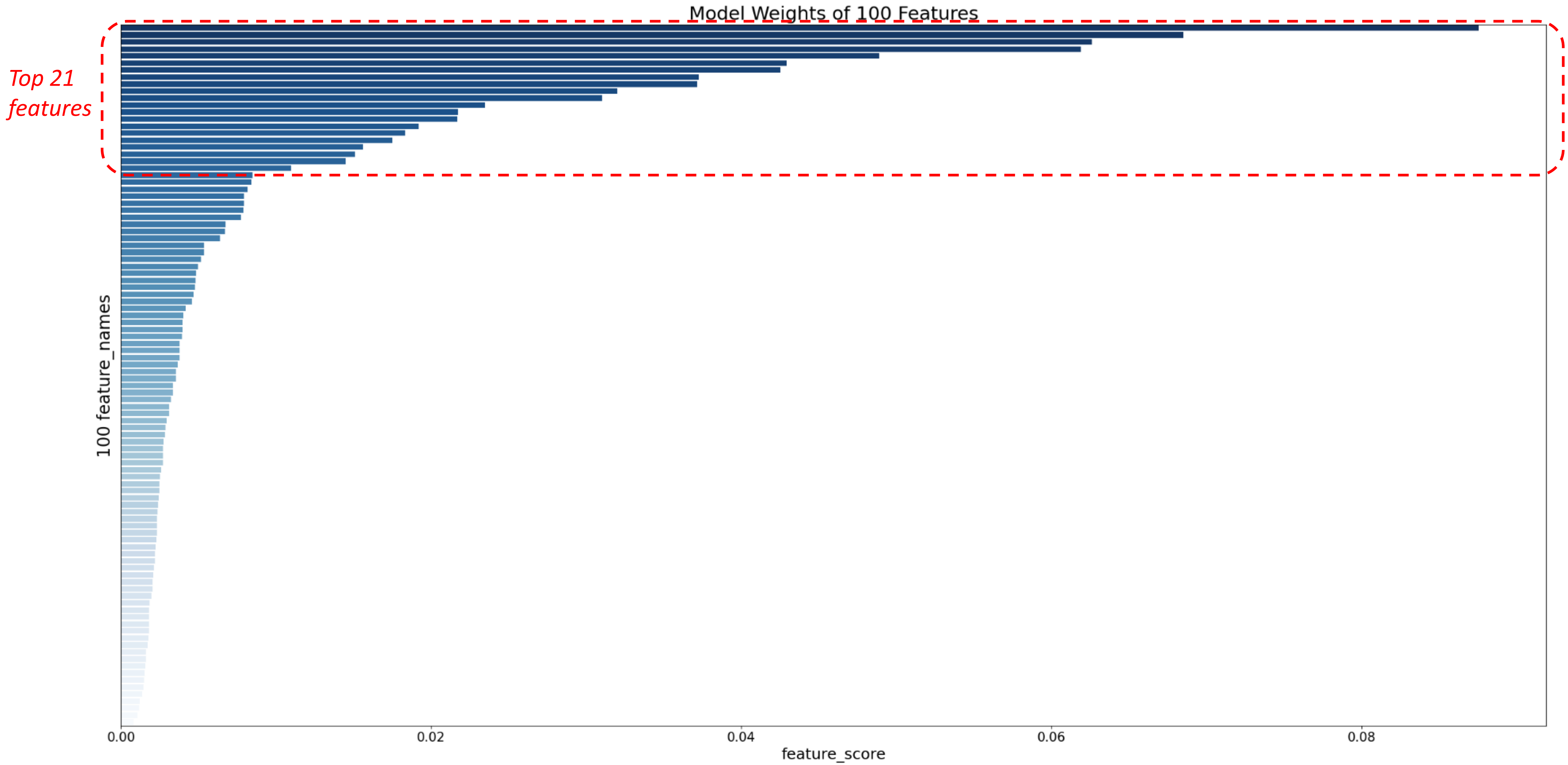
- Accuracy score: 80%
- Precision score: 81%
- Sensitivity score: 71%
- F1 score: 76%
- *Baseline score: 50%*



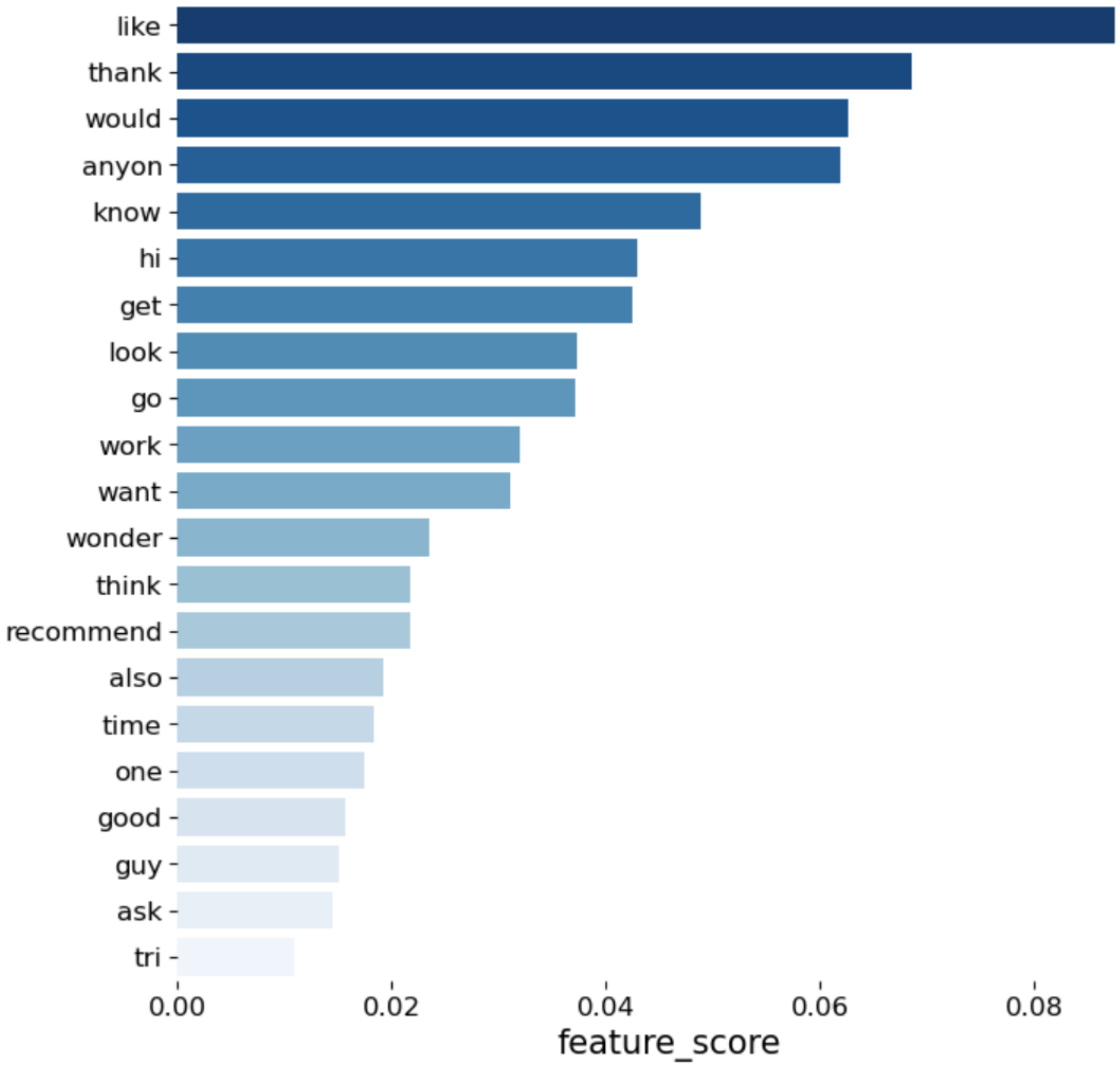
True label \ Predicted label	singapore	askSingapore
singapore	1894	398
askSingapore	690	1718

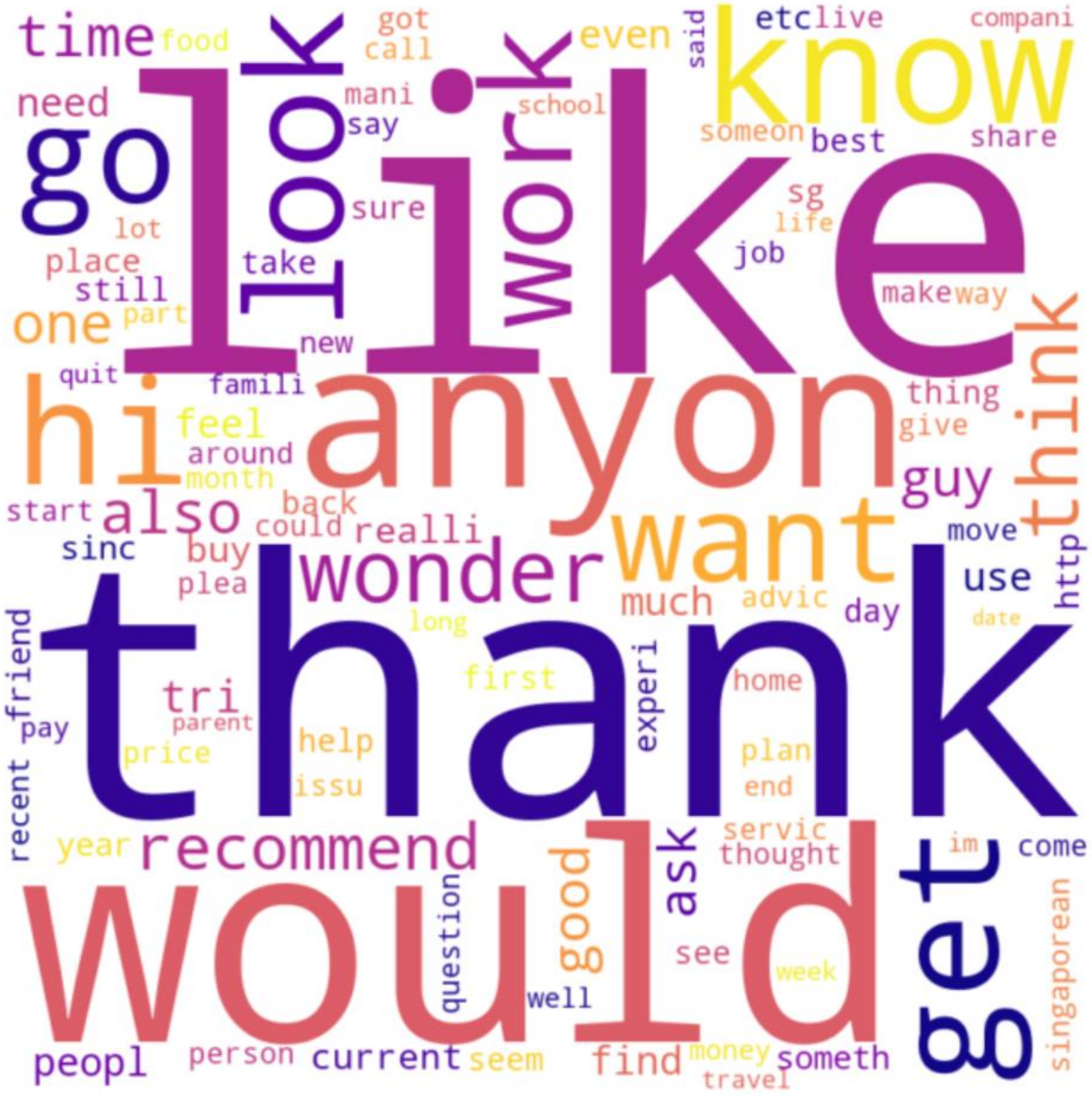
Slightly higher rate of False Negatives

The **top 21** features contributed **73%** to the total weight of 100 features

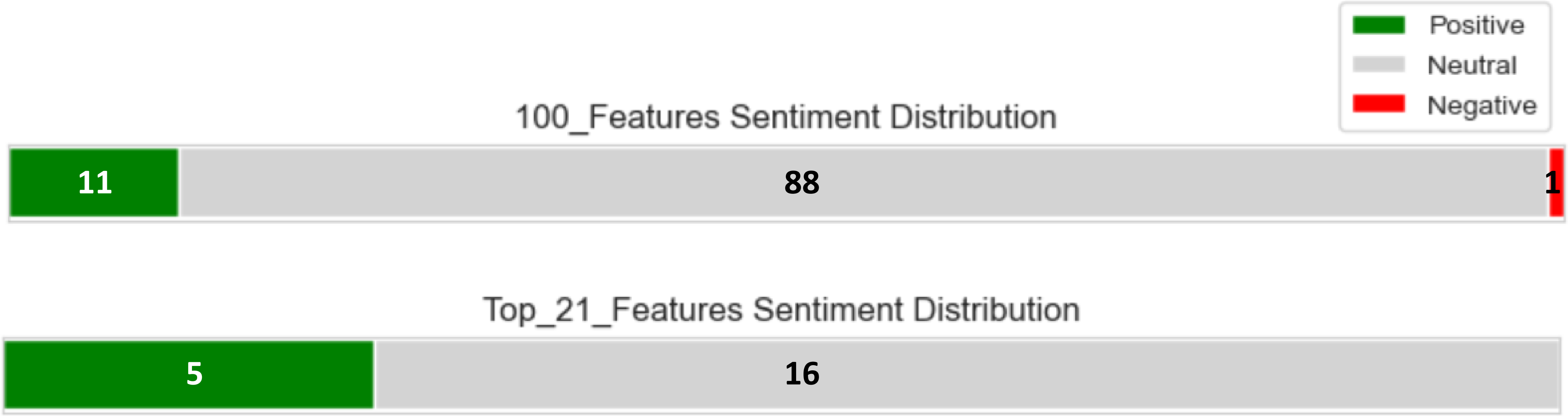


Top 21 Features





A **concentration of positive tokens** in the top 21 features is being overweighted to sharpen model's accuracy



**Note that words are taken out of context and assessed individually using SentimentIntensityAnalyzer() from nltk library*

Conclusions and Recommendations

Use TF-IDF Vectorizer + RandomForestClassifier

- ***Highly Simple***
  - Just 100 features only
- ***Highly Accurate***
  - Scored ~80% both in-sample and out-sample
- ***Highly Relevant***
  - Top 21 features contribute ~60% to the model’s accuracy
- ***Not Highly Perfect!***
  - Slightly higher tendency to predict False Negatives due to overlap