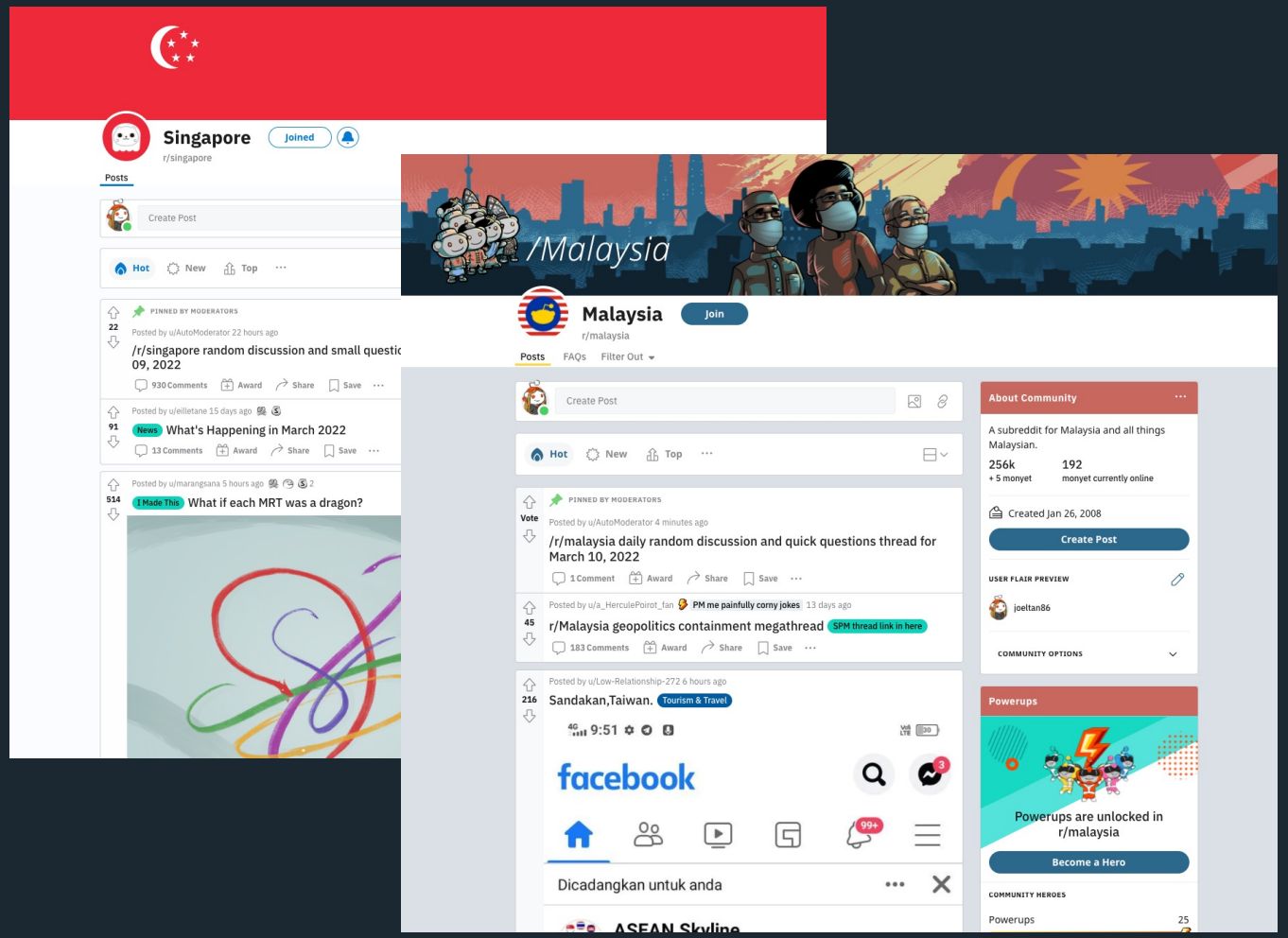
Singapore and Malaysia subreddits: A Comparison

Project 3: Web APIs & NLP

Problem Statement



Project Scenario: A foreign company, is looking to establish its presence into Singapore and Malaysia. As part of a team of data analysts and data scientists , I've been tasked by the marketing department to focus on the platform: Reddit, to explore what are the main concerns and topics of the day of the two countries. And using NLP to train model to identify if a user is from Singapore or Malaysia via the post made.

Stakeholders

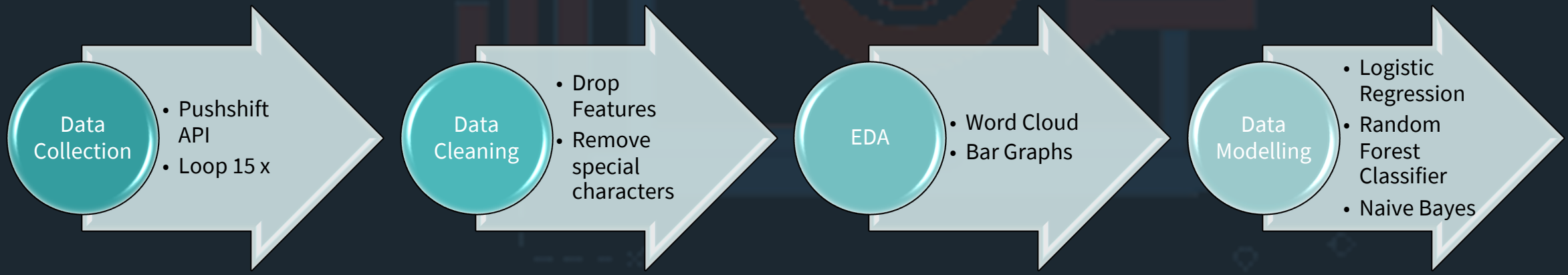
- Primary

The marketing department looking to gather insights and leverage on NLP models to assist in the company's expansion into the region.

- Secondary

Anyone in the company looking for insights into the two subreddits

Work Process



Data Collection

	removed_by_category	num_comments	over_18	selftext	title	media_embed	subreddit
0	still_live	0	False	Why non Slavs still decide to be in the Russia...	Why non Slavic ethnic in the Russian Federatio...	0	malaysia
1	still_live	0	False	How much revenue can hardware shop generate Se...	How much revenue can hardware shop generate	0	malaysia
2	still_live	0	False	NaN	PRN Johor Perdana Menteri Tinjau Keadaan Anggo...	1	malaysia
3	still_live	0	False	NaN	PRN Johor Program Gotong Royong Bersama Pendud...	1	malaysia
4	automod_filtered	0	False		Boleh pcaya ka puasa dan aidilfitri tak PKP	0	malaysia

- **Total post collected: 2,999**

1,500 from r/Singapore

1,499 from r/Malaysia

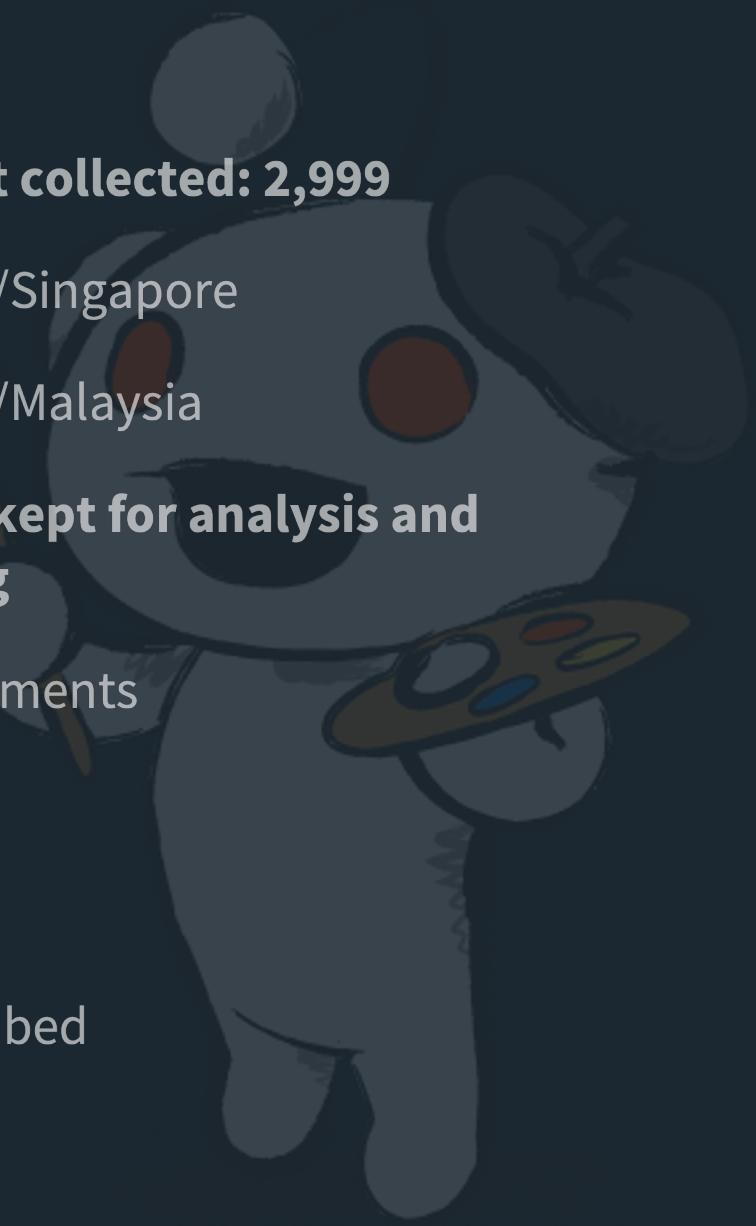
- **Features kept for analysis and modelling**

✓num_comments

✓Over_18

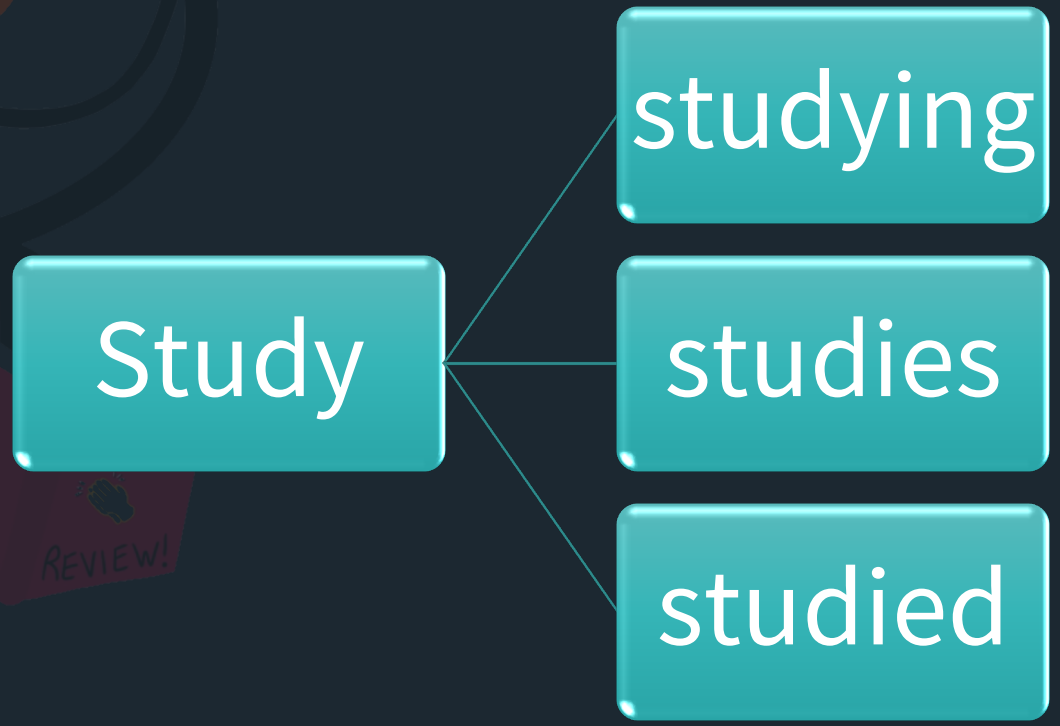
✓Title

✓Media_embed



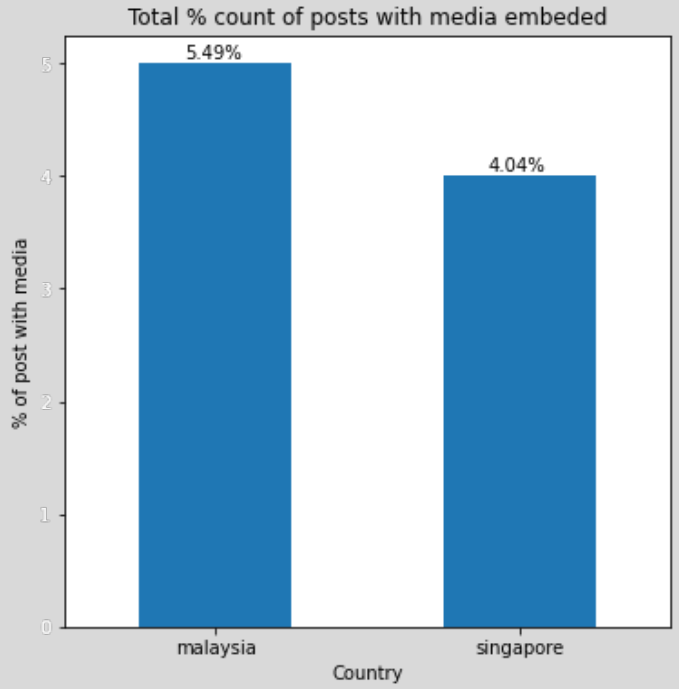
Data Cleaning

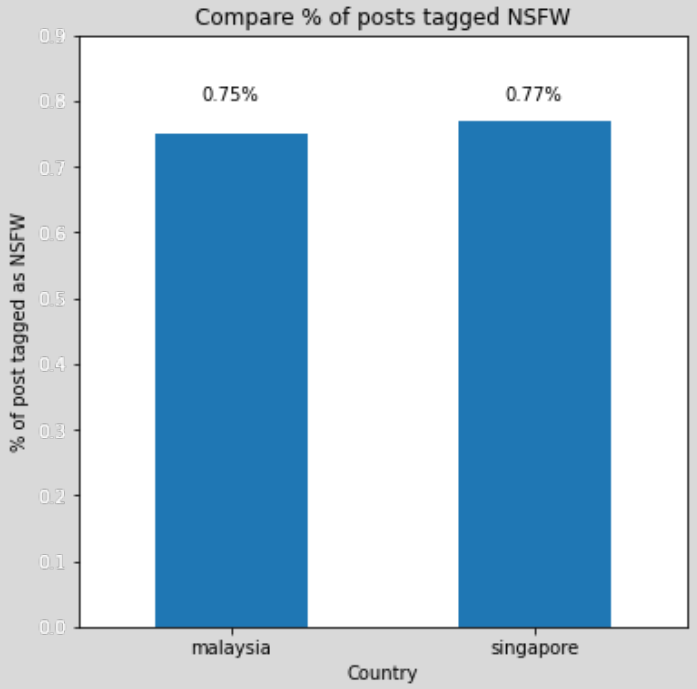
- Drop mismatch features from both subreddits
- Drop unused features
- Drop special characters and digits
- Drop text rows with single alphabets
- Drop words base on current global events
- Lemmatize words

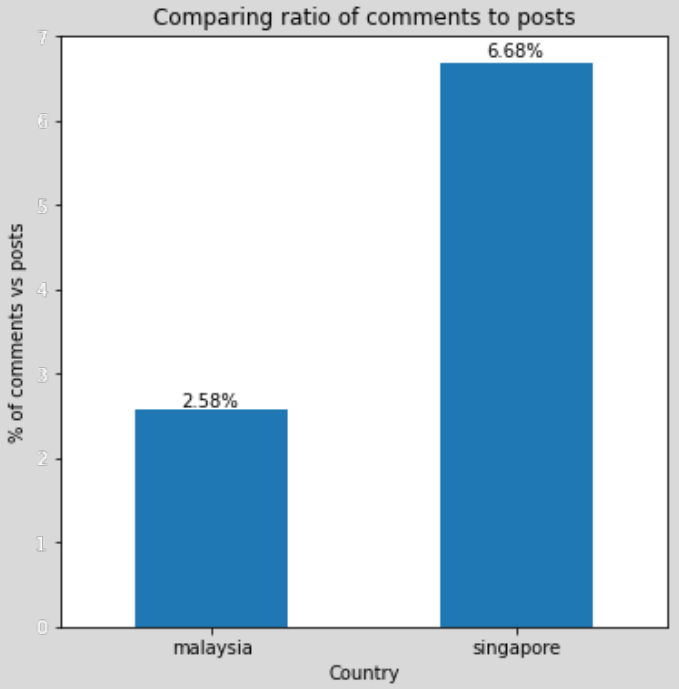


EDA

Comparing both subreddits



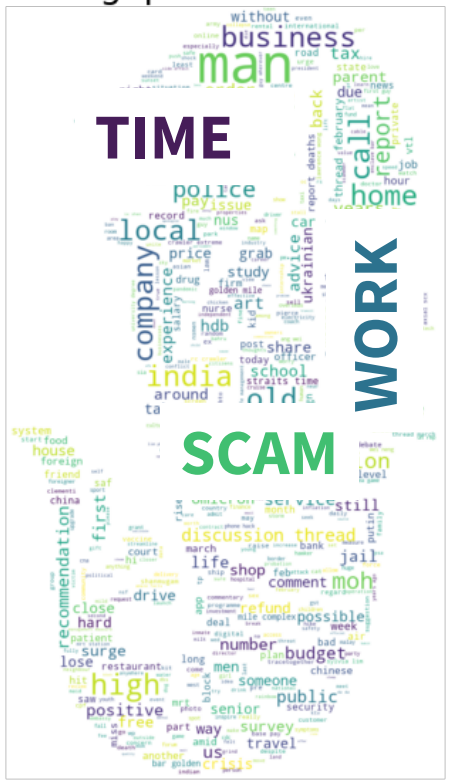




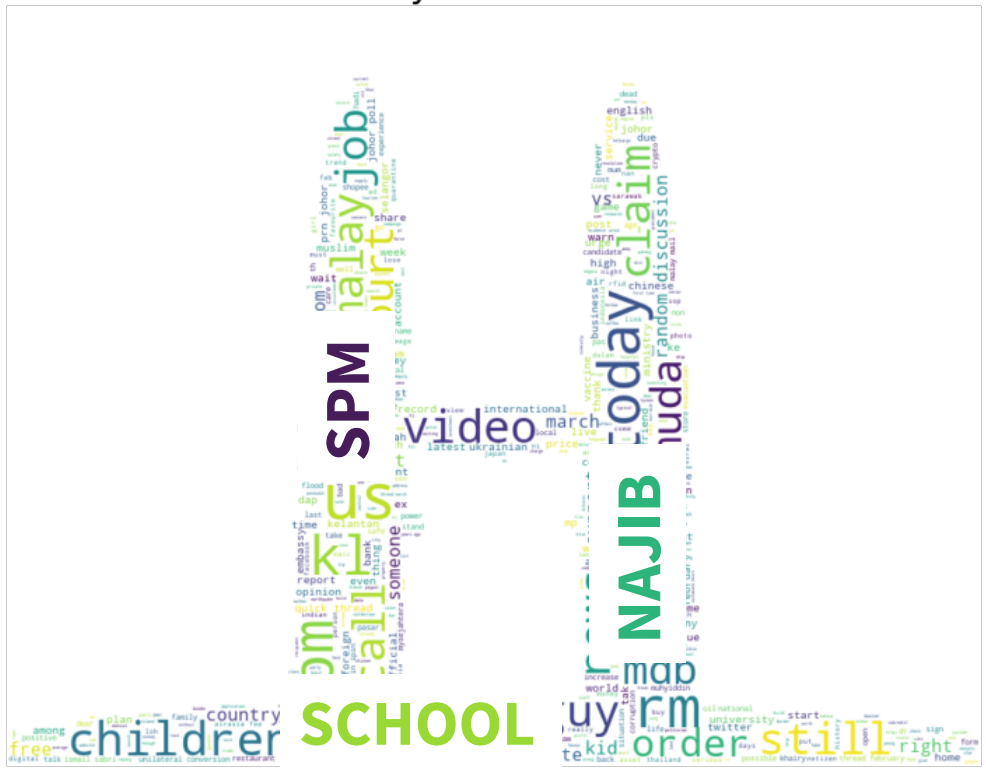
EDA

Word Cloud of both subreddits and the top 10 word count

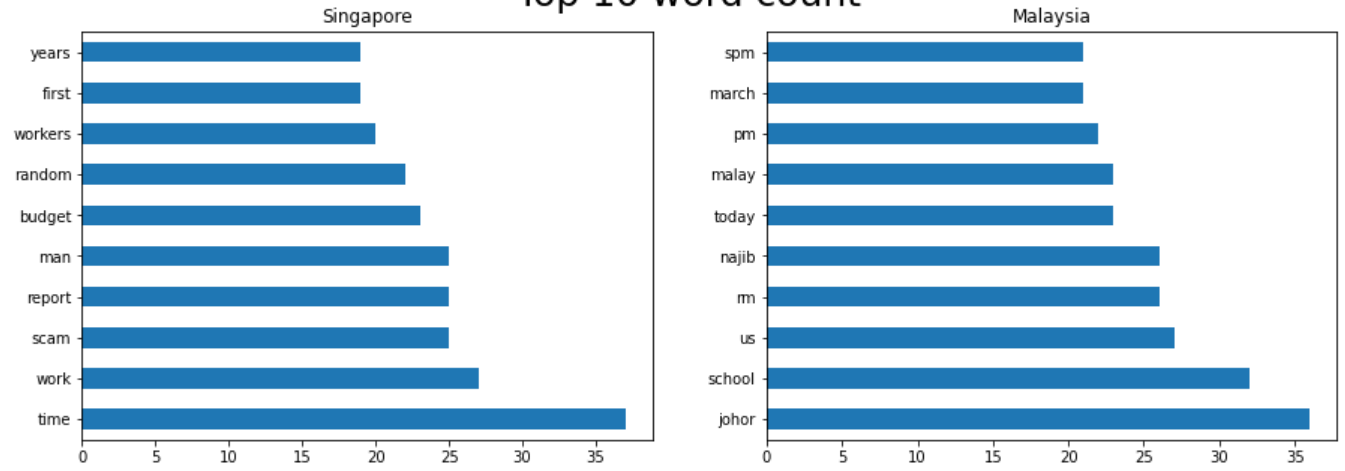
Singapore Subreddit



Malaysia Subreddit



Top 10 word count

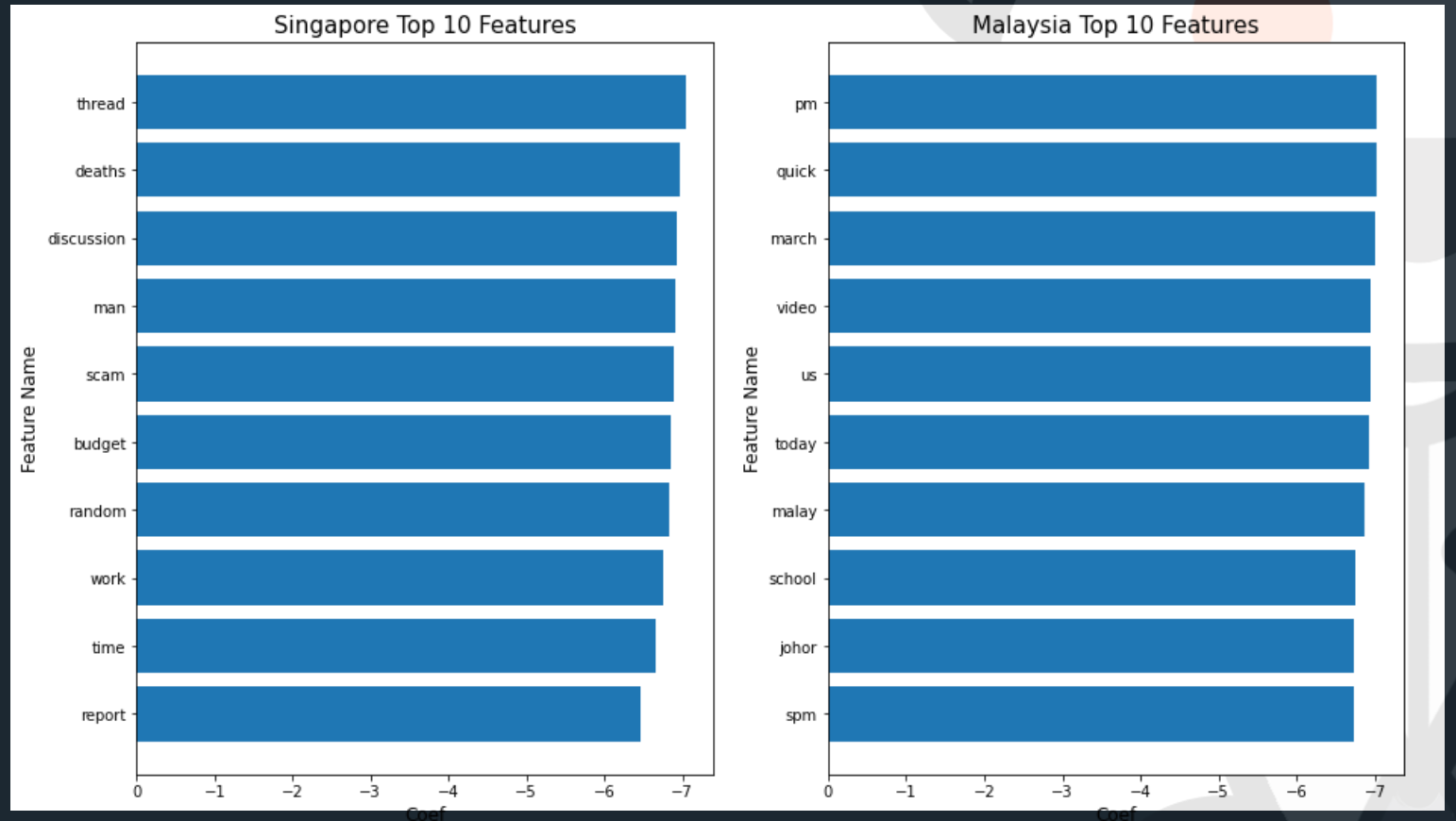


Data Modelling

Measures	Logistic Regression	Random Forest	Multinomial	Complement	Bernoulli
train_score	0.973822	0.997818	0.952007	0.928447	0.931937
test_score	0.686411	0.668990	0.712544	0.707317	0.688153
specificity	0.685246	0.718033	0.681967	0.727869	0.596721
recall	0.687732	0.613383	0.747212	0.684015	0.791822
balanced_accuracy	0.686489	0.665708	0.714590	0.705942	0.694271
best_score	nan	0.661421	0.677983	0.682783	0.681051

Model Evaluation

Mulitnomial classifier top features



Sample post the multinomial classifier did not predict correctly

	title	actual	predicted
0	online museum scientist malaya trip	malaysia	singapore
7	double life wive duke nus researcher spy us	singapore	malaysia
9	kl bangkok hsr idea better economic sense expert	malaysia	singapore
10	kdrama vs drama	malaysia	singapore
14	house waterfall basement walk wine cellar	singapore	malaysia
16	putin pm lee pres yacob	singapore	malaysia
18	upcoming sale scoot	singapore	malaysia
19	infantcare teacher admit hit old boy slam face	singapore	malaysia
21	true	malaysia	singapore
23	language origin mrt station name	singapore	malaysia
26	man move injure juvenile wild boar sit middle ...	singapore	malaysia
27	pfizer pills weeks khairy	malaysia	singapore
29	lift rostered routine sectors financial admini...	singapore	malaysia
30	price increase fav prata stall due chain disru...	singapore	malaysia
33	asos miss order	singapore	malaysia
38	airlines post first quarterly profit since sta...	singapore	malaysia
39	shoot	singapore	malaysia
40	workers party crisis adversity quotient back p...	singapore	malaysia
42	mahkamah tolak permohonan najib irwan supaya d..	malaysia	singapore
43	lah study hard straits time	singapore	malaysia

Conclusion and Final Thoughts

- Once words that were linked to current global events were removed, there was a distinct trend in both subreddits. For Singapore recurring words of note were: scam, work and time. For Malaysia it was: Najib, School, SPM and children.
- While clear distinctions could be made during EDA, the classification models did not fare well managing only a max accuracy of 71%.
- Additional data collection and Feature Engineering would help in improving the effectiveness of the model (translating Malay to English, working with a linguist, etc.)