Classifying Humour

British Problems vs. British Success

Ben Poh, DSI24

Structure

- Context and Problem Statement
- Data Cleaning/EDA
- Modelling (Logistic Regression/ Naïve Bayes / Random Forest)
- Analysis
  - Feature Analysis
  - Misclassification

Context and Problem Statement

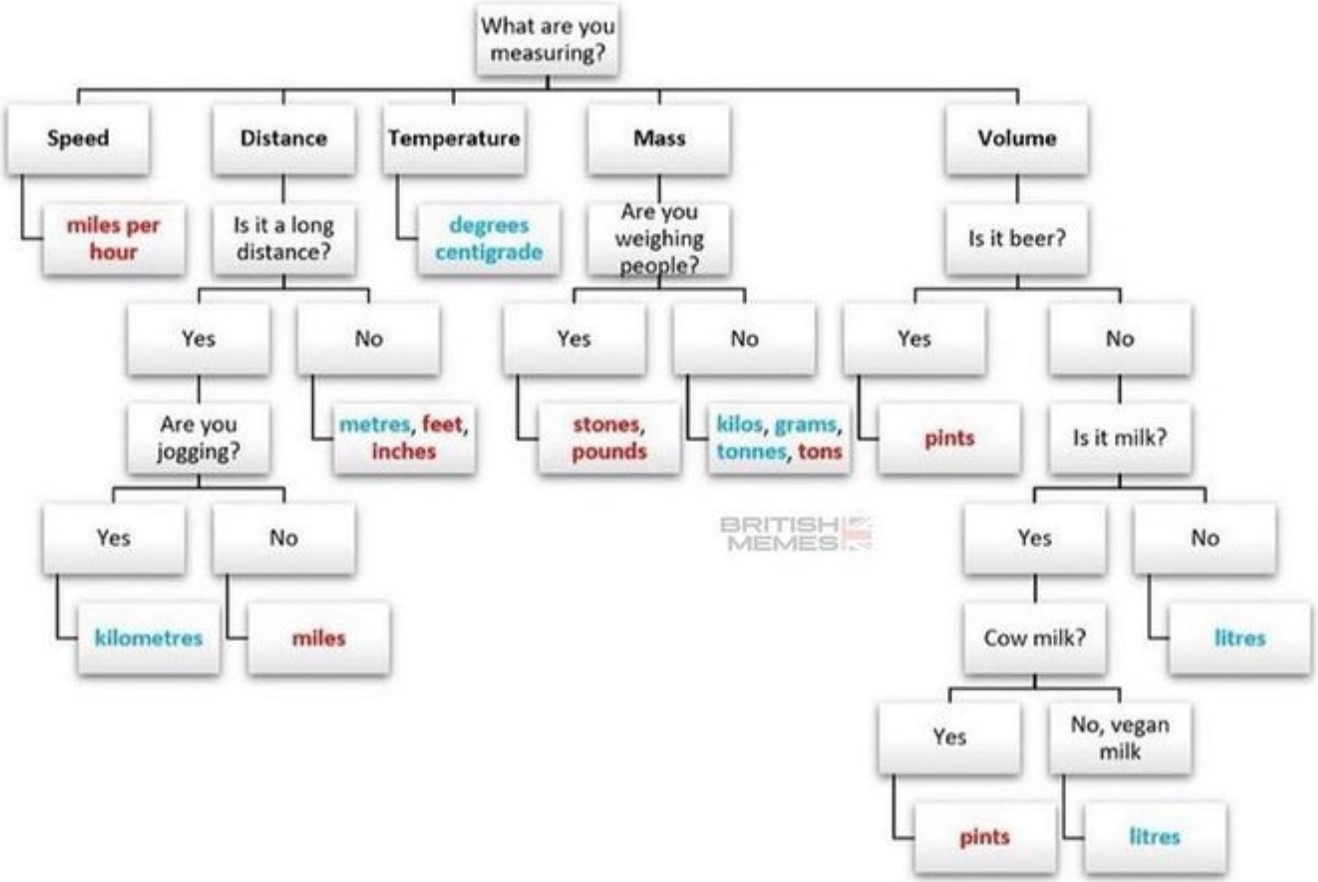
Context

- You work for SGAG – your manager to explore new engagement channels via Reddit
- You want to start sub-reddits on SingaporeanProblems and SingaporeanSuccess – aim to laugh at daily life
- British Problems and British Success sub-reddits – similar on surface, but linguistic complexities of British humour could make them tricky to distinguish (irony, sarcasm, self-deprecating humour)
- *'I am excited to be taking a nice stroll in the rain again...'*

Problem Statement

- You are thinking of setting up both sub-reddits but worry that they might cannabalise viewership if posts are similar.
- Are there differentiating features (words) if humour can actually be classified

How to measure as a Brit.

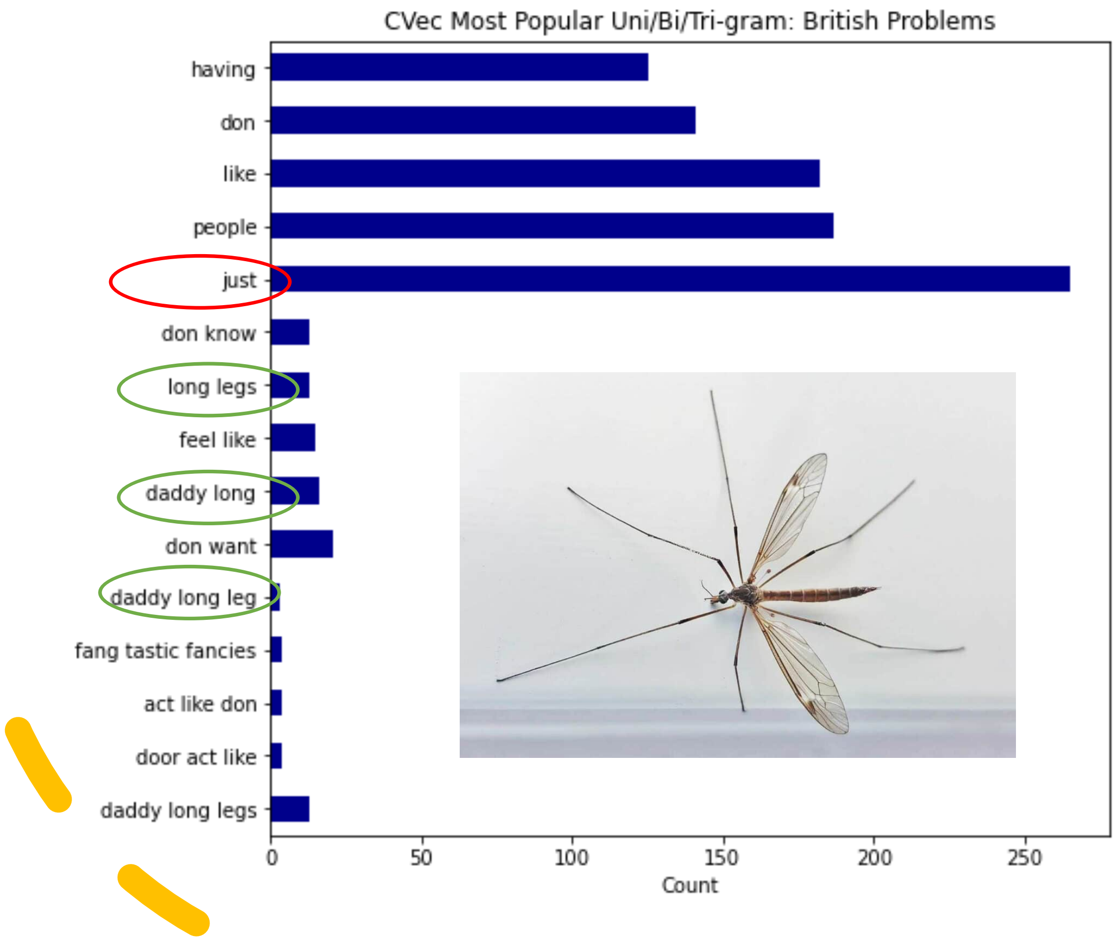


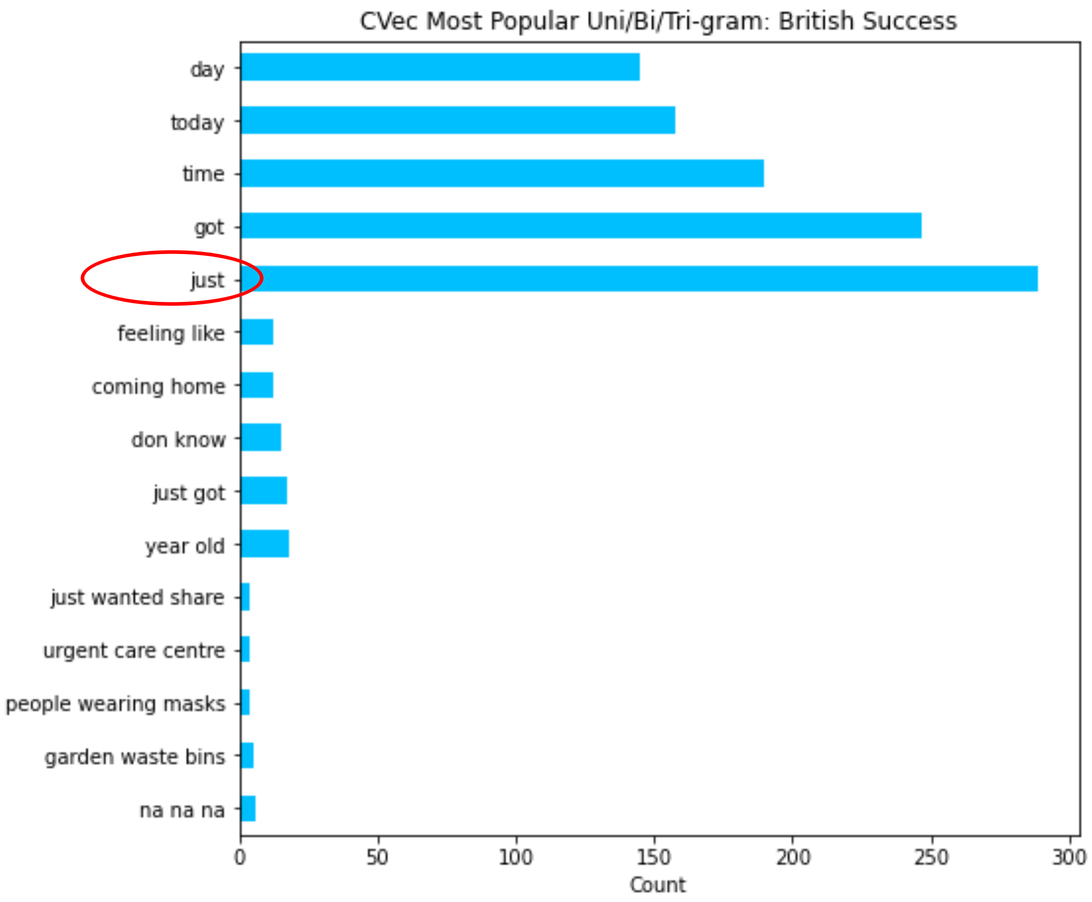
- What are you measuring?
  - Speed
    - miles per hour
  - Distance
    - Is it a long distance?
      - Yes
        - Are you jogging?
          - Yes
            - kilometres
          - No
            - miles
      - No
        - metres, feet, inches
  - Temperature
    - degrees centigrade
  - Mass
    - Are you weighing people?
      - Yes
        - stones, pounds
      - No
        - kilos, grams, tonnes, tons
  - Volume
    - Is it beer?
      - Yes
        - pints
      - No
        - Is it milk?
          - Yes
            - Cow milk?
              - Yes
                - pints
              - No, vegan milk
                - litres
          - No
            - litres

Data Cleaning and EDA

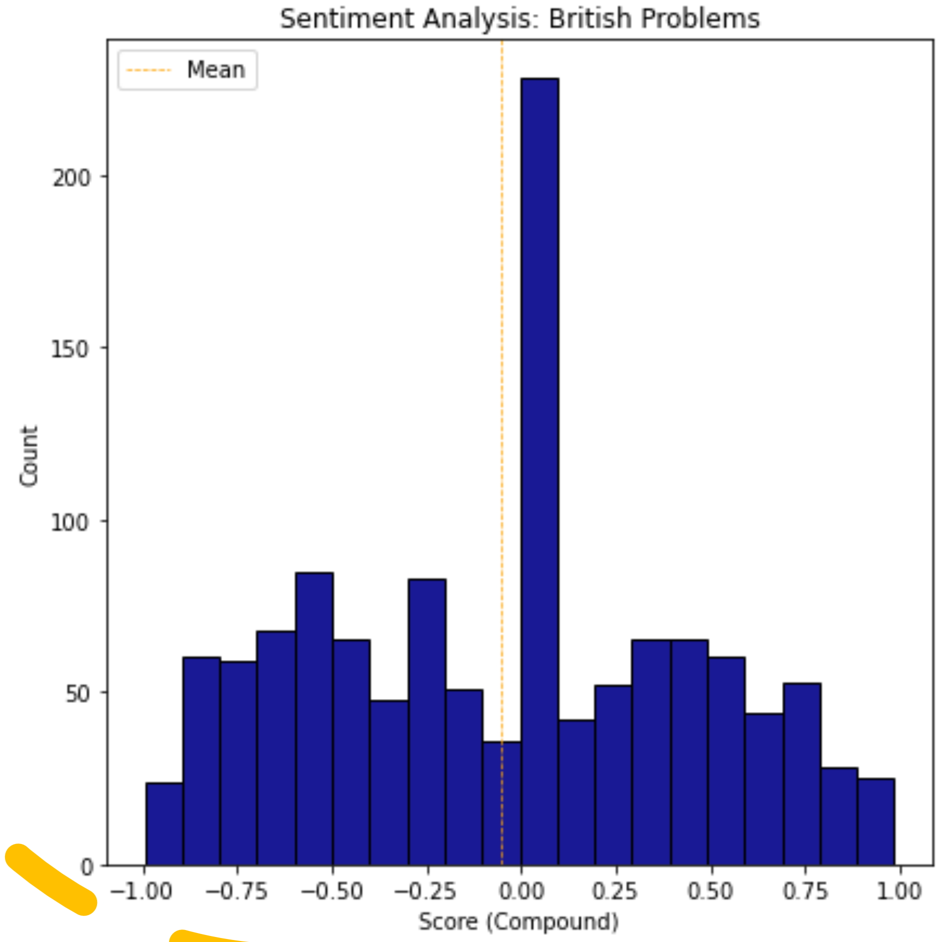
- Missing ‘selftext’ field – a lot of one-liners in ‘title’ for comedic effect
- Remove dupes / posts deleted by Reddit
- Remove URLs, \n \r with Regex
- Popular N-grams
- Sentiment Analysis

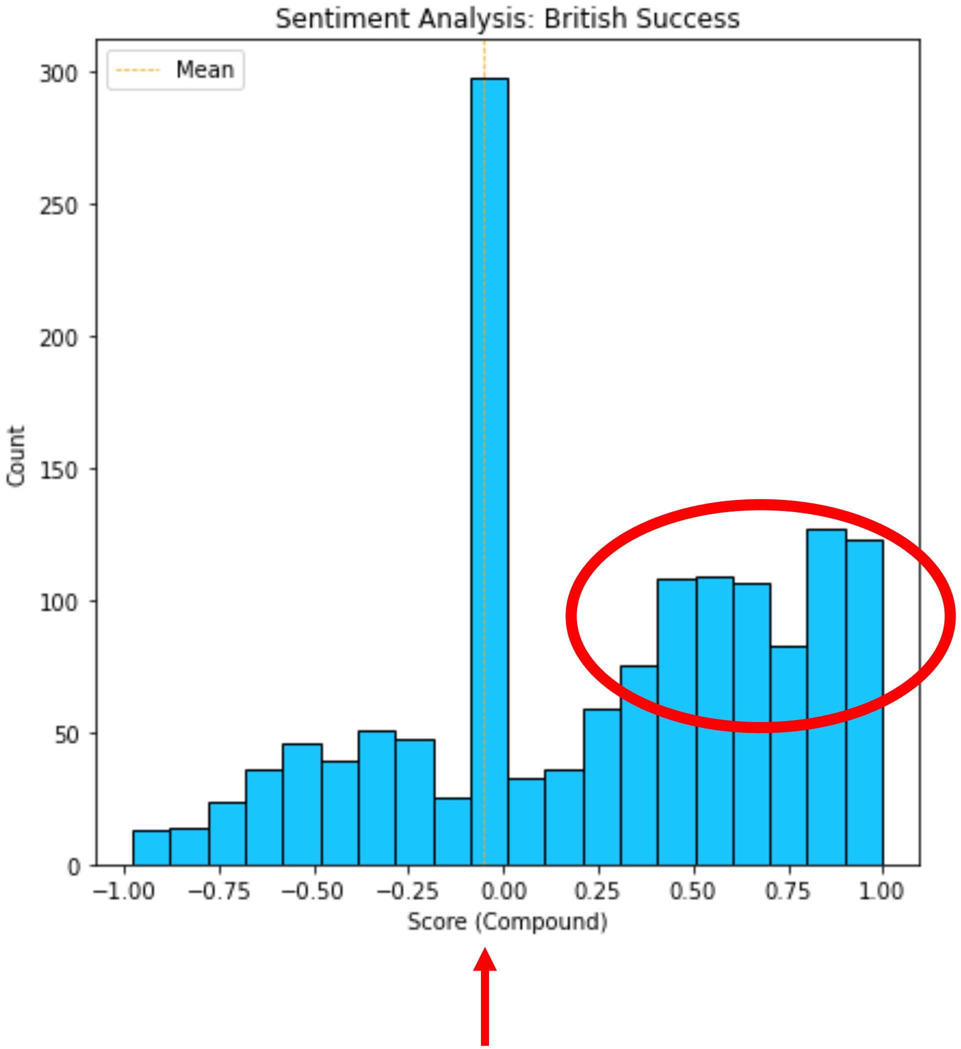
Popular Uni / Bi / Tri-grams



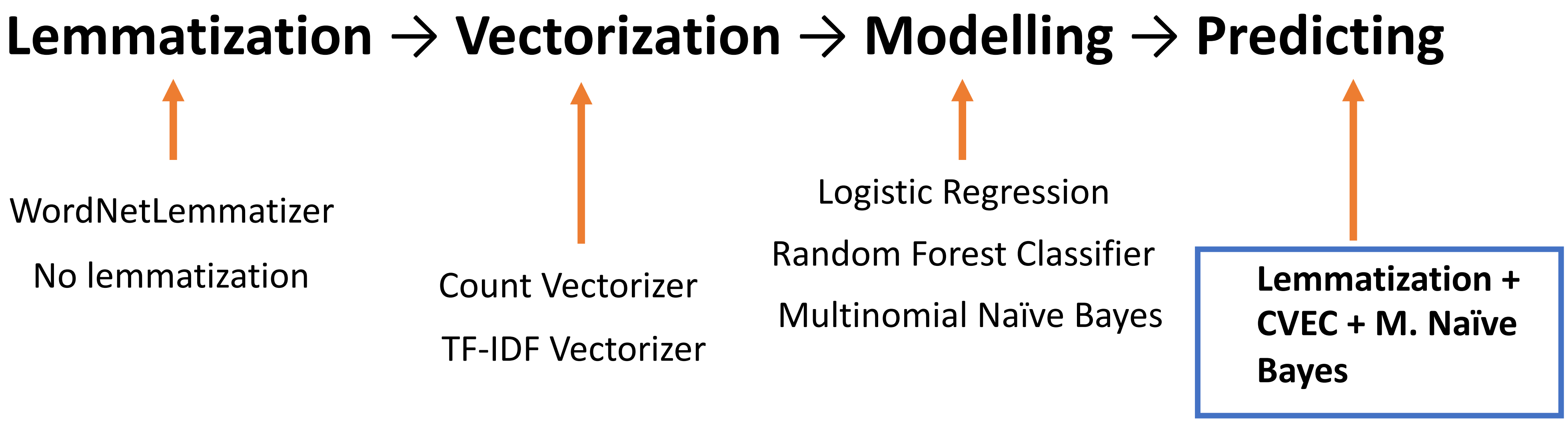


Sentiment Analysis


Sentiment Analysis: British Problems


Sentiment Analysis: British Success

Modelling



Lemmatization → Vectorization → Modelling → Predicting

WordNetLemmatizer

No lemmatization

Count Vectorizer

TF-IDF Vectorizer

Logistic Regression

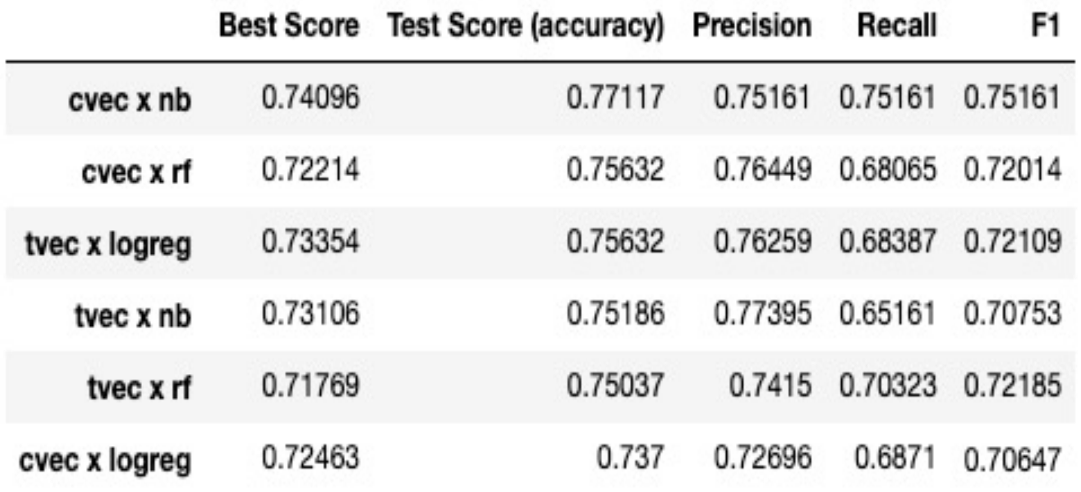
Random Forest Classifier

Multinomial Naïve Bayes

Lemmatization + CVEC + M. Naïve Bayes

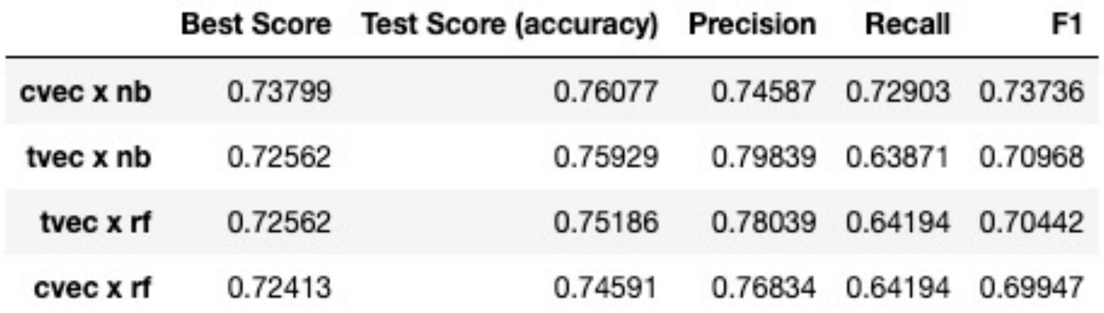
Results

Part 1: With Lemmatization



	Best Score	Test Score (accuracy)	Precision	Recall	F1
cvec x nb	0.74096	0.77117	0.75161	0.75161	0.75161
cvec x rf	0.72214	0.75632	0.76449	0.68065	0.72014
tvec x logreg	0.73354	0.75632	0.76259	0.68387	0.72109
tvec x nb	0.73106	0.75186	0.77395	0.65161	0.70753
tvec x rf	0.71769	0.75037	0.7415	0.70323	0.72185
cvec x logreg	0.72463	0.737	0.72696	0.6871	0.70647

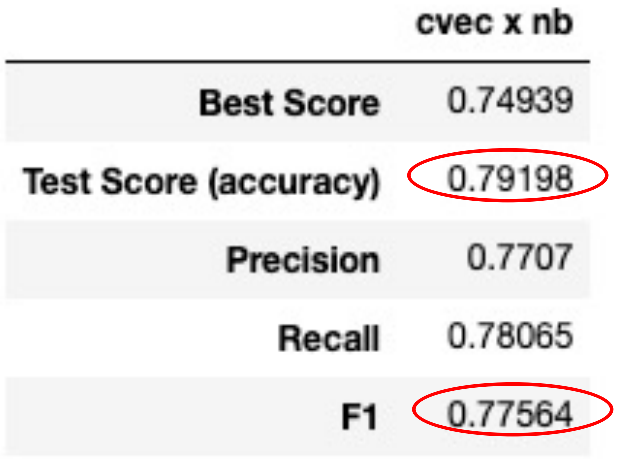
Part 2: w/o Lemmatization



	Best Score	Test Score (accuracy)	Precision	Recall	F1
cvec x nb	0.73799	0.76077	0.74587	0.72903	0.73736
tvec x nb	0.72562	0.75929	0.79839	0.63871	0.70968
tvec x rf	0.72562	0.75186	0.78039	0.64194	0.70442
cvec x rf	0.72413	0.74591	0.76834	0.64194	0.69947

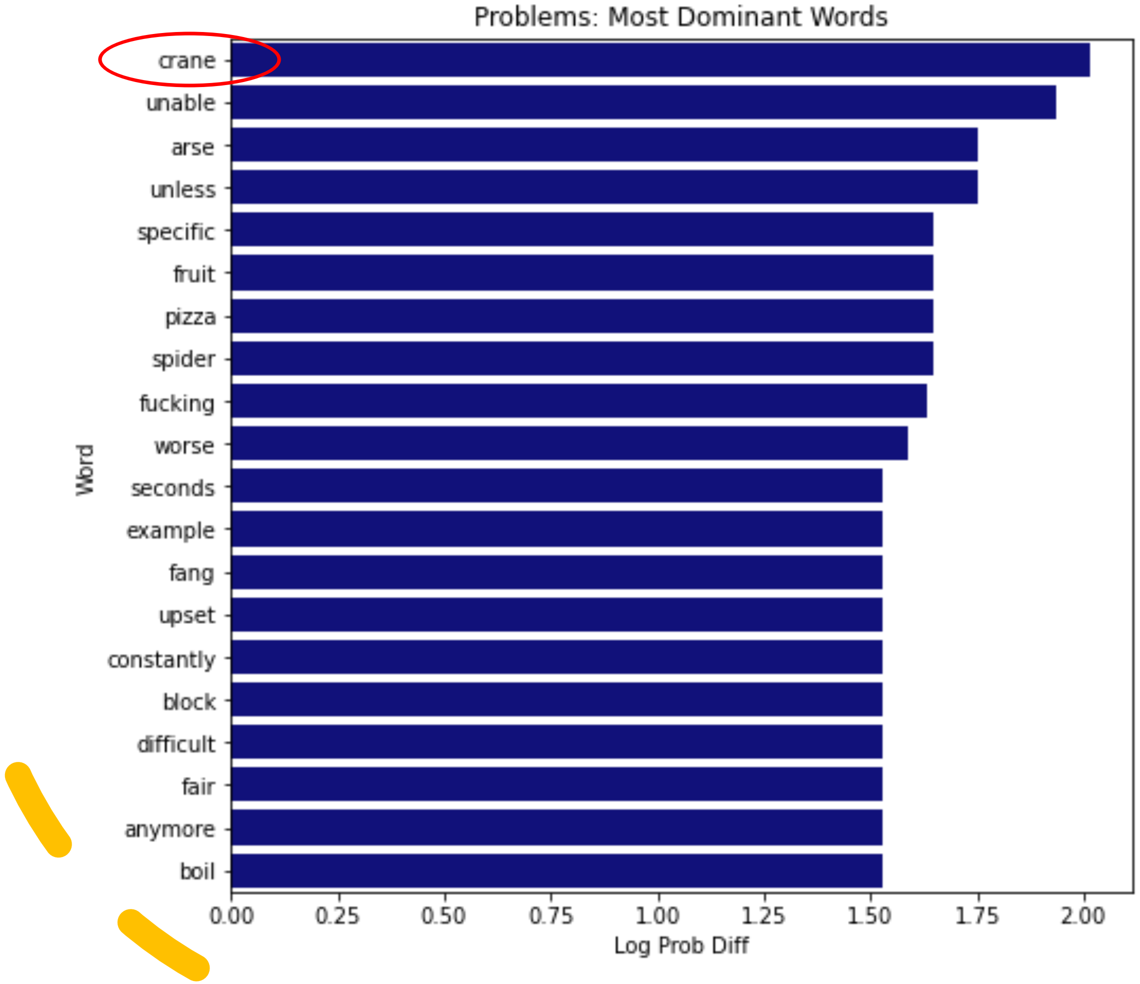
Final Model

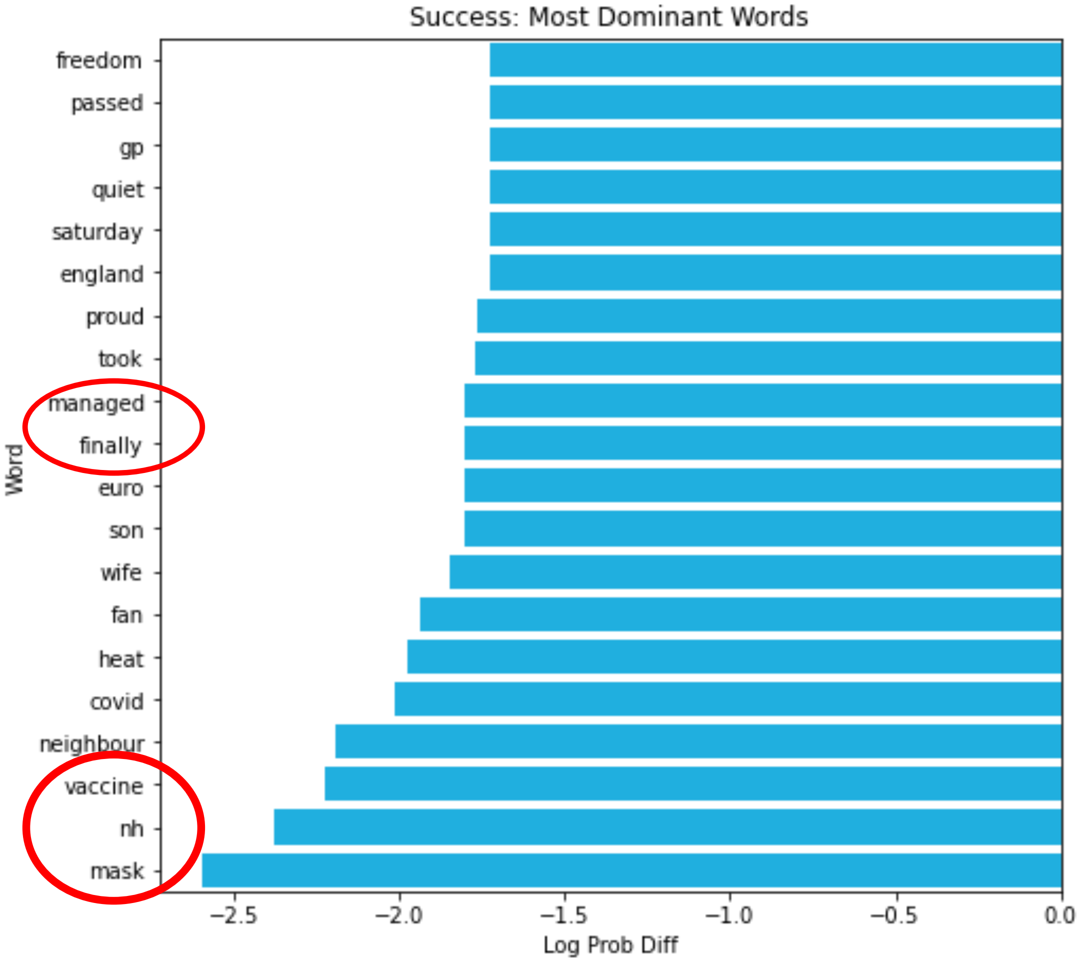
- Lemmatization
- CVEC
- Multinomial Naïve Bayes



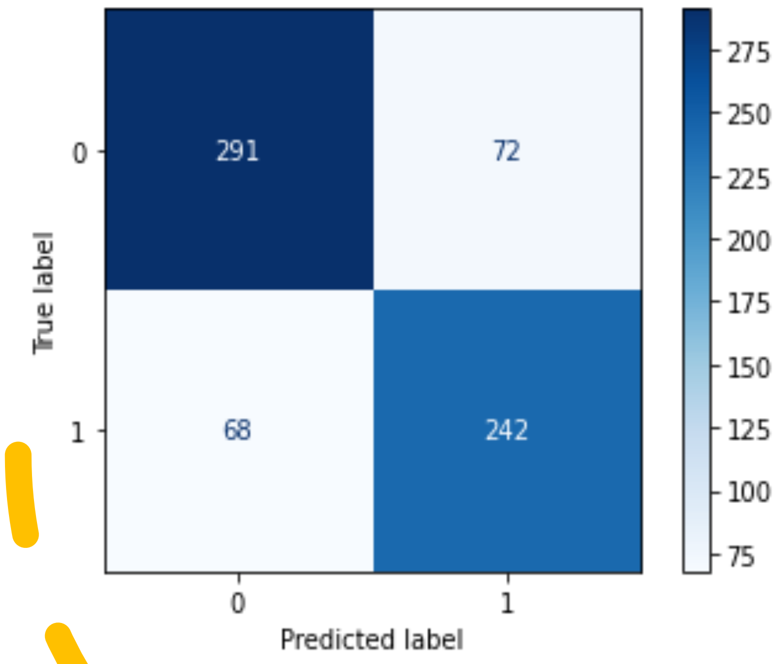
	cvec x nb
Best Score	0.74939
Test Score (accuracy)	0.79198
Precision	0.7707
Recall	0.78065
F1	0.77564

Differentiating Words





Gone wrong somewhere?



Sarcasm

'getting mildly <u>excited</u> buying a new vacuum cleaner head and realising you've reached <u>peak</u> middle age at 34.'

- Self deprecating humour – positive words in post

Irony/Punchline

'while in a queue of car going the speed limit in a 30, i got overtaken by a <u>numpty</u> in a black bmw hatchback going around 50…he rev up, go nowhere and a plume of smoke come out of the back… i don't usually enjoy the mishap of others, but <u>this really cheered up my day</u>!'

- Irony at play, bulk of the post was complaining about a fellow motorist – positive punchline is only delivered at the end
- Sentence embedding (e.g. BERT) instead of word

Gone wrong somewhere?

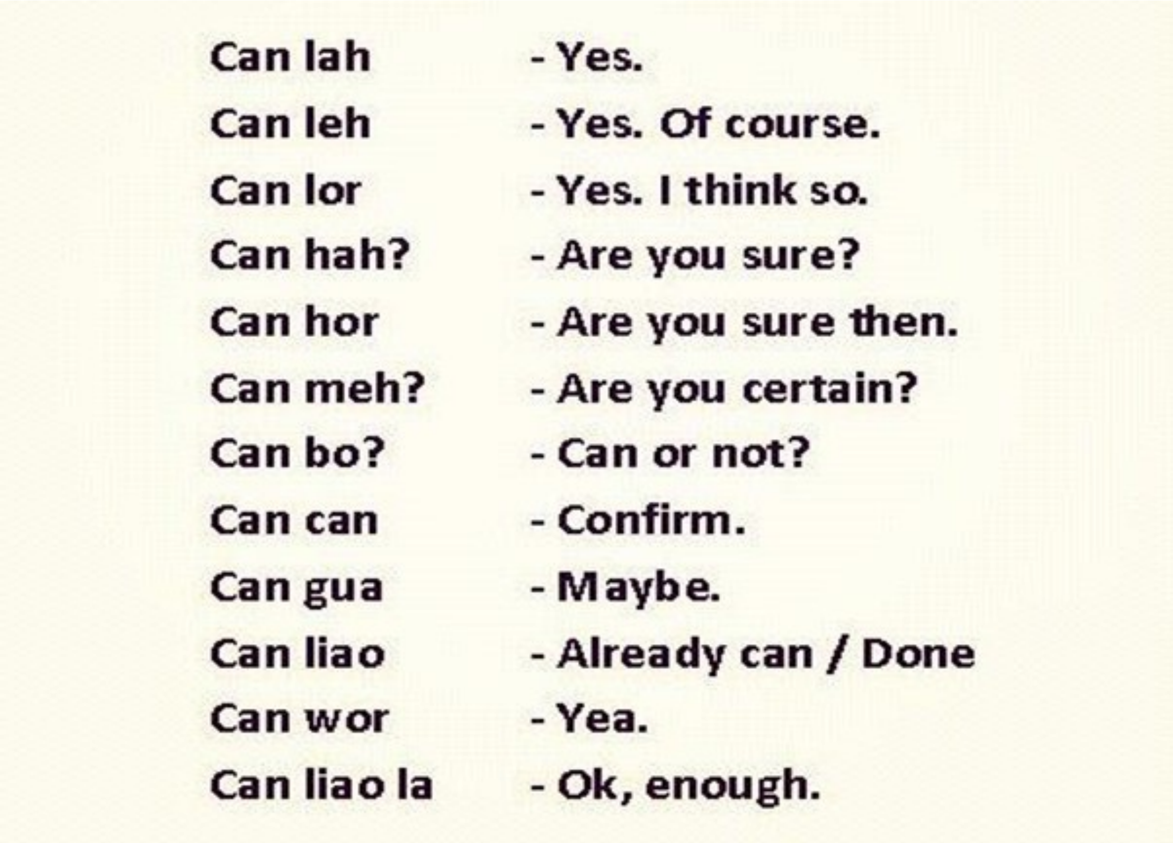
Coincidence

‘it get hoovered several time a week. is this some late-stage capitalism insanity, or am i just out of the loop? the sun is finally out after a wet saturday morning, perfect to enjoy some time in the garden, to the maddening soundtrack of next door hoovering their artificial grass. ‘

- Misclassified as a ‘success’.
- A few positive words (‘enjoy’, ‘finally’, ‘perfect’)
- ‘Saturday’ – showed up in most ‘different’ words associating it with British Success
- All 4 posts from British Problems are in test set and none in train set
- Larger dataset would help

Conclusion

- Good accuracy scores when predicting posts of two sub-reddits of 79%
- F1 score = 77.5%
- Sarcasm/Irony – sentence embedding
- Problem statement – SingaporeSuccess and SingaporeProblems could co-exist
- But…



Can lah	- Yes.
Can leh	- Yes. Of course.
Can lor	- Yes. I think so.
Can hah?	- Are you sure?
Can hor	- Are you sure then.
Can meh?	- Are you certain?
Can bo?	- Can or not?
Can can	- Confirm.
Can gua	- Maybe.
Can liao	- Already can / Done
Can wor	- Yea.
Can liao la	- Ok, enough.