reddit

PROJECT 3

CLASSIFICATION OF SUBREDDIT POSTS

Aw Boon Jun

PROBLEM STATEMENT

To classify whether a post is an uplifting news for reddit to boost positivity within the community.

DATA

r/UpliftingNews

A place to read and share positive and uplifting, feel good news stories.

r/worldnews

A place for major news from around the world, excluding US-internal news.

DATA

Baseline Accuracy: 0.62

r/UpliftingNews

Y=0

410 Posts

410 Titles

0 Selftext

r/worldnews

Y=1

672 Posts

672 Titles

0 Selftext

WORKFLOW

Web Scraping With Json API

Data Cleaning

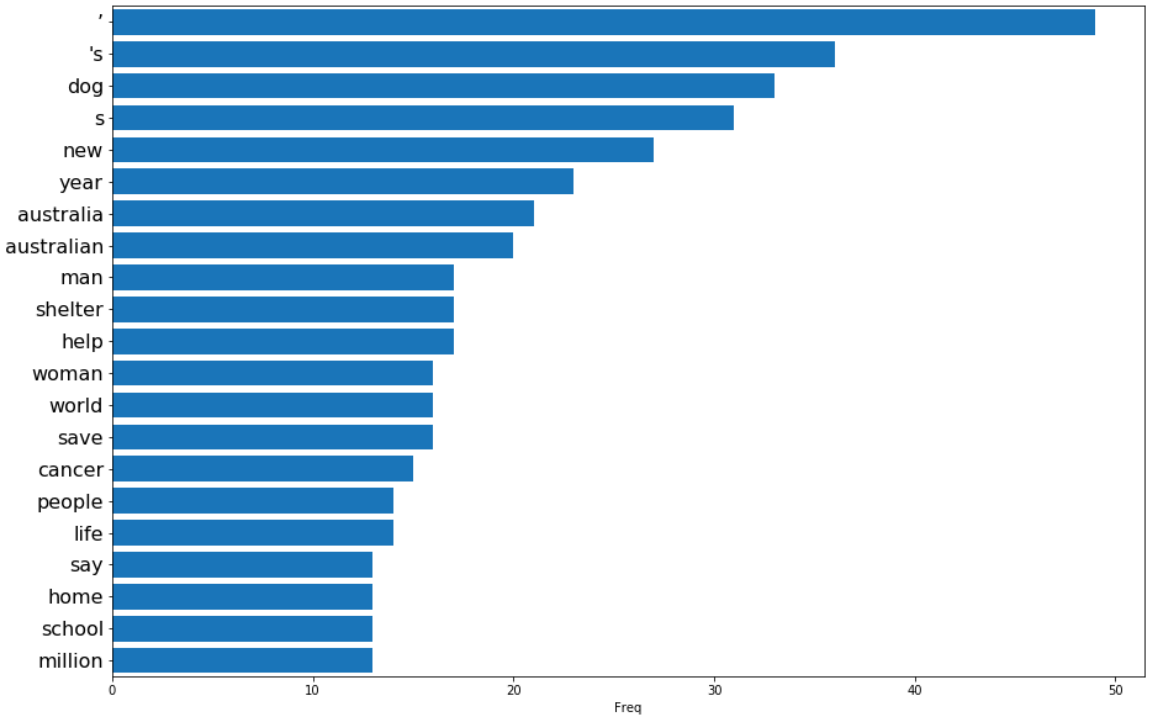
EDA

Model Validation: Logistic Regression & Multinomial Classification

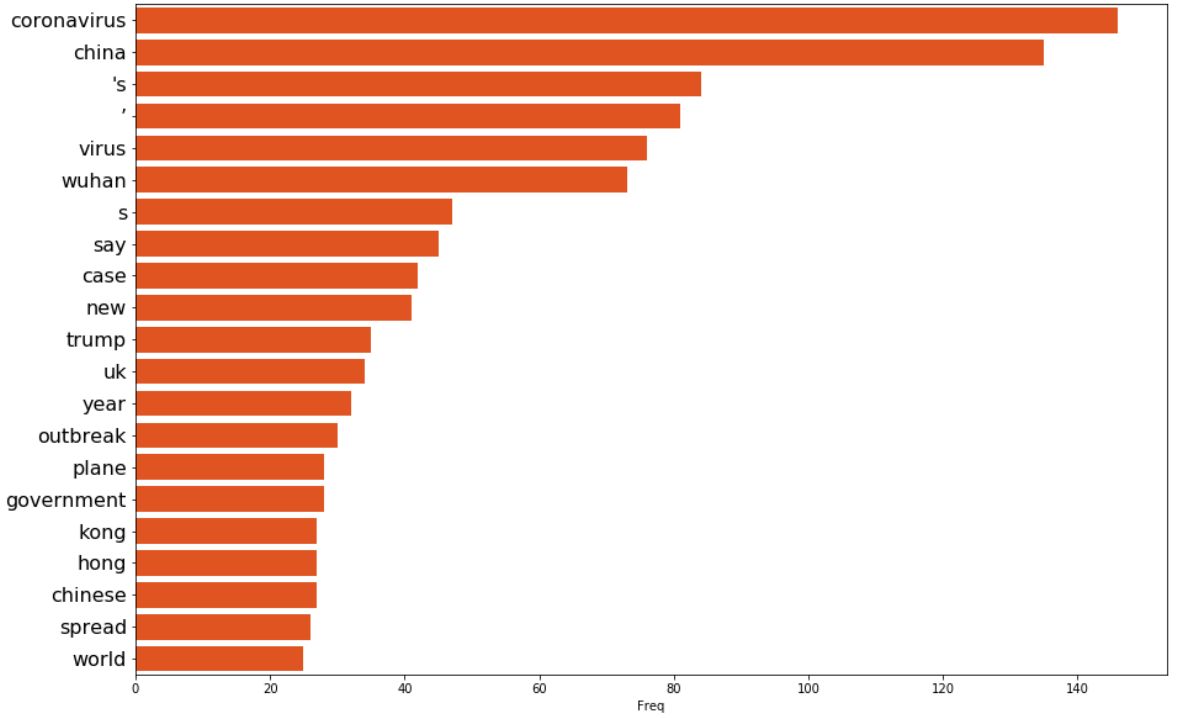
Model Fitting & Prediction

WORKFLOW - EDA

Uplifting News



World News



Words related to current events, which make sense.

A few ‘dirty’ words remain after count vectorizing: ‘s, ‘, & s

WORKFLOW – MODEL VALIDATION

Logistic Regression

VS

Multinomial

Vectorizing Method	Score
Count Vectorizer	**0.89**
TFIDF	0.83

Vectorizing Method	Score
Count Vectorizer	0.88
TFIDF	0.87

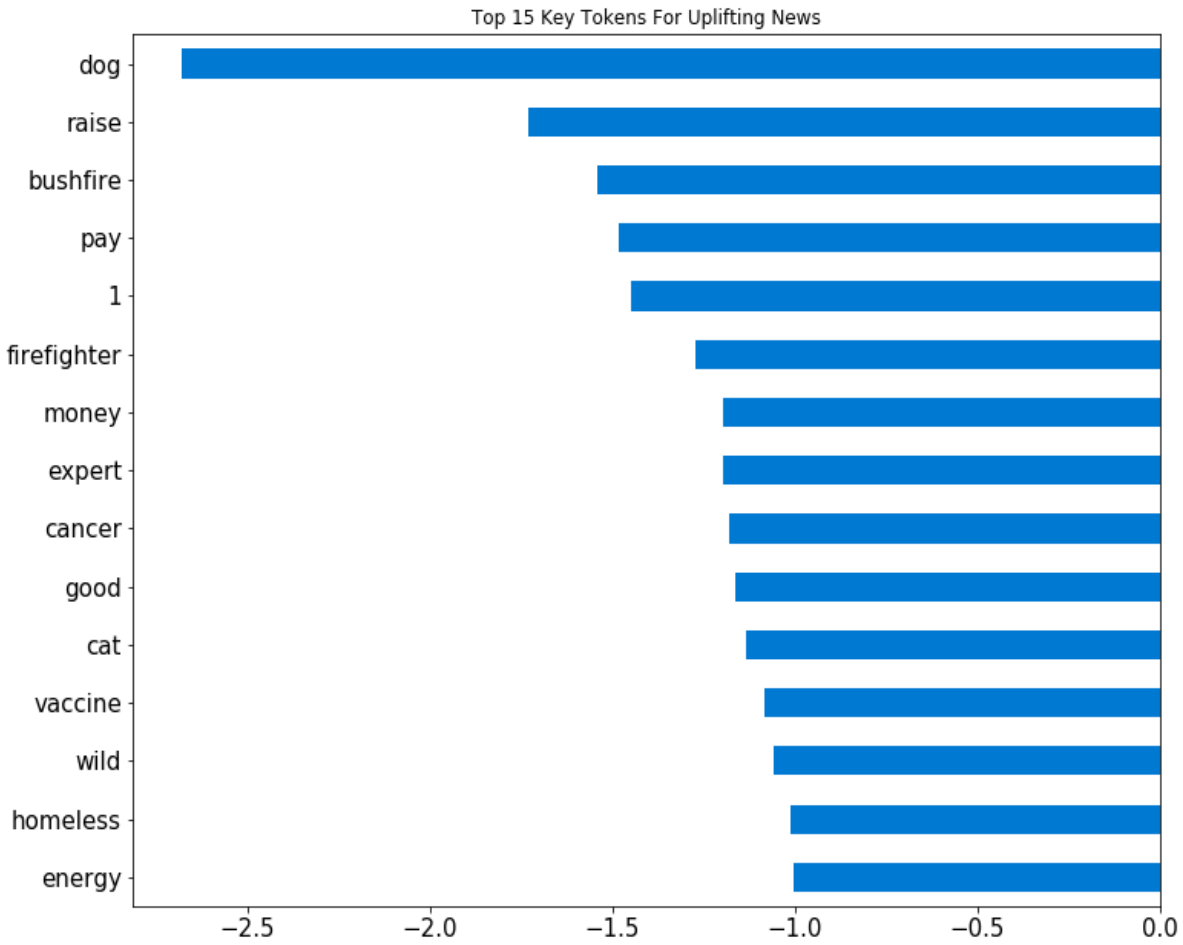
WORKFLOW – MODEL FITTING

Logistic Regression

Cvec params	Max features = 400 Max_df = 0.5 Min_df = 2 Ngram_range = (1,2)
Penalty	l1
Training Score	0.89
Testing Score	0.82

WORKFLOW – MODEL EVALUATION

Logistic Regression

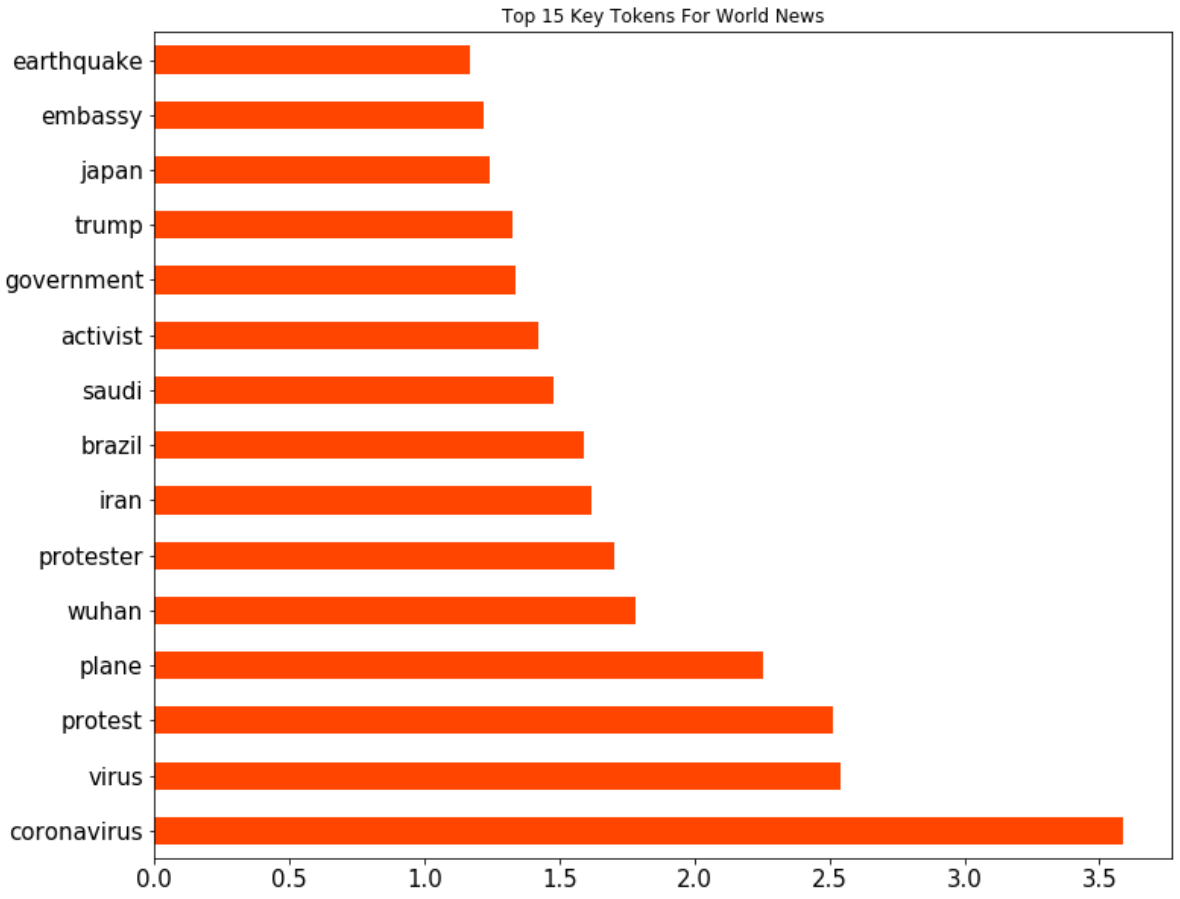


- Key Tokens make senses.
  - Dogs are men's best friend
  - Cats are the king of internet
  - Australian bushfire is getting better

WORKFLOW – MODEL EVALUATION

Logistic Regression

- Key Tokens make senses.
  - Correspond to recent current event



WORKFLOW – MODEL EVALUATION

Logistic Regression

- Model perform as expected in classifying the 2 subreddit posts

Limitation

- Classifying news related subreddit is very time specific
- Tokens such as Corona Virus are not relevant for news last year

Confusion Matrix	
True Negatives	92
False Positives	11
False Negatives	40
True Positives	128

Accuracy Comparison	
Baseline	0.62
Model	**0.81**

RECOMMENDATIONS

- Only can be trained on the most recent 500-700 posts due to API limitation
- Would be more accurate and less time-dependent if more historic data of the posts can be used to train the LR model

reddit

THANK YOU