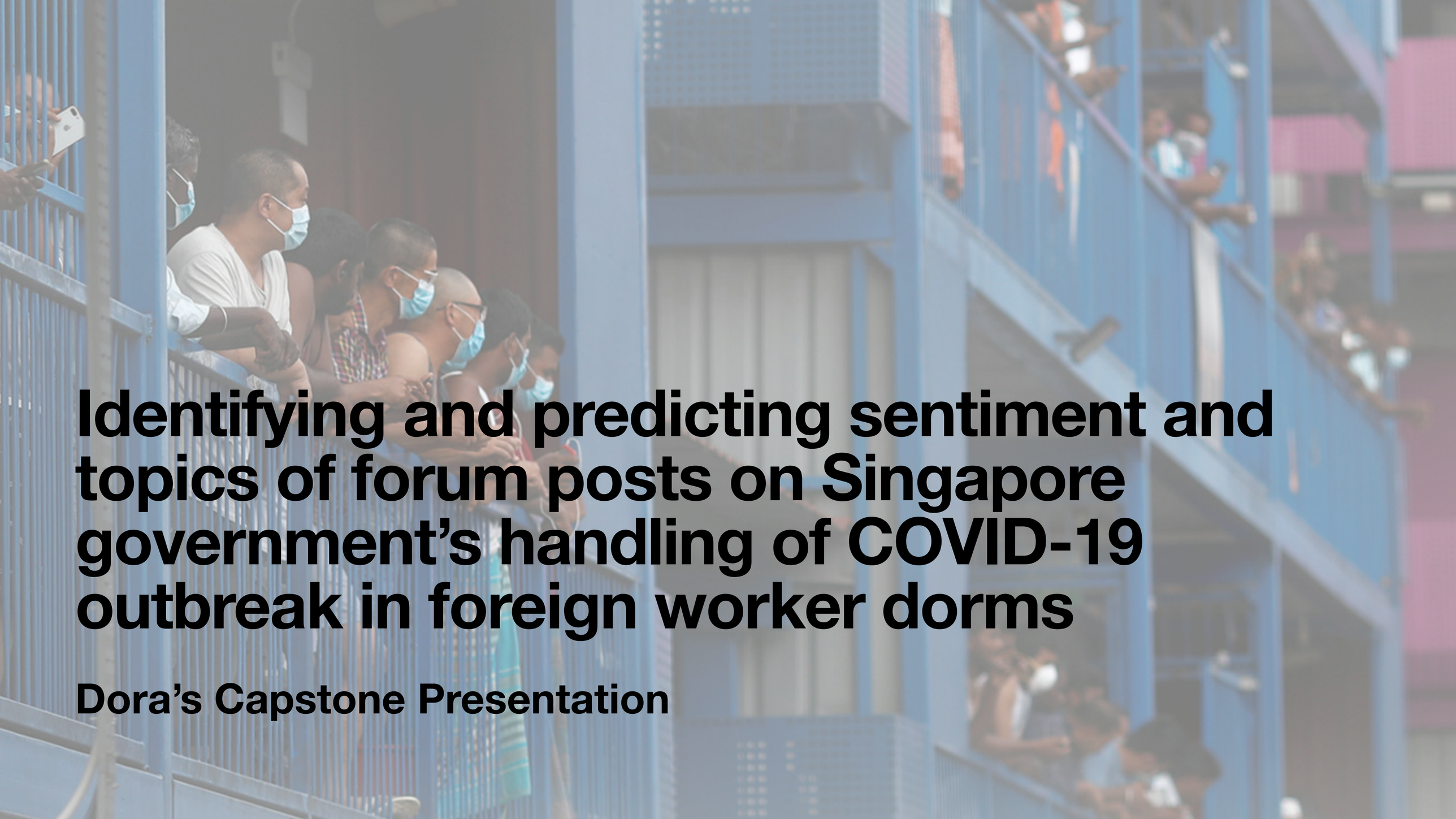


Identifying and predicting sentiment and topics of forum posts on Singapore government's handling of COVID-19 outbreak in foreign worker dorms

Dora's Capstone Presentation

Background

- As of 6 Aug 20, 51,633 of 54,555 COVID-19 cases belonged to foreign workers residing in dorms
- First COVID-19 cluster in foreign worker dorms found on 1 Apr 20
- Government has started testing all workers living in foreign dorms, isolating infected workers and restricting their movement beyond the dorms
- The government has received criticism for not acting swiftly enough to curb COVID-19 spread in dorms since the first foreign worker case was detected on 9 Feb 20
- Netizens have expressed their opinions about the government’s handling of the issue on forums

Problem Statement

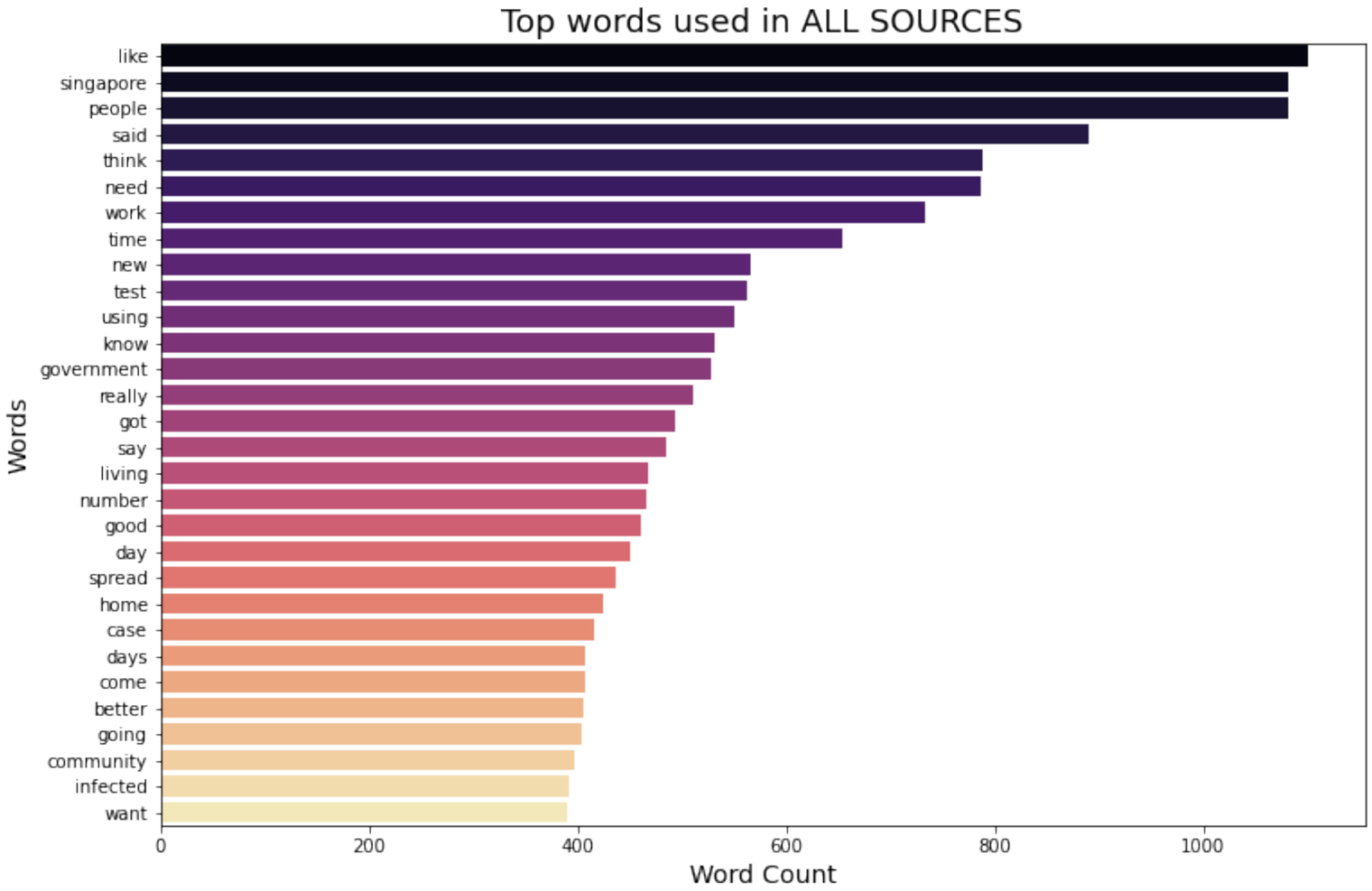
- Develop a model to identify sentiment and key topic of forum posts on government's handling of COVID-19 cases in foreign worker dorms
  - Sentiment Analysis
  - Topic modeling
- Using Supervised Machine Learning and Recurrent Neural Networks
- Performance of models evaluated on accuracy, F1 and ROC AUC scores (sentiment analysis) as well as coherence scores (topic modeling)

Data Collection and Cleaning

- Scraped posts from Reddit, Hardware Zone, SG Talk
  - Based on search term “foreign worker dormitories COVID-19”
- Preliminary data cleaning
  - Removing http addresses and usernames
  - Removing non-english text
- Preprocessing - tokenize, lemmatize, removed stopwords

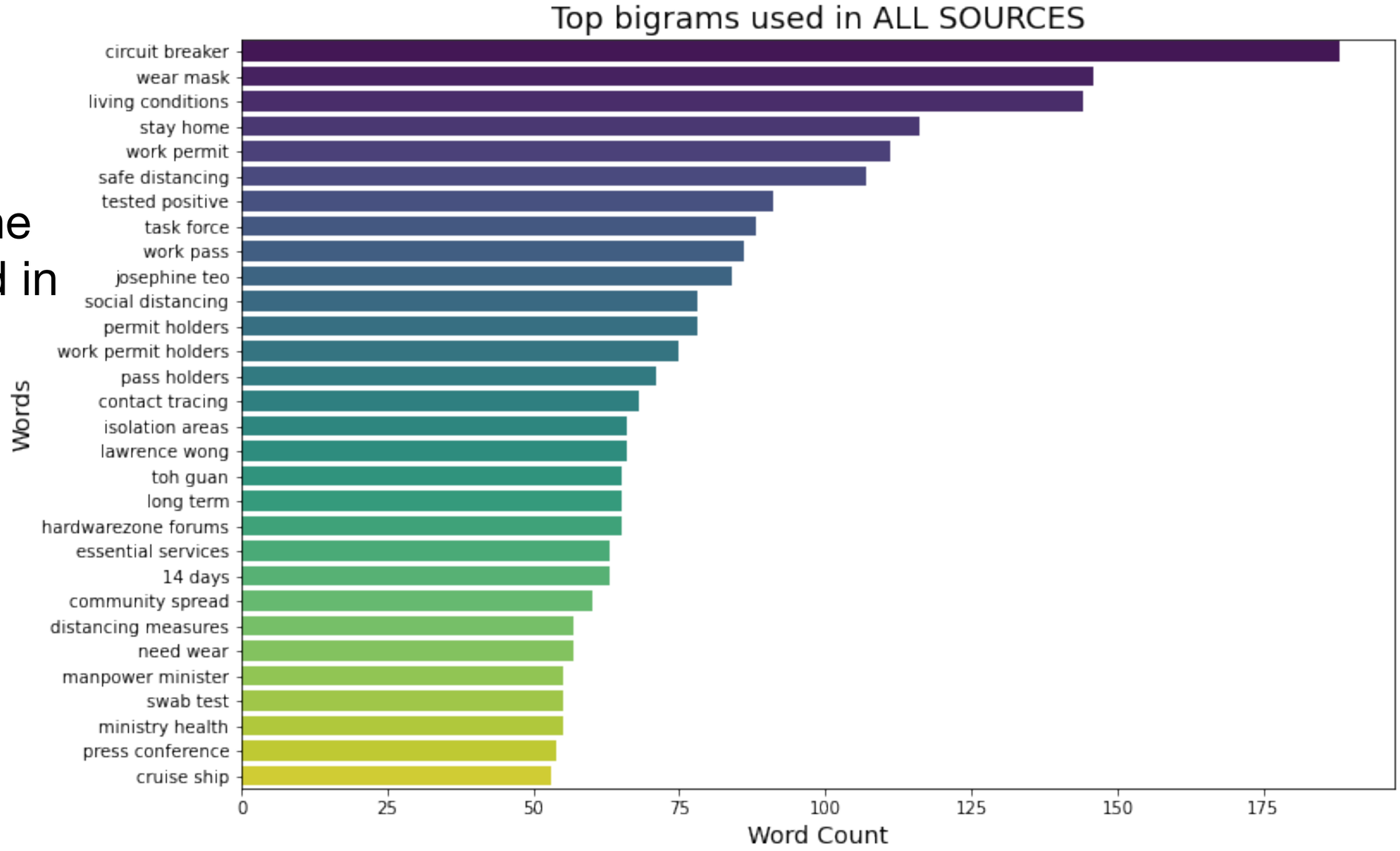
EDA: Top words

- test, government, numbers, spread, community, infected
- Indicates forum posts are about the topic of the government's handling of the COVID-19 outbreak in dorms



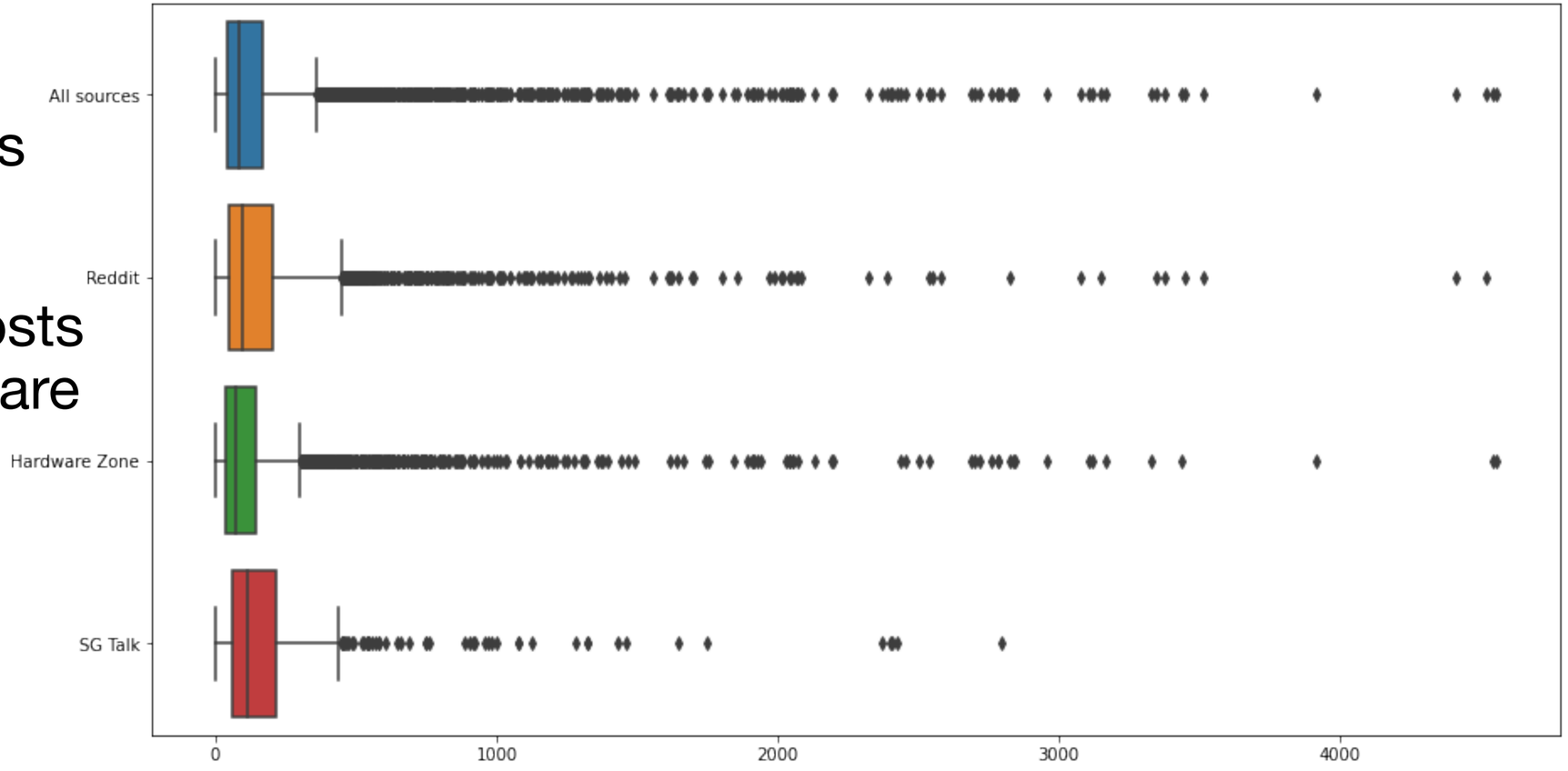
EDA: Top bigrams

- Bigrams more informative on the topics discussed in the forum



EDA: Length of posts

- Mean number of words per post is 177
- Reddit and SG Talk posts are longer than Hardware Zone posts in general



Labelling sentiment on forum posts

- Preprocessing: Tokenize and Lemmatize
- Different treatment of stop words
  - LSTM RNN model: Did not remove as sequence of words is important
  - Multinomial NB: stopwords removed
    - Left in words indicating sentiment: No, not, against, shouldn’t…
    - Removed in “common” words: foreign, migrant, worker, dormitory, covid…
- Issue of unbalanced classes: 86% neutral, 11% negative, 3% positive
  - Oversampling and undersampling of classes did not improve performance of model significantly

Labelling sentiment on forum posts

- Tried unsupervised learning models; TextBlob and VADER did not work well
- Sentiment of posts in train set had to be labelled through supervised learning models
- Multinomial Naive Bayes model gave best results
- Process:
  - Split data into train (8214 posts) and test (2056 posts)
  - Manually label sentiment of first 1000 posts in train set
  - Train first 1000 labelled posts with Multinomial NB model
  - Predict sentiment for next 1000 unlabelled posts
  - Check for accuracy and make changes to incorrect labels
  - Collate labelled posts and train Multinomial NB model
  - Label next 1000 posts
  - Repeat until all posts in train set have been labelled

Sentiment Analysis Production Model

- Several models assessed - Multinomial Naive Bayes, LSTM RNN, BERT
- Multinomial Naive Bayes performed best
  - Test F1 Score: 0.873
  - Test ROC AUC Score: 0.726
  - Better able to predict minority classes (positive/negative sentiment)

Sentiment Analysis Insights

- Top words indicating positive or negative sentiment posts had many similarities: people, government, singapore, numbers
- Unique top words for negative sentiment posts
  - ‘PAP' (People's Action Party) and 'MOM' (Ministry of Manpower) - called out specific party, government agency
  - 'Still' - indicates a sentiment in these posts that COVID-19 cases in dorms remain high
  - Mask' - perhaps taking issue with the Government's delay in asking people to wear masks in public
- Unique top words for positive sentiment posts
  - 'Would', 'should' - indicating some kind of suggestion in the post

Topic Modeling

- Topic modeling to find key topics of posts with positive and negative sentiment
- Preprocessing and removing stop words
  - Only removed “noisy” stopwords (lah, liao, ah, gagt), left keywords in (dormitories, foreign, worker, covid, government)
- Latent Dirichlet Allocation with MALLET implementation used to identify key topics
  - Better coherence score than vanilla LDA model
  - Topics are more coherent and better separated

Insights from Topic Modeling

Negative Sentiment

- **dormitory, government, operators, jo, pay, conditions, living, money, years, problem**
- **dormitory, cases, covid, situation, fw, community, spread, home, case, end**
- **workers, foreign, mom, pap, work, employers, worker, teo, covid, feb**
- mask, people, masks, singapore, pap, wear, blame, healthy, care, world
- government, time, sg, ministers, good, people, long, give, point, standard

Insights from Topic Modeling

Positive Sentiment

- **cases, community, numbers, dormitory, news, forum, big, spread, cb, world**
- **conditions, pay, migrant, thing, issue, country, cost, distancing, operators, singaporeans**
- **government, sg, high, citizens, companies, care, hard, day, agree, lot**
- people, masks, time, mask, lockdown, weeks, home, hindsight, long, wearing
- work, make, clusters, measures, start, point, days, making, made, life

Limitations and Future Steps

- Generalisibility
  - Relatively small dataset (10,000 posts)
  - May not be able to accurately predict sentiment and topics of posts on another topic
- Unable to use pretrained text analysis libraries (BERT, Textblob, VADER)
  - Different words and syntaxes
- Going forward, train models with more forum posts on different topics

Deployment Model for Sentiment Analysis

“Announce the stricter measurements, but wont implement it until 4 days later. You jolly well know that the number of cases is spiking yet you still dont want to implement it immediately? Jolly well know Singaporeans will play around. Its like your parents want to cane you, but they allow you to go play games before caning you. Classic PAP blunder yet again. Decisive bold gold standard ok”

Thanks for your attention :)