BC Stats

Text Analytics:

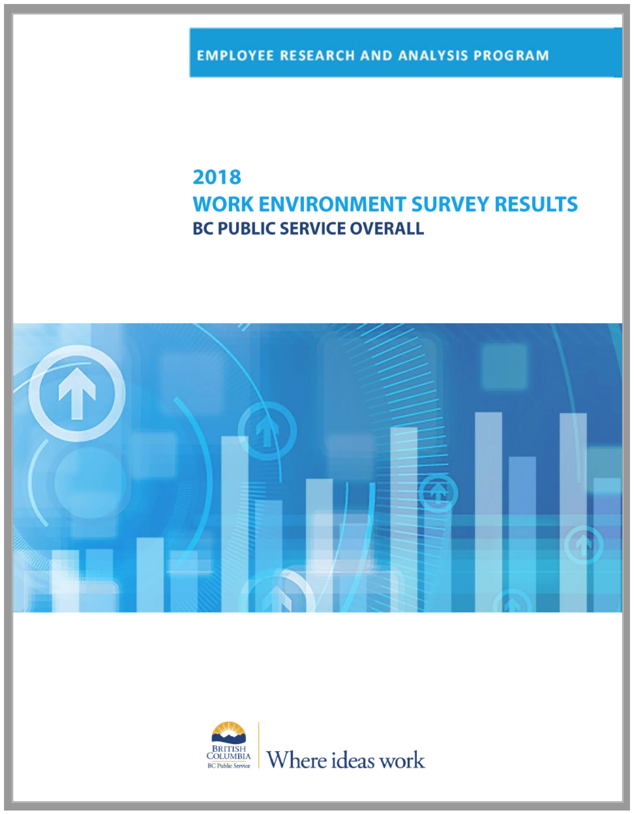
Quantifying the Responses to Open-Ended Survey Questions

Carlina Kim, Karanpal Singh, Sukriti Trehan, Victor Cuspinera
Partner: BC Stats | Mentor: Varada Kolhatkar

2020-06-19

Introduction

The Survey



Work Environment Survey (WES)

- conducted by BC Stats for employees within BC Public Service
- measures the health of work environments and identifies areas for improvement
- ~80 multiple choice questions (5 point scale) and 2 open-ended questions

Data

Open-ended Question Responses

Question 1

- **What one thing would you like your organization to focus on to improve your work environment?**

Example: *"Better health and social benefits should be provided."*

Question 2

- **Have you seen any improvements in your work environment and if so, what are the improvements?**

Example: *"Now we have more efficient vending machines."*

*Note: these examples are fake comments for privacy reasons.

Data Example of Question 1

What one thing would you like your organization to focus on to improve your work environment?

Comments*	CPD	CB	EWC	...	CB_Improve_benefits	CB_Increase_salary
Better health and social benefits should be provided	0	1	0	...	1	0

Theme: CB = Compensation and Benefits

Sub-theme: CB_Improve_benefits = Improve benefits

Question 1: +31,000 labelled comments for 2013, 2018, 2020, +12,000 additional comments from 2015

Question 2: +6,000 labelled comments for 2018, +9,000 additional comments from 2015, 2020

*Note: this is a fake comment as an example of the data.

Objectives

1) Build a model to automate multi-label text classification that:

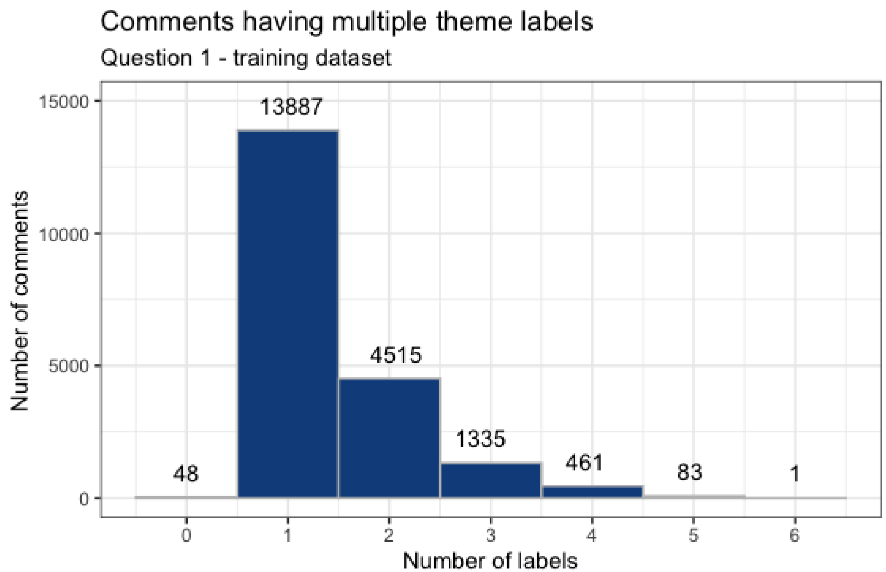
- predicts label(s) for Question 1 and 2's main themes
- predicts label(s) for Question 1's sub-themes

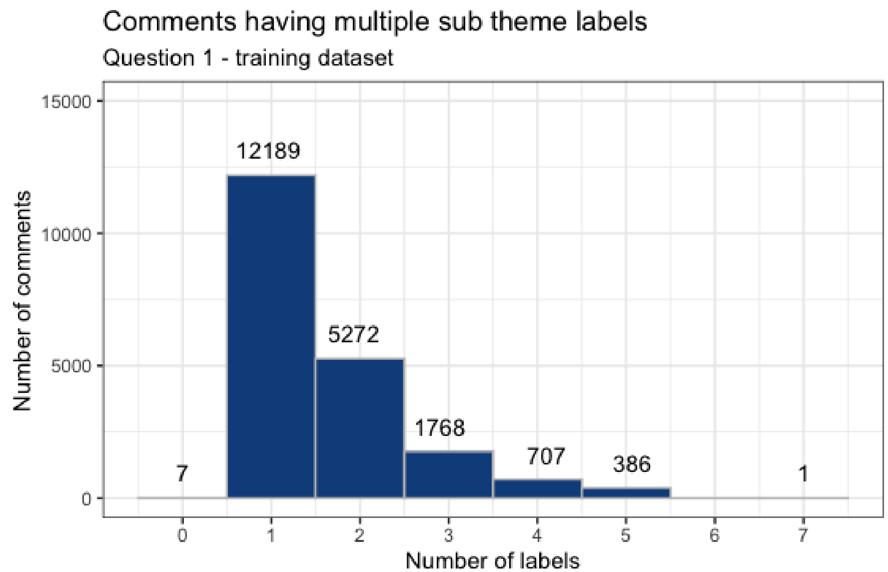
2) Visualizations on discovery of text analysis:

- mapping words for both questions to identify common texts
- identify potential needs & resolutions using sentimental analysis
- identify theme trends across ministries over given years

Challenges with data

Sparsity



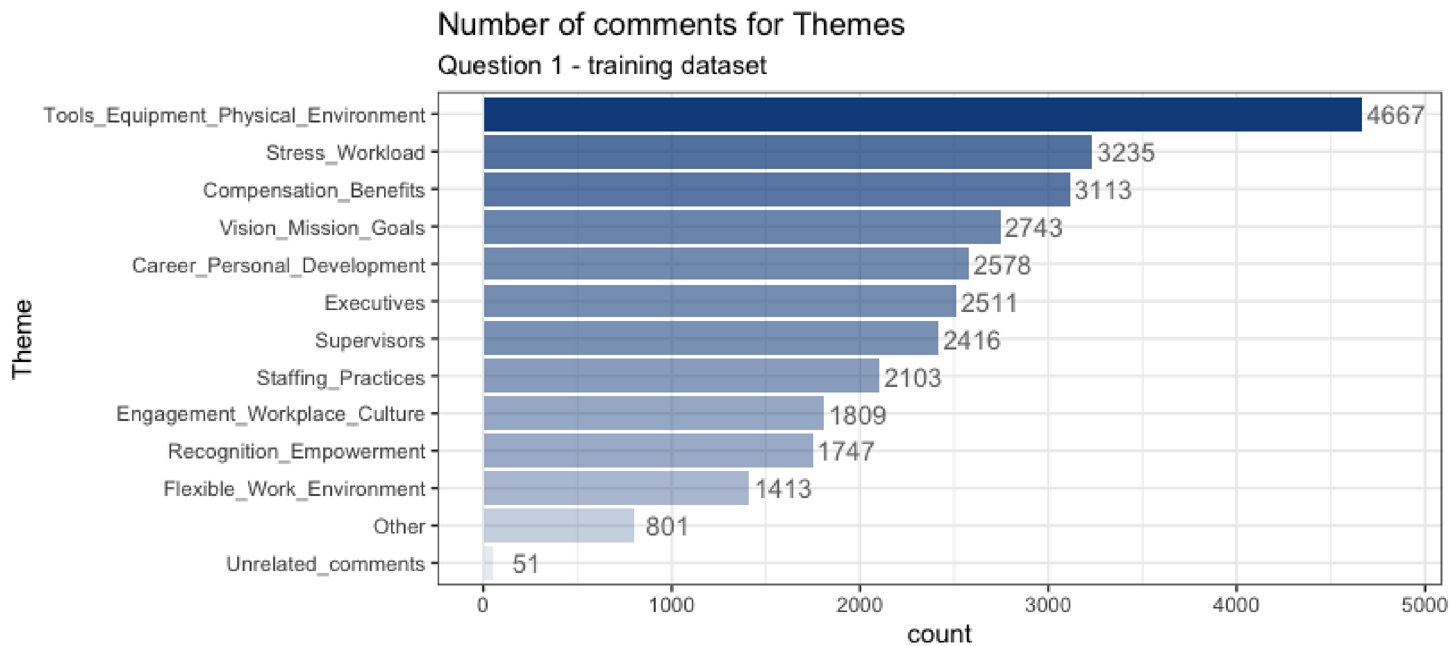


There are 12 themes and 63 subthemes that comments can be encoded into.

- Label cardinality for **themes: ~1.4** and for **subthemes: ~1.6**

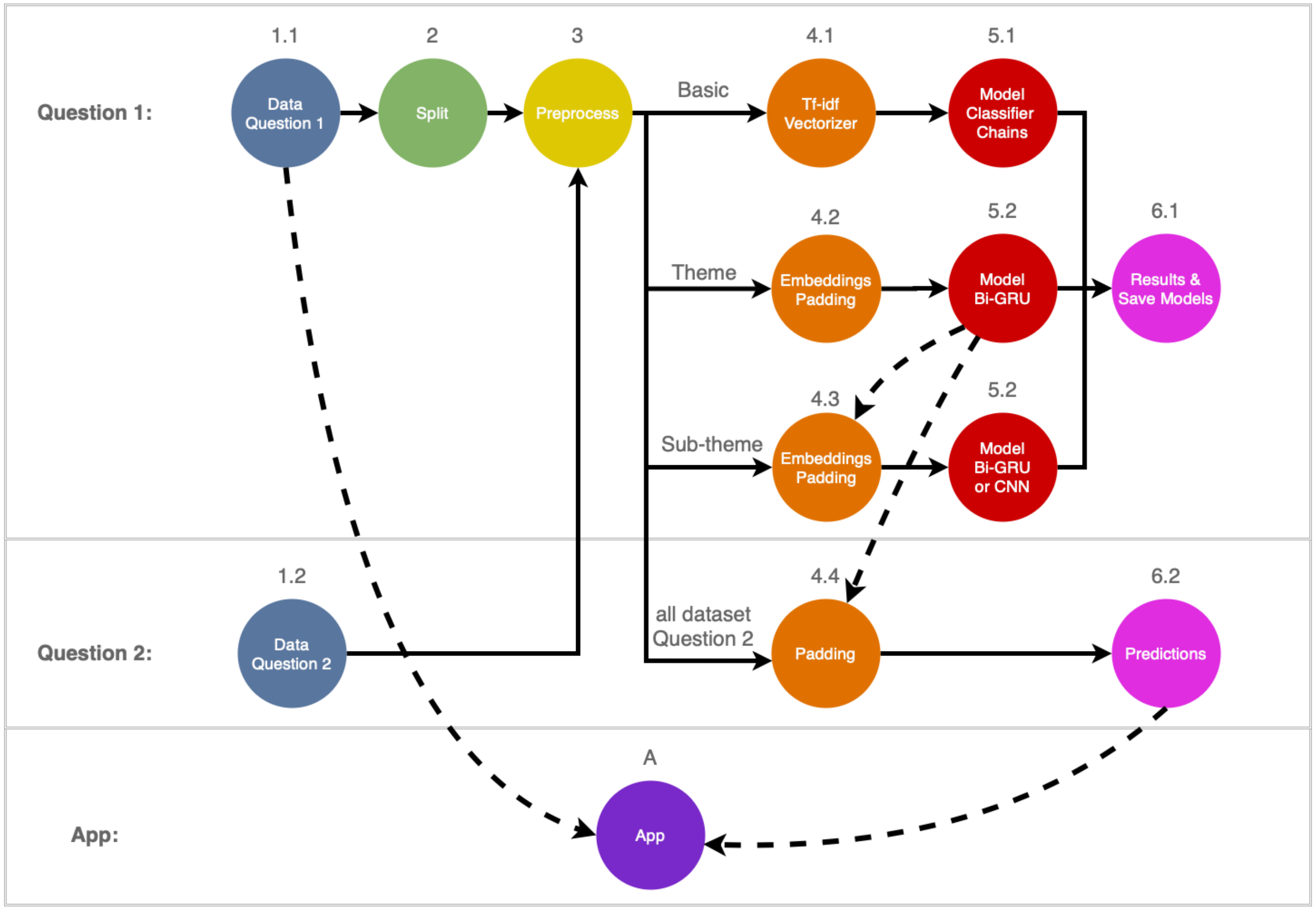
Challenges with data

Class Imbalance



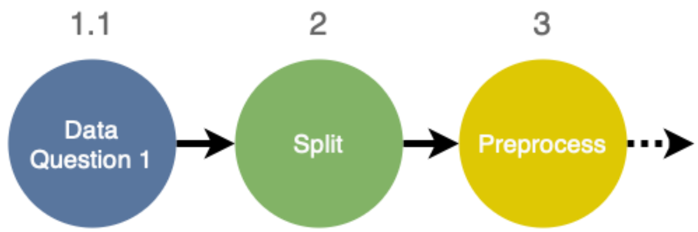
Imbalanced data in each theme

Text classification methodology



8/23

Data Split & Preprocessing

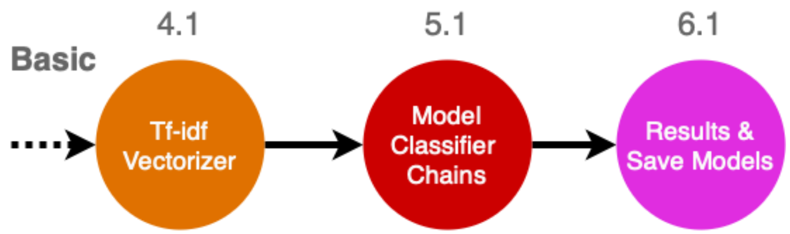


- Raw -> 80% train, 20% test.
- Training -> 80% train, 20% validation

- remove sensitive information using Named Entity Recognition (NER) to remove person, organization, location, and geopolitical entity from data
- all social media handles to "social media" instead of "Facebook", "Instagram", "Twitter"
- removed punctuation and lowercase for tokenization

Example comment to get flagged: **"George and I love when the deparment gives us new coupons!"**

Baseline Model: Classifier Chains



- **TF-IDF Vectorizer** uses weights instead of token counts (CountVectorizer)
- **Classifier Chains** preserves order and occurence of labels
  - multiple scikit-learn base classifiers tried (RandomForest, GaussianNB, etc)
  - best result with LinearSVC

X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

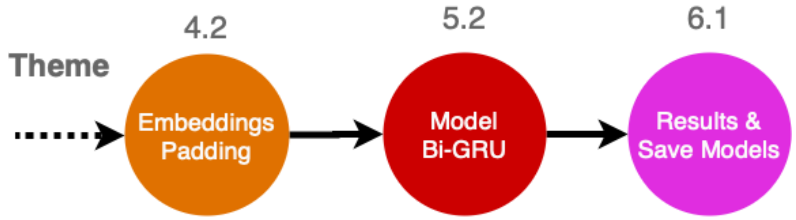
Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

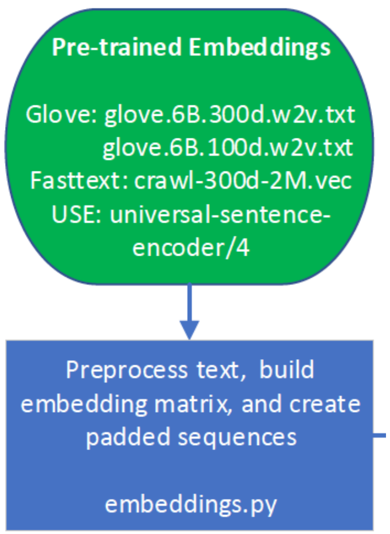
Classifier 4

Source: Multi-Label Classification: Classifier Chains, by Analytics Vidhya

Advanced Model: Pre-Trained Embeddings



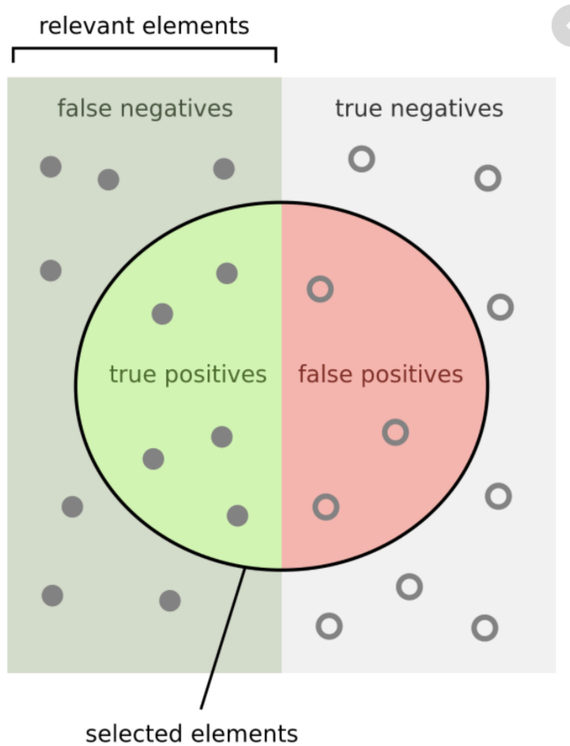
Fasttext, Glove, Universal Sentence Encoder



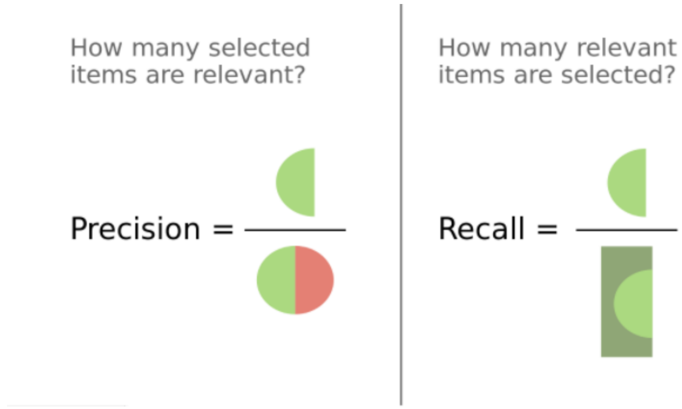
- explored several embeddings on various models
- built embedding matrix & maximized vocab coverage for each embedding
- transformed comments to padded data to fit into embedding size
- removed sensitive data using embeddings to upload into public cloud services for our advanced models

How we measured success

Precision & Recall

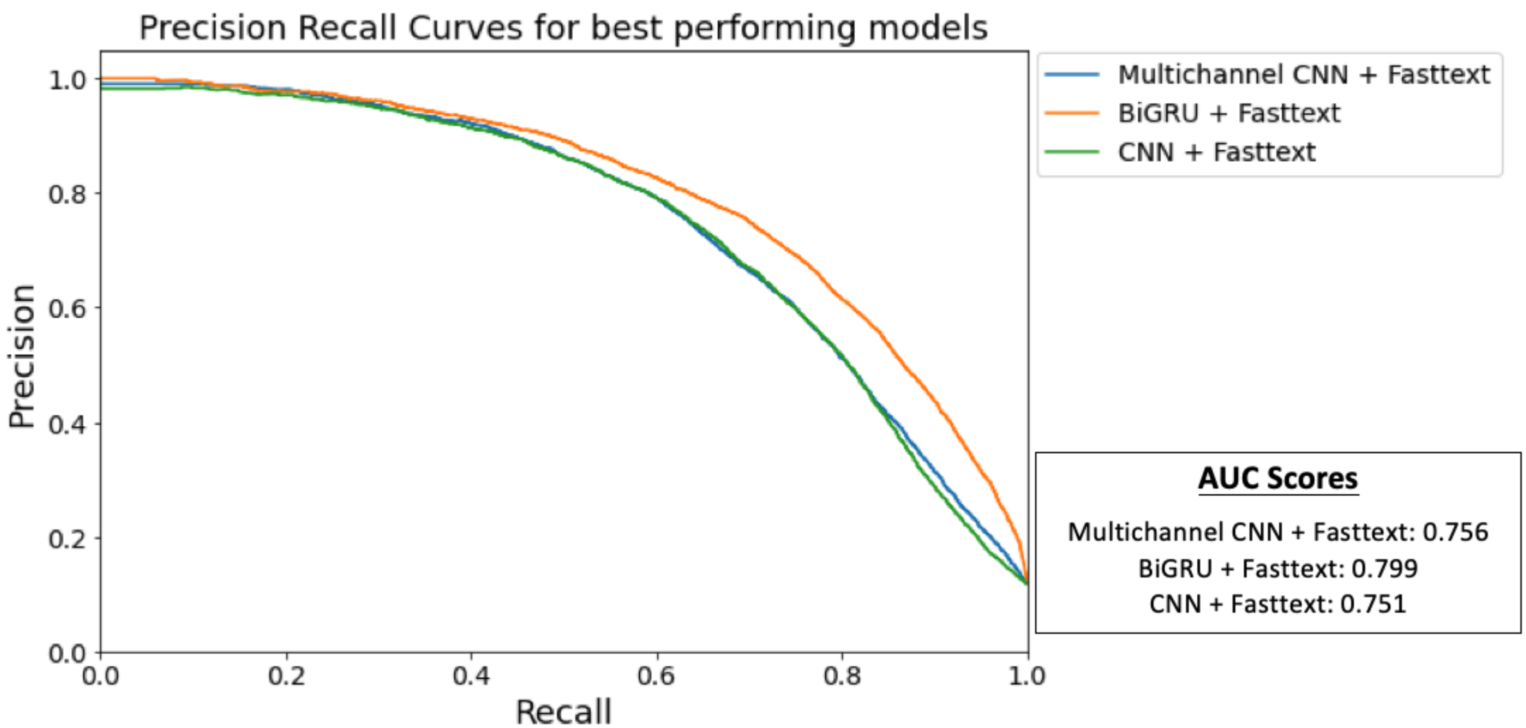


Source: Precision and Recall



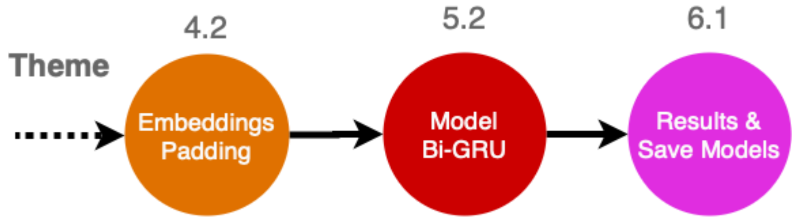
- Precision Recall curve: plotting precision vs recall at various threshold rates
- Micro-average: weighted average of the precision and recall

Precision Recall Curve for Q1 Themes



13/23

Our advanced model: Fasttext + BiGru



Threshold	Accuracy	Precision	Recall	F1
0.3	0.513	0.714	0.744	0.7287
0.4	**0.531**	**0.751**	**0.709**	**0.7293**
0.5	0.534	0.781	0.674	0.7234
0.6	0.534	0.811	0.638	0.6979
0.7	0.526	0.836	0.599	0.6726

Model Results for Theme Labelling

Model	Accuracy	Precision	Recall	F1
TFID + LinearSVC	0.50	0.79	0.63	0.70
Fasttext + BiGru	0.54	0.75	0.71	0.73

2019 Capstone team's results

Model	Accuracy	Precision	Recall
Bag of Words + LinearSVC	0.45	0.74	0.64
Fasttext + BiGru	0.53	0.83	0.66

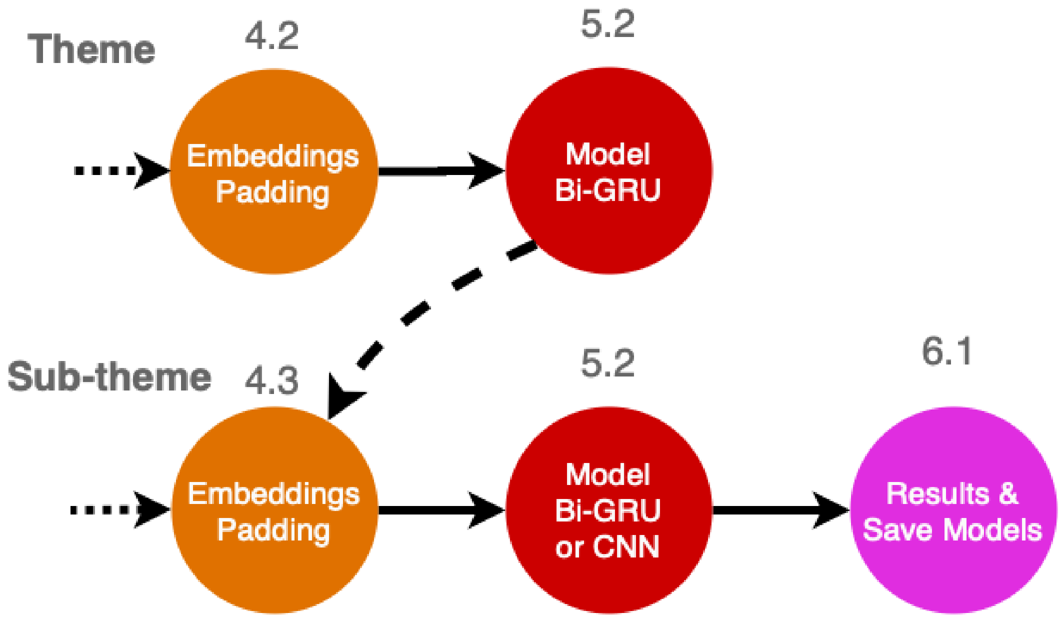
Source: BC Stats Capstone 2019-Final Report, by A. Quinton, A. Pearson, F. Nie

Results for Fasttext + BiGru

Predicting each theme

Theme	Accuracy	Precision	Recall	Theme	Accuracy	Precision	Recall
CPD	0.94	0.77	0.79	**RE**	0.94	0.69	0.51
CB	0.97	0.90	0.90	**Sup**	0.92	0.66	0.57
EWC	0.94	0.69	0.56	**SW**	0.92	0.74	0.65
Exec	0.92	0.64	0.71	**TEPE**	0.95	0.92	0.85
FEW	0.97	0.73	0.77	**VMG**	0.90	0.62	0.66
SP	0.95	0.76	0.75	**OTH**	0.96	0.43	0.29

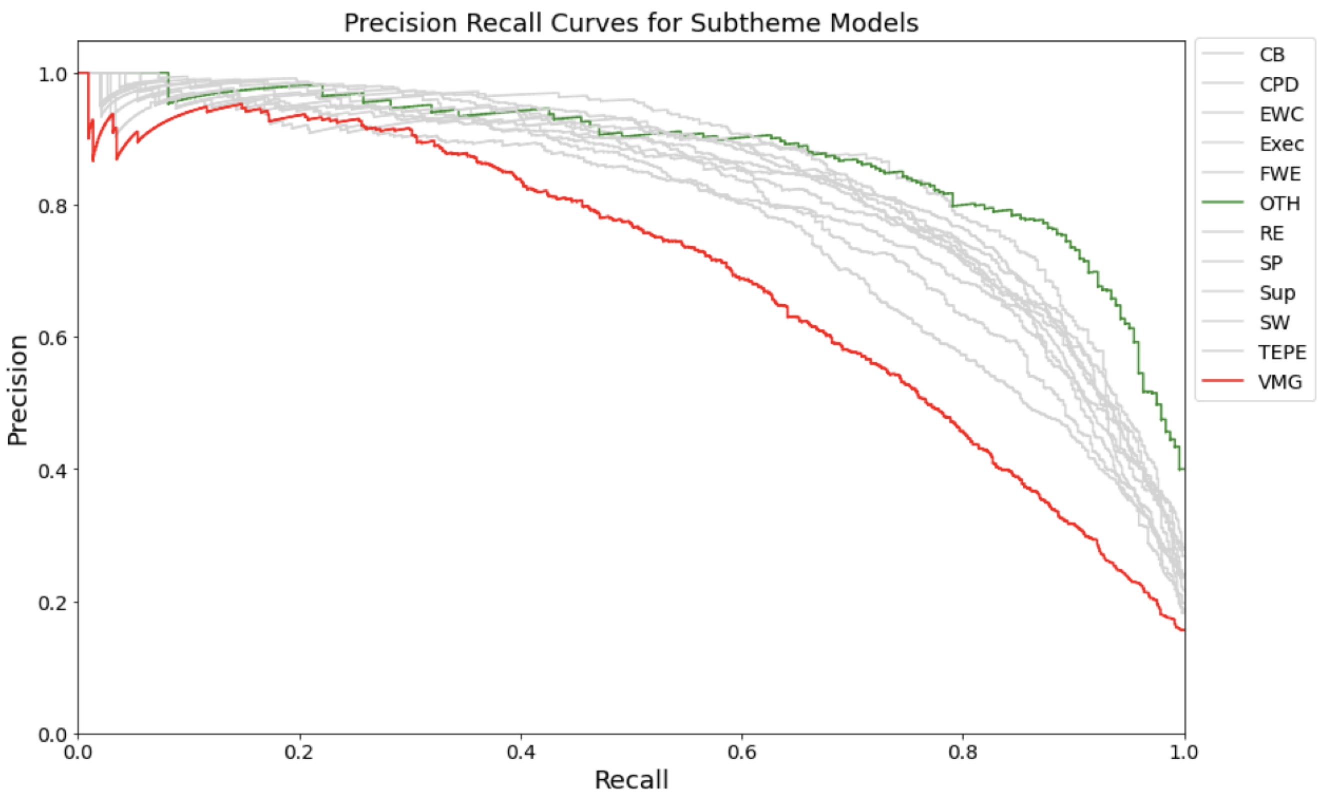
Labelling Subthemes



Subthemes are predicted based on the theme(s) our model has assigned to the comment.

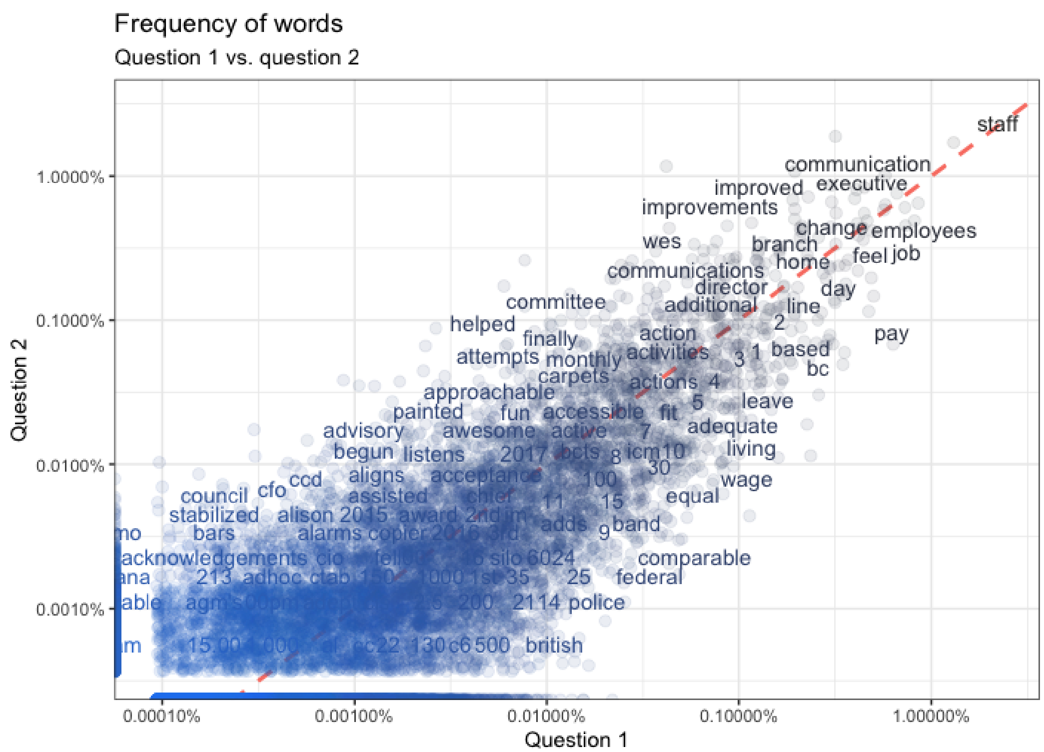
17/23

Precision Recall Plot for Subtheme Models



	label	auc
11	OTH	0.879922
4	FWE	0.860080
5	SP	0.853332
0	CPD	0.843512
9	TEPE	0.841459
8	SW	0.839033
6	RE	0.833848
3	Exec	0.830526
7	Sup	0.797128
2	EWC	0.794497
1	CB	0.777544
10	VMG	0.697208

Question 2: Predicting Themes

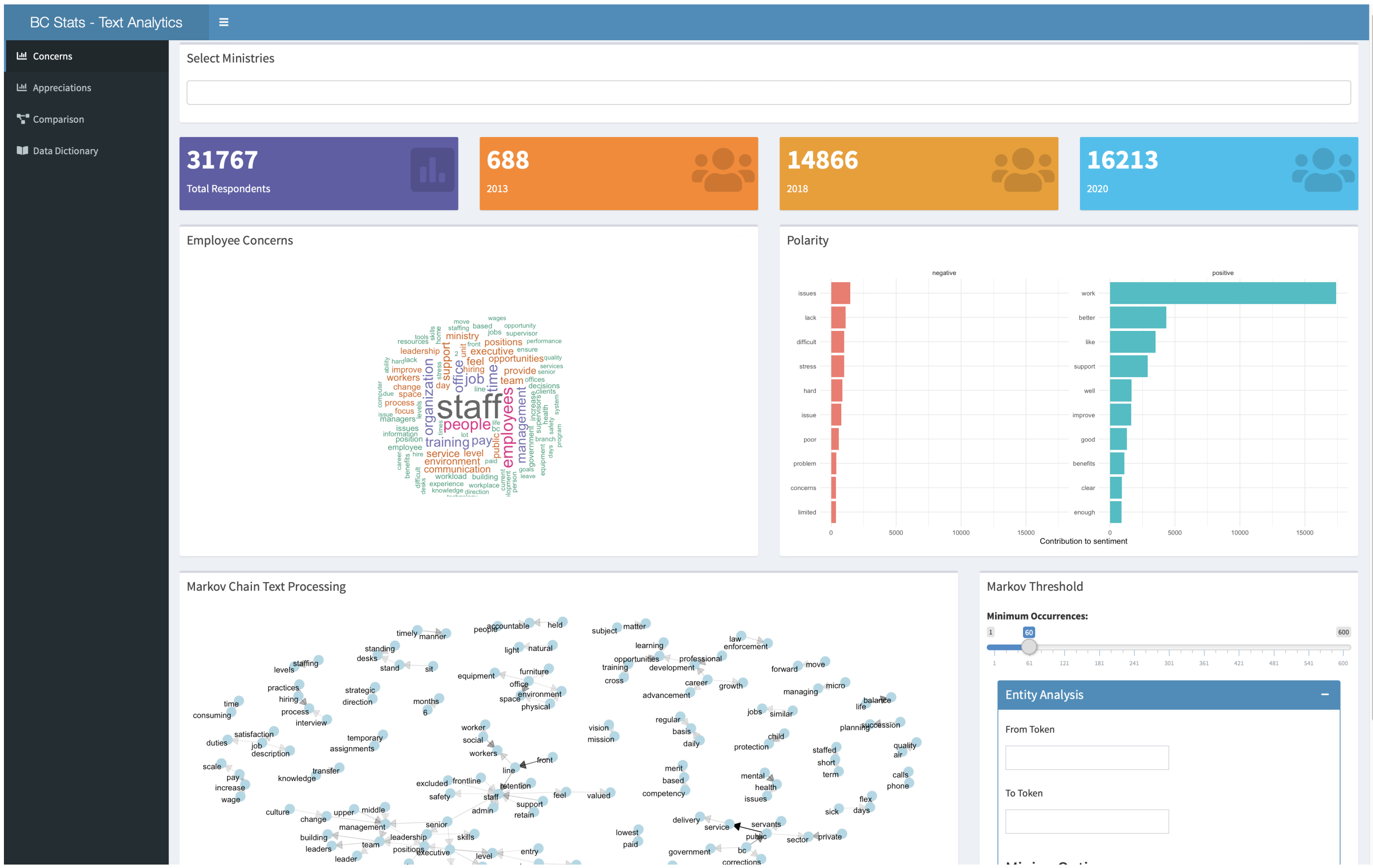


Results using sample data of Question 2 manually encoded by BC Stats (at 0.4 threshold):

Accuracy	Precision	Recall	F1
0.46	0.77	0.63	0.69

19/23

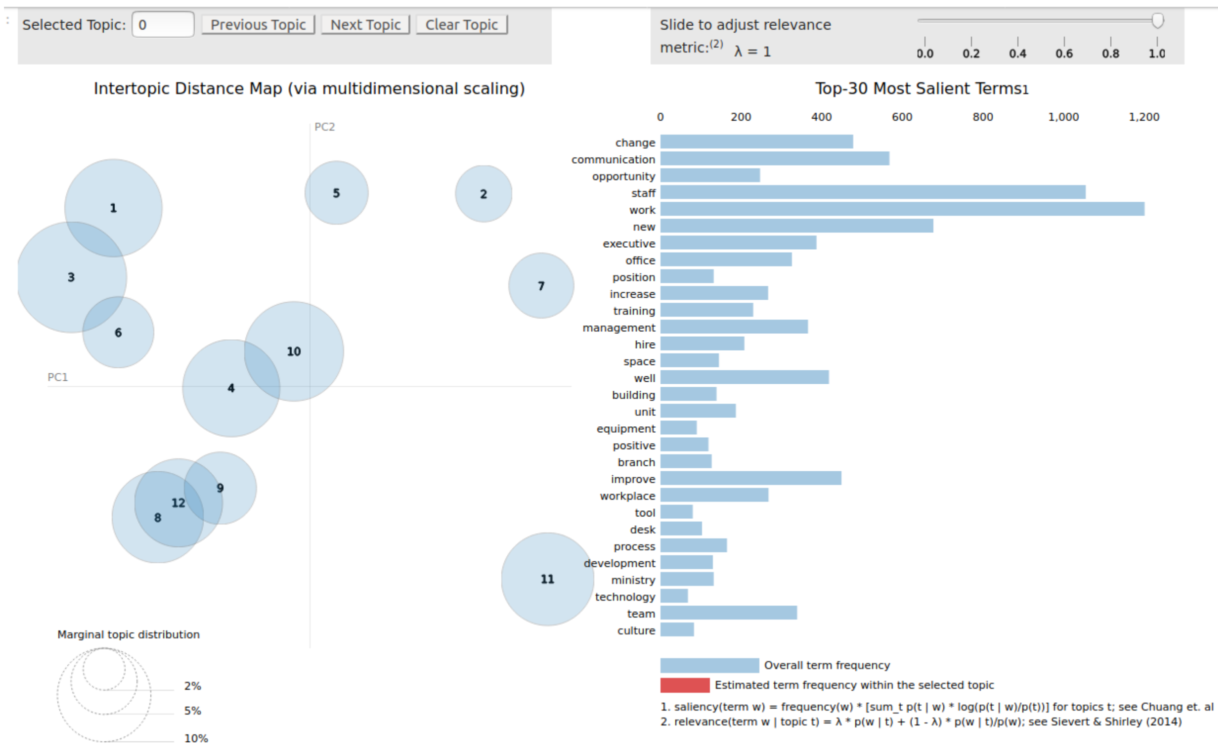
Dashboard



20/23

Methodologies that did not work

- **overfitting** in **CNNs** and **multi-channel CNNs**
- **USE** and **BERT** embeddings
- **Topic modelling** for Question 2 (too much overlap in words, ambiguity)



21/23

Recommendations

- observed better results with more **more data**
- Try **BERT** (could not get embeddings due to sentitive data not being able to upload to cloud platforms)
- using embeddings and padded training & validation data on **public cloud services** (Google Gollab, AWS) which can pave way for applying more complex machine learning algorithms on sensitive data
- Topic modelling for Question 2 can be tried out after removing commonly repeated words

Thank you!