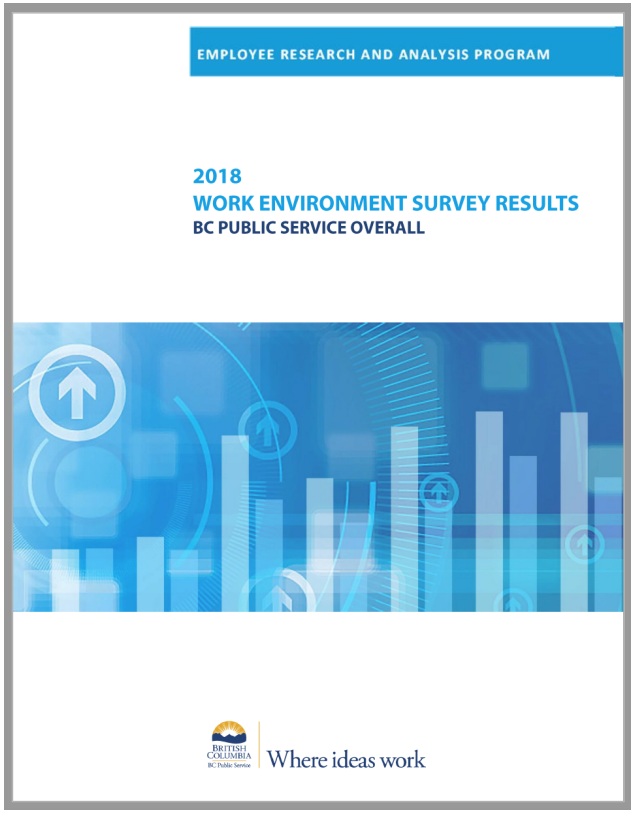
BC Stats Proposal

Text Analytics:
Quantifying the Responses to Open-Ended Survey Questions

Carlina Kim, Karanpal Singh, Sukriti Trehan, Victor Cuspinera
Partner: Nasim Taba | Mentor: Varada Kolhatkar

2020-05-07

Introduction



Work Environment Survey (WES)

- Survey conducted by BC Stats for employees of BC Public Service.
- Measures the health of the work environments.
- 80 multiple choice questions (5 point scale) and 2 open-ended questions.
- 2013, 2015, 2018, and 2020 across 26 Ministries.

Introduction

Open-ended Questions

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 are fake comments as examples of the data.

Objectives

Overarching goal:
Use automated multi-label theme classification of comments to themes and sub-themes.

Question 1

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

- Build a model for predicting label(s) for main themes.
- Build a model for predicting label(s) for sub-themes.
- Scalability: Identify trends across ministries and over the four specified years.

Question 2

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

- Identify labels for theme classification and compare with existing labels.
- Build a model for predicting label(s) for themes.
- Create visualizations for executives to explore the results.

Existing Solution for Question 1

Last year's Capstone

Objective: - To build a model for predicting label(s) for themes

Table 2. Results from the Base Model the Chosen Model

Model	Accuracy	Precision	Recall
Bag of Words | LinearSVC	45%	0.74	0.64
Deep Learning Ensemble	53%	0.83	0.66

Source: BC Stats Capstone 2019-Final Report, by A. Quinton, A. Pearson, F. Nie (https://github.com/aaronquinton/mds-capstone-bcstats/blob/master/reports/BCStats_Final_Report.pdf)

We aim to improve accuracy for predicting label(s) for main themes respective of previous capstone project results.

Getting Familiar with the Data

- Separate Data for each question, and each year.
- Comments with sensitive information.
- Files in XLSX -Excel format-.

Question 1

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

- **Labeled data from 2013, 2018, 2020**, added to around 32,000 respondents.

Question 2

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

- **Labeled data from 2018**, which add around 6,000 respondents.
- **Unlabeled data from 2015 and 2020**, that respresent 9,000 additional comments.

EDA

Question 1

Dataset format
Responses for this question are captured and labeled (theme and sub-theme) by hand:

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

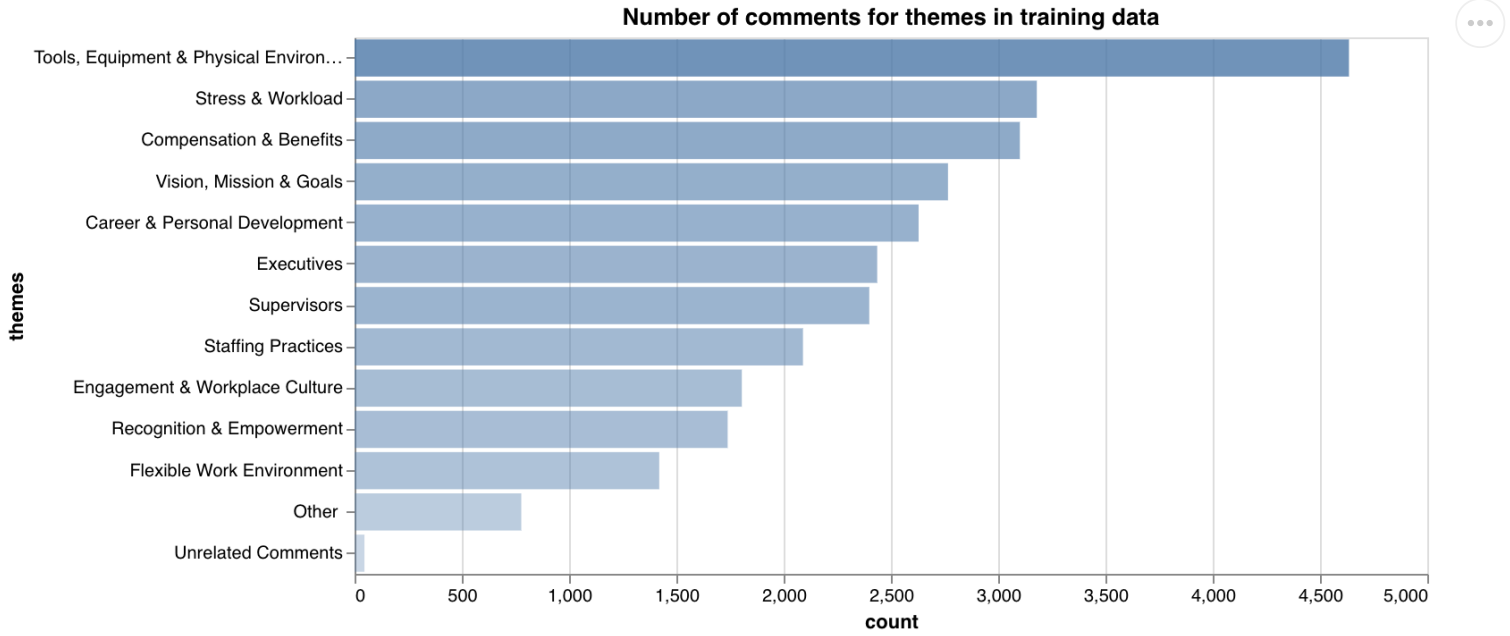
Sub-theme: CB_Improve_benefits = Improve benefits

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

EDA

Question 1

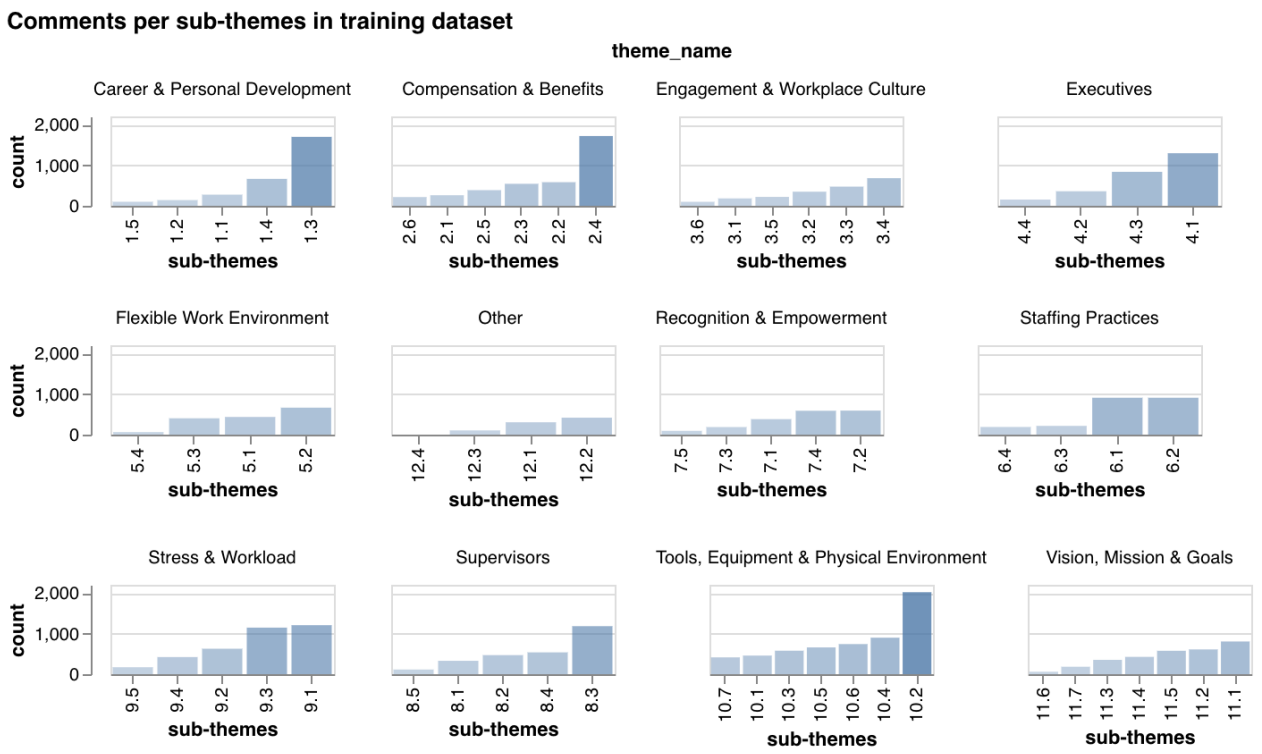
Labels: 13 themes and 63 sub-themes.



Label cardinality for themes: **~1.4**

EDA

Question 1

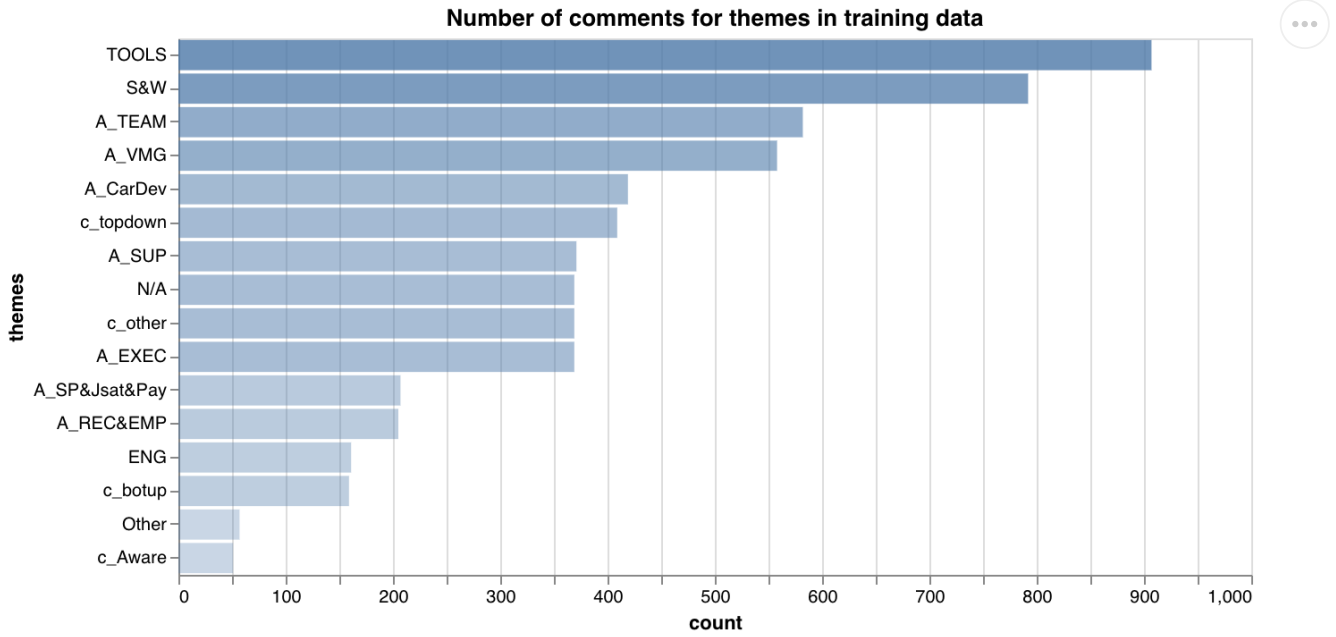


Label cardinality for sub-themes: **~1.6**

EDA

Question 2

Labels for 2018: 6 themes and 16 sub-themes



Label cardinality: **~1.6**

Challenges

- Decide appropriate metric for evaluating accuracy (considering partial correctness) for multi-label prediction problem.

- Low label cardinality indicating sparsity in training data
- ~2 labels per comment from ~60 labels.

- Build a model with increased performance - higher label precision and recall- than the MDS team last year so that it can be deployed by BC Stats.

- Class Imbalance in the data
- skeweness in number of comments per label.

Techniques

Question 1

Binary Relevance - Base Model from last year's Captsone

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Source: Multi-Label Classification: Binary Relevance, by Analytics Vidhya (https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/)

Techniques

Question 1

Classifier Chains - Proposed Base Model

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

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

- Multi-Label Classification using TF-IDF Vectorizer with Classifier Chain.

Source: Multi-Label Classification: Classifier Chains, by Analytics Vidhya (https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/)

Techniques

Question 2

Theme Identifications

- Use clustering algorithms like PCA and Topic Modelling

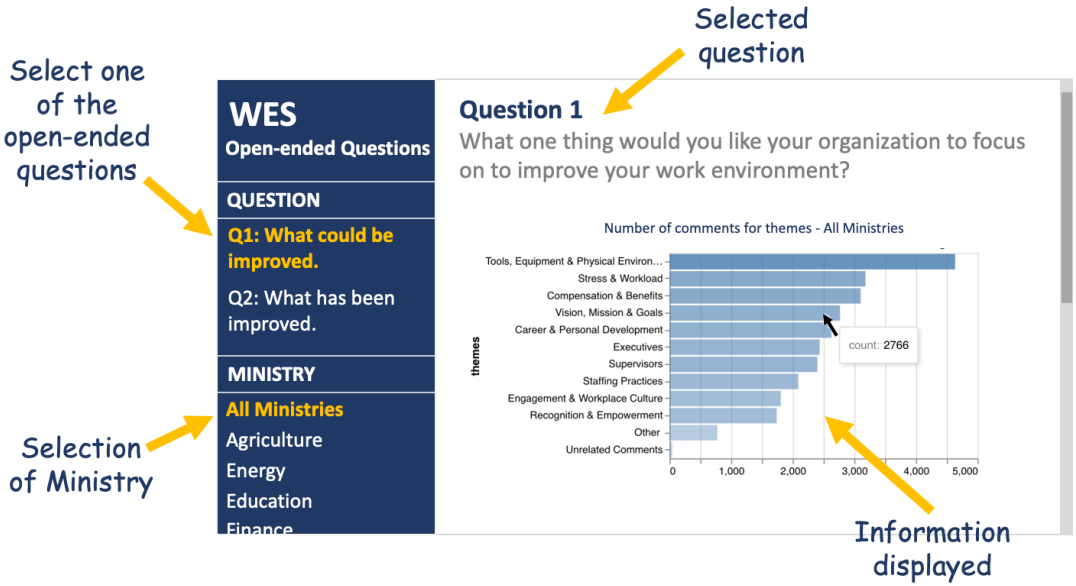
Scalability

- Descriptive Statistics using Matplotlib, Altair and Plotly
- Identify trends over the years
- Identify trends across Ministries

Deliverables

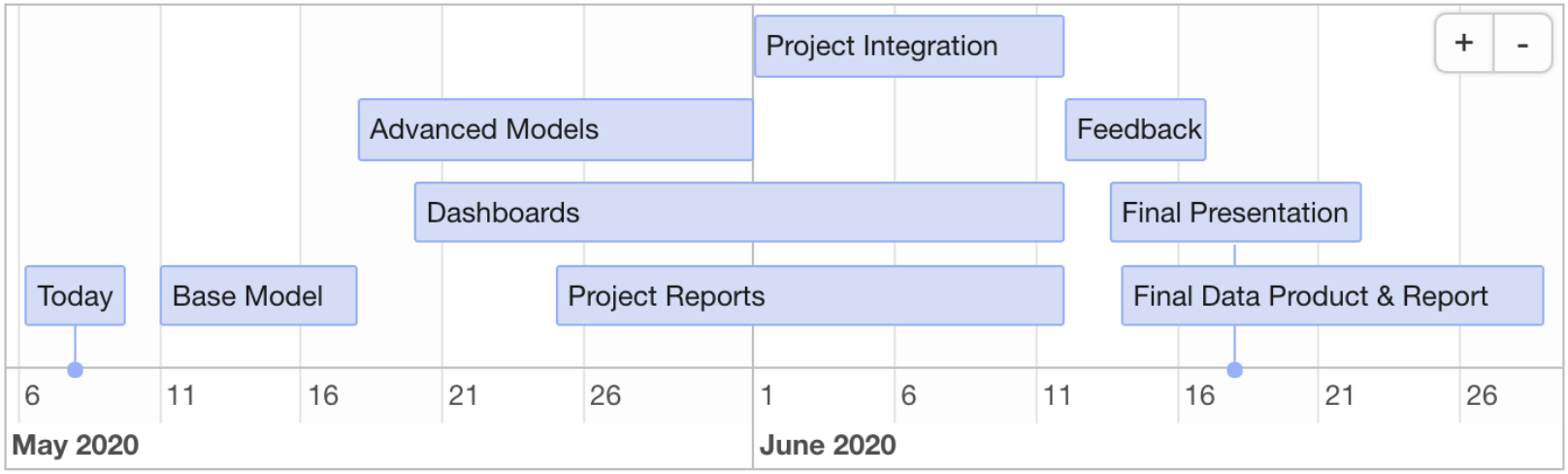
- **Data pipeline with the documentation for our models**

- **Dash app that displays the trends across ministries for both qualitative questions**



Source: Dash app's sketch, based in app developed by BC Stats for the Workforce Profiles Report 2018 (https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/).
Note: This figure is just for illustrative purpose, the final version of the app could differ from the sketch.

Timeline


Project Integration
+
-
Advanced Models
Feedback
Dashboards
Final Presentation
Today
Base Model
Project Reports
Final Data Product & Report
6
11
16
21
26
1
6
11
16
21
26
May 2020
June 2020