

BC Stats Proposal

Text Analytics:

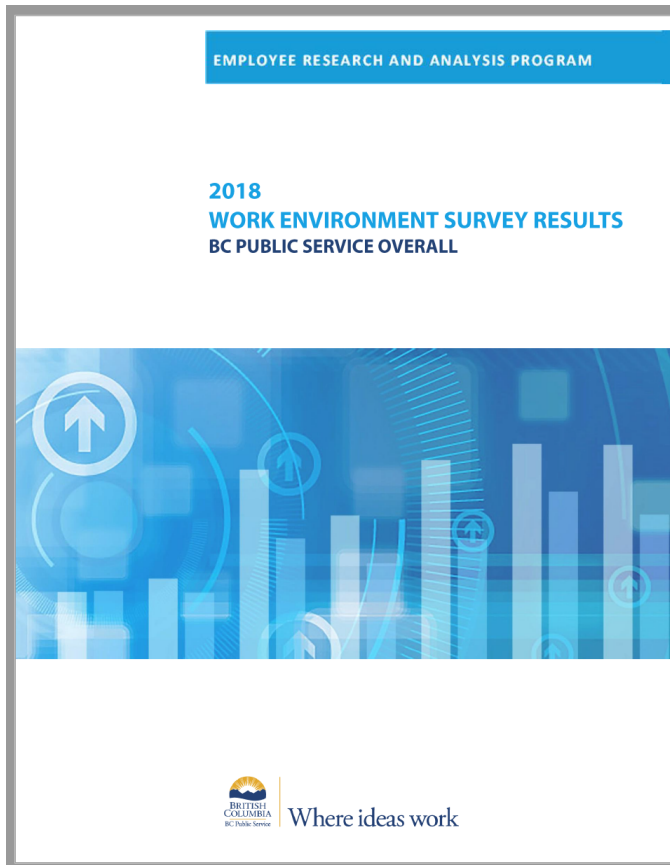
Quantifying the Responses to Open-Ended Survey Questions

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?

Question 2

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

Objectives

Overarching goal:

Use automated multi-label theme classification of comments to themes and subthemes.

Question 1

- Improve accuracy for predicting label(s) for main themes respective of previous capstone project results.
- Build a model for predicting label(s) for sub-themes.
- Scalability: Identify trends across ministries and over the four specified years.

Question 2

- Identify labels for theme classification and compare with existing labels.
- Create visualizations for executives to explore the results.

EDA

Question 1

Data

Labeled data from 2013, 2018, 2020. Added to around 32,000 respondents.

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

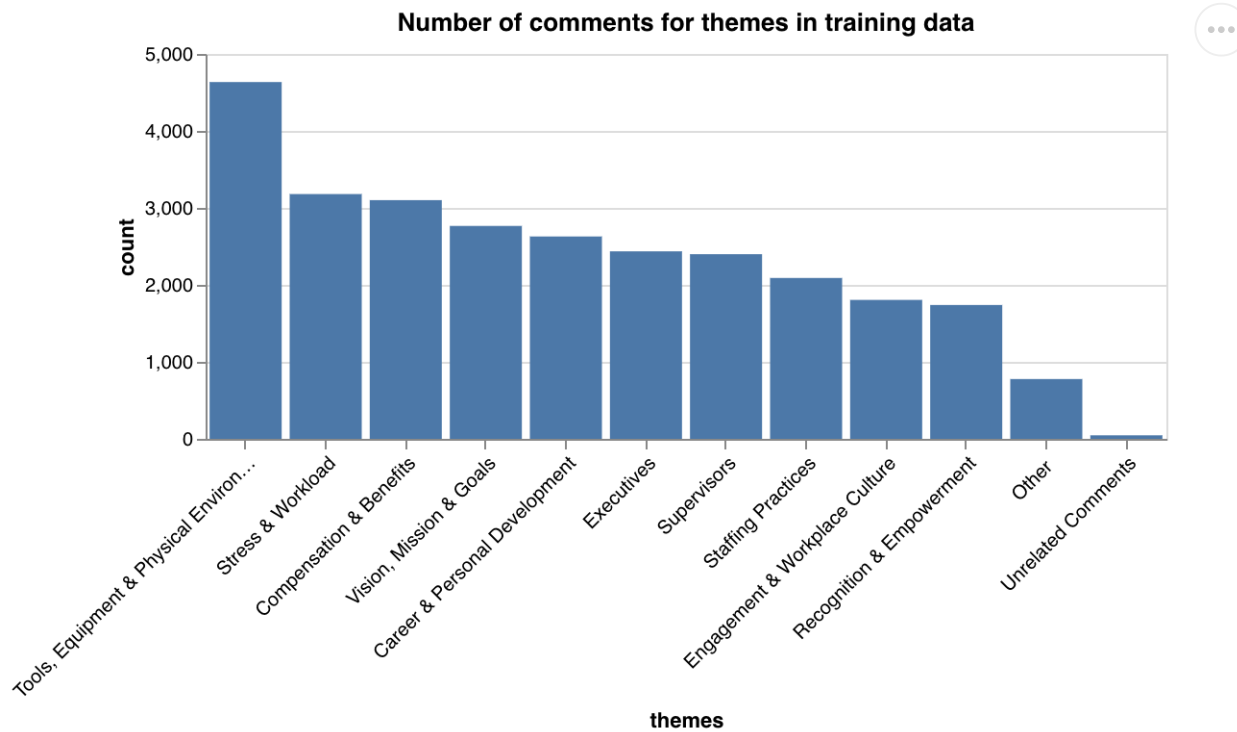
Sub-theme: CB_Improve_benefits = Improve benefits

*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



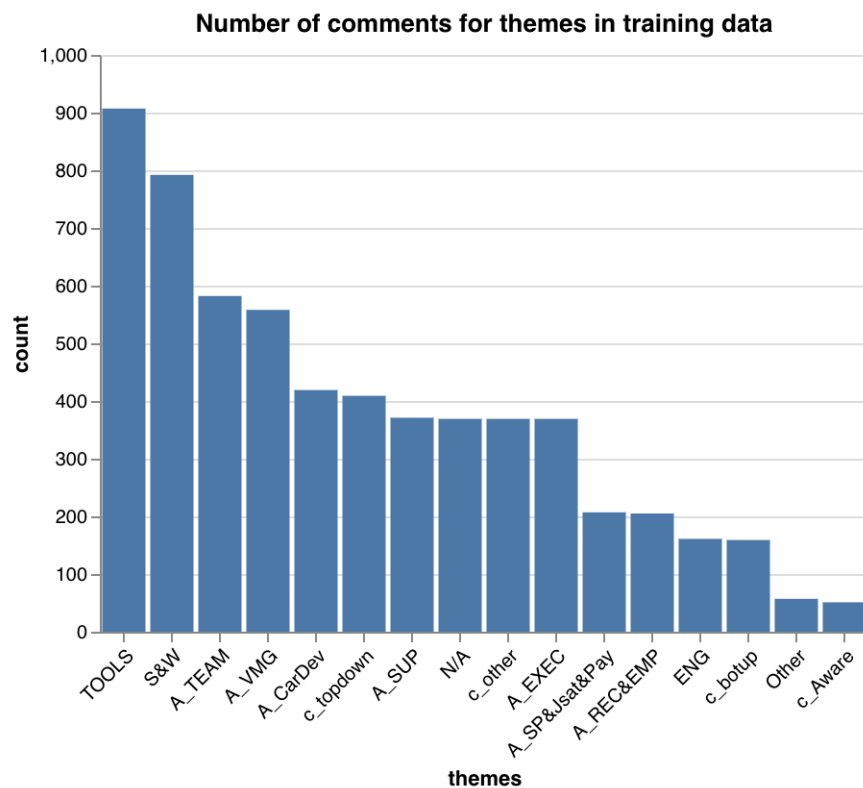
Label cardinality for sub-themes: ~1.6

EDA

Question 2

Data

- Labeled data from 2018 (around 6,000 respondents).
- Unlabeled data from 2015 and 2020 (9,000 additional comments).



Labels for 2018: 6 themes and 16 sub-themes

Label cardinality: ~1.6

Challenges

- Achieve desired accuracy with Multi-label classification model having high number of labels.
- Class Imbalance in the data
 - skeweness in number of comments per label.
- Low label cardinality indicating sparsity in training data
 - ~2 labels per comment from ~60 labels.

Techniques

Question 1

Binary Relevance - Base Model from last year's Captstone

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

\mathbf{X}	Y_1	\mathbf{X}	Y_2	\mathbf{X}	Y_3	\mathbf{X}	Y_4
$\mathbf{x}^{(1)}$	0	$\mathbf{x}^{(1)}$	1	$\mathbf{x}^{(1)}$	1	$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1	$\mathbf{x}^{(2)}$	0	$\mathbf{x}^{(2)}$	0	$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	1	$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1	$\mathbf{x}^{(4)}$	0	$\mathbf{x}^{(4)}$	0	$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	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 the qualitative questions

Timeline

