BC Stats Proposal

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

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

2020-05-07

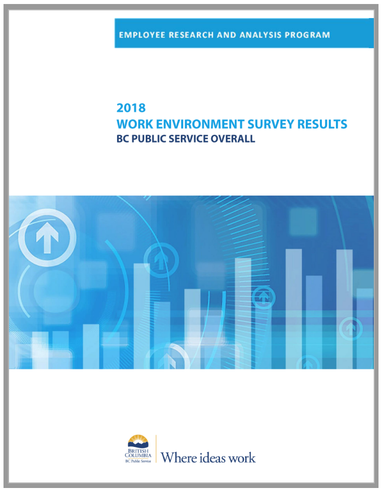
Executive Summary

A brief and high level summary of the project proposal.
We should write this section after we finish the proposal.

Introduction

*"I believe a well performing government, one that meets
the service expectations of British Columbians, can only
be achieved through a strong, highly competent and
committed public service."* - Wayne Strelioff, Auditor General of British Columbia

Work Environment Survey (WES)



EMPLOYEE RESEARCH AND ANALYSIS PROGRAM

2018
WORK ENVIRONMENT SURVEY RESULTS
BC PUBLIC SERVICE OVERALL

Where ideas work

Since 2006, the BC Public Service has conducted the Work Environment Survey (WES) with the goal of understanding their employees' experience, celebrating their successes, and identifying areas for improvement. The survey consists of ~80 multiple choice questions, in 5-point scale, and two open ended questions:

Question 1. What one thing would you like your organization to focus on to improve your work environment?

Example[1]*: "Better health and social benefits should be provided."*

[1] This is a fake comment as examples of the data.

Question 2. Have you seen any improvements in your work environment and if so, what are the improvements?

Example[2]*: "Now we have more efficient vending machines."*

The responses to the first question have been manually coded into 13 themes and 63 sub-themes, and for the second question it has been manually coded into 6 themes and 16 sub-themes.

Objectives

Our project aims to apply natural language processing and machine learning classification on these open-ended questions to automate the process.

The specific objectives for each question are:

Question 1

- 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

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

It is important to mention that the first question has been addressed by previous Capstone projects of MDS Students. In specific, the BC Stat's Capstone of 2019 (Quinton, Pearson, Nie), built a model that predicts the labels of the main themes, and reached the following results:

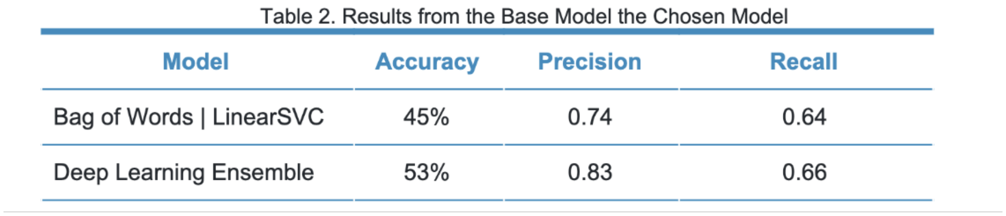


Table 2. Results from the Base Model the Chosen Model

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

Source: Final Report of BC Stats Capstone 2019, by Quinton,Pearson and Nie.

In this case, our aim is to improve the accuracy for predicting labels for main themes respective the results of the 2019 BC Stats Capstone Project.

Getting Familiar with the Data

The Data consist of separated files for each question, and for each of the years (2013, 2015, 2018, 2020), in Microsoft Excel format (.xlsx), and contain sensitive information from employees of BC Public Services.

In specific, the information that we would use for this project for the first question corresponds to the labeled data from 2013, 2018, 2020, that added to around 32,000 respondents.

In the following we can see an example[3] of how this question is presented in the database.

[2] Idem.

[3] This is a fake comment as an example of the data.

Comments[4]	CPD	CB	EWC	...	CB_Improve_benefits	CB_Increase_s
Better health and social benefits should be provided	0	1	0	...	1	0

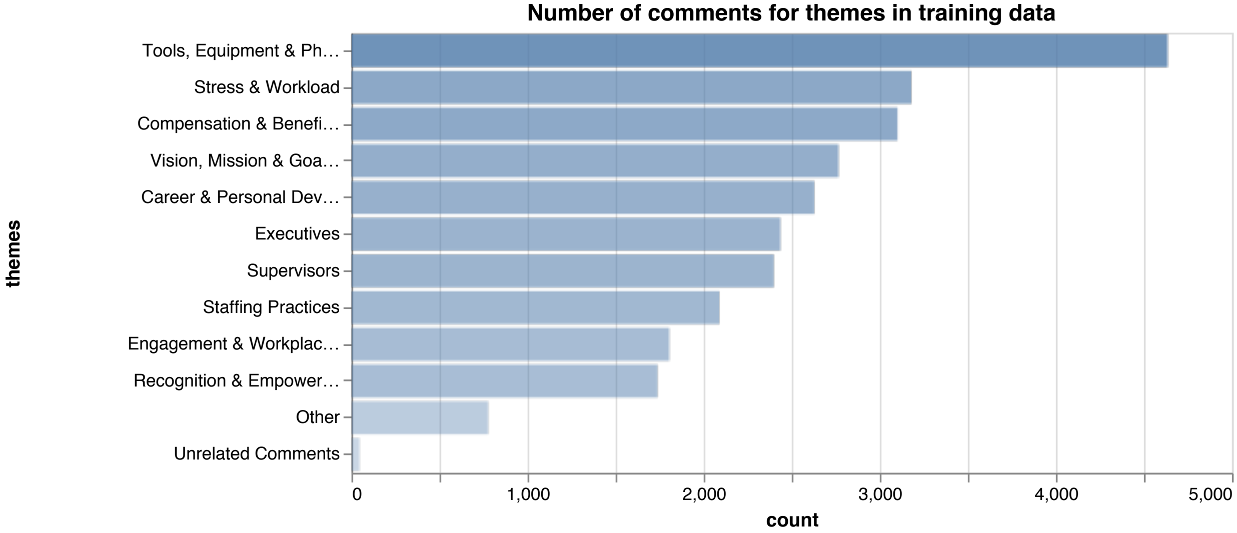
The classification of the previous comment is CB (Compensation and Benefits) for theme, and CB_Improve_benefits (Improve benefits) as sub-theme.

For the second question, we have labeled data from 2018, which add around 6,000 respondents. Also, we have unlabeled data from 2015 and 2020, that represent 9,000 additional comments.

Exploratory Data Analysis

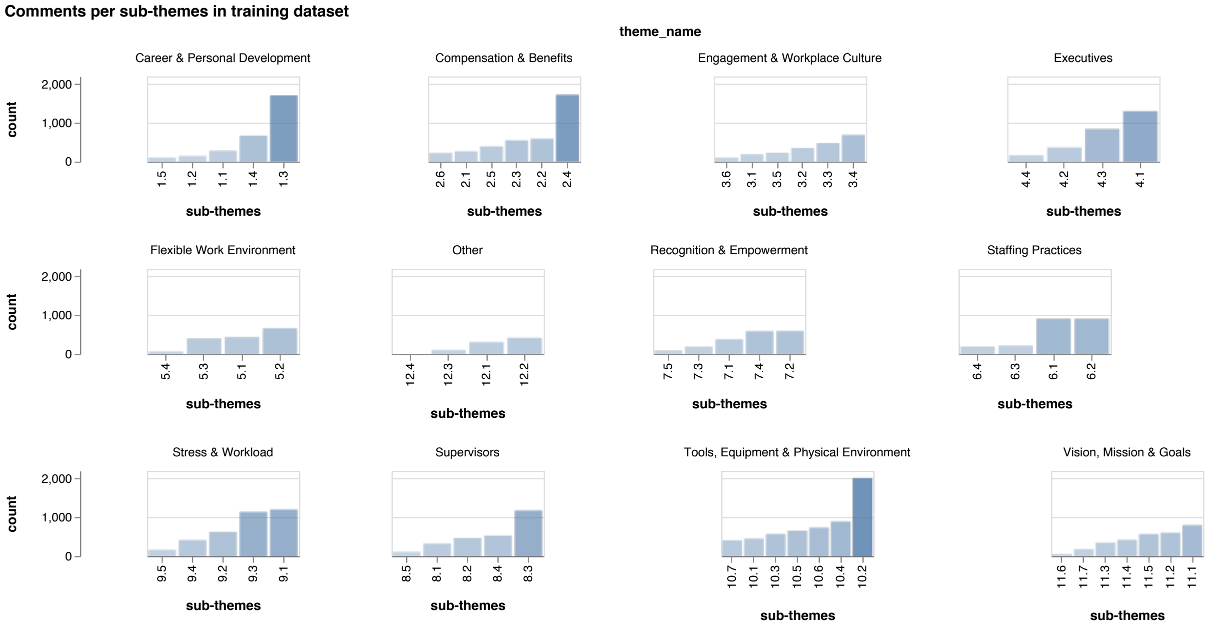
Question 1. What one thing would you like your organization to focus on to improve your work environment?

Labels: 13 themes and 63 sub-themes.



Label cardinality for **themes**: ~**1.4**

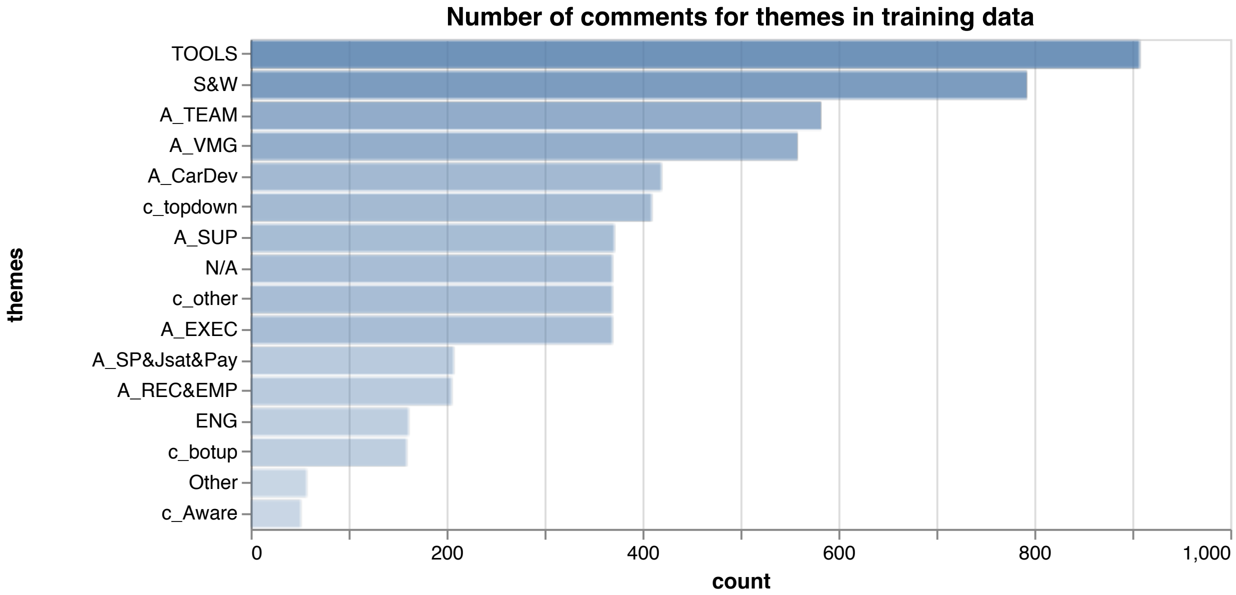
[4] This is a fake comment as an example of the data



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

Question 2. Have you seen any improvements in your work environment and if so, what are the improvements?

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.

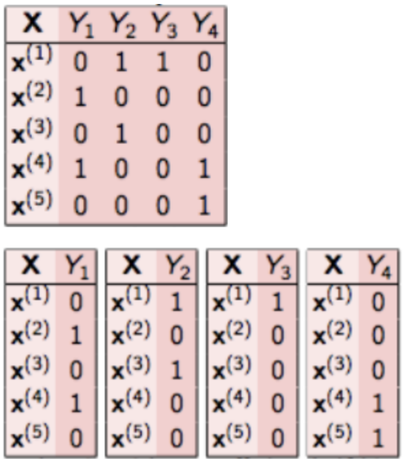
Data Science Techniques

Expectatives of this section: Describe how you will use data science techniques in the project. Be sure to discuss the appropriateness of the data for the proposed data science techniques, as well as difficulties the data might pose. It is recommended to include a description of the data (variables/features and observational units) and some examples/snippets of what the data looks like (as a table or a visualization). Be sure to always always start with simple data science techniques to obtain a simple version of your data science product. There are two benefits to this approach. First, the simple method gives you a baseline to which you can compare future results. Second, the simple method may solve the problem, in which case you don't need something more complicated. For example, your first model should not be an LSTM.

What we had previously in Google Docs Our first question includes labeled data from 2013, 2018, 2020, while the second question has labeled data from 2015, 2018, 2020. Each survey has around ___ respondents across ___ ministries. Sparsity of the data (cardinality) Multi-label problem Create fake comment and labels (we can use real labels) (insert image of class imbalance figure) Initial observations we notice are patterns of class imbalance with the themes and subthemes that may pose an issue with recall and precision in the future. We also observe a low level of label cardinality (average number of labels per comment) so we are dealing with sparsity in our dataset. For automated classification to themes and subthemes tasks, our baseline approach will be to run TF-IDF vectorizer and a Classifier Chains model. The past capstone group used Binary Relevance for their multi-label classification model which treats each label as a separate single class classification problem. In Classifier Chains, the model forms chains in order to preserve label correlation and believe this would be a better choice. For theme identification, our baseline is to use a standard LDA approach. For the dashboard, we will be using Matplotlib, Altair and Plotly. The visualizations are focused on: Identifying trends across the years Identifying trends across ministries

Question 1

Binary Relevance - Base Model from last year's Captsone



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

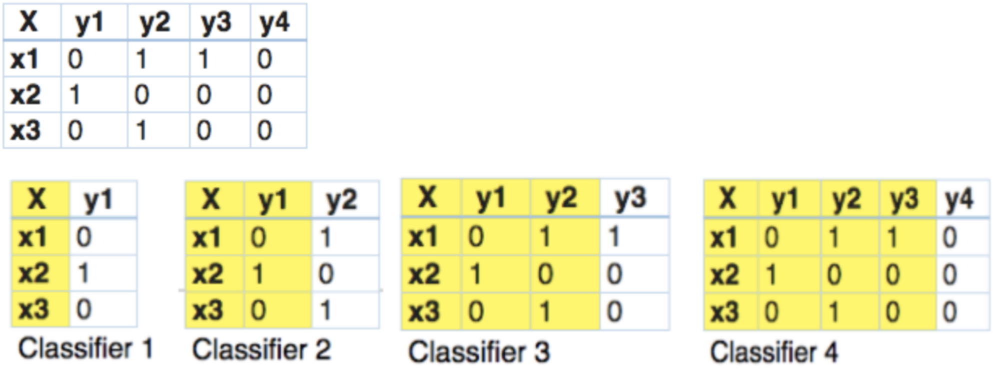
$\mathbf{X}$	Y_2
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

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

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

Source: Multi-Label Classification: Binary Relevance, by Analytics Vidhya

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

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

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

Question 2

Theme Identifications

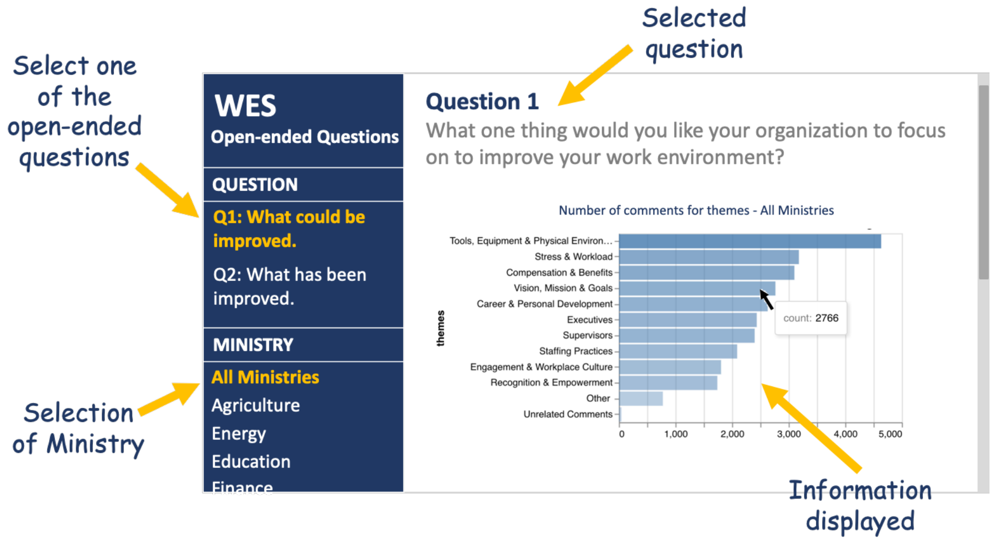
- Use clustering algorithms like PCA and Topic Modelling

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**



Selected question

Select one of the open-ended questions

Selection of Ministry

Information displayed

WES

Open-ended Questions

QUESTION

Q1: What could be improved.

Q2: What has been improved.

MINISTRY

All Ministries

Agriculture

Energy

Education

Finance

Question 1

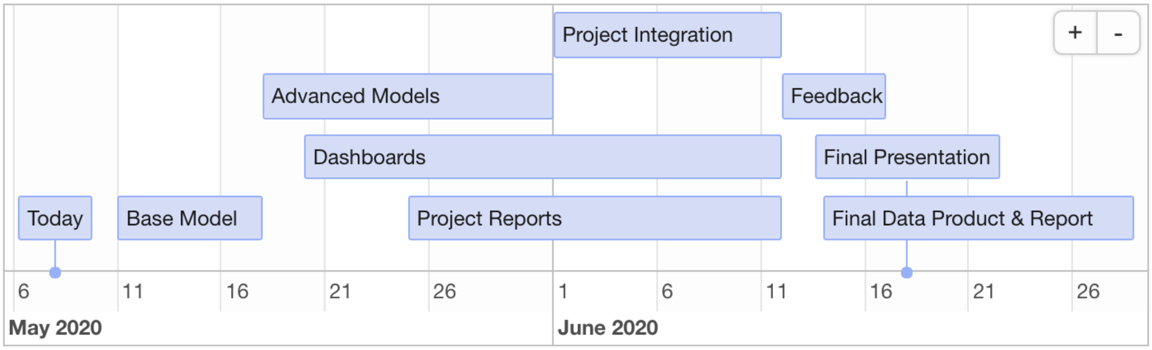
What one thing would you like your organization to focus on to improve your work environment?

Number of comments for themes - All Ministries
themes
Tools, Equipment & Physical Environ...
Stress & Workload
Compensation & Benefits
Vision, Mission & Goals
Career & Personal Development
Executives
Supervisors
Staffing Practices
Engagement & Workplace Culture
Recognition & Empowerment
Other
Unrelated Comments
count: 2766
0
1,000
2,000
3,000
4,000
5,000

Source: Dash app's sketch[5]*, based in app developed by BC Stats for the Workforce Profiles Report 2018.*

[5] This figure is just for illustrative purpose, the final version of the app could differ from the sketch.

Timeline



Week 1 (May 11-15) - Base Models: classification and topic modelling.
Week 2 (May 18-22) - Begin working on advanced models & visualizations for dashboard.
Week 3 (May 25-29) - Continue working with advanced models & start project reports.
Week 4 (Jun 1-5) - Deliverables & Pipelines continuous integration.
Week 5 (Jun 8-12) - Report Writing & Documentation.
Week 6 (Jun 15-16) - Feedbacks & Submissions.

References

- BC Stats. (August 2018). 2018 Work Environment Survey Driver Guide. Site: https://www2.gov.bc.ca/assets/gov/data/statistics/government/wes/wes2018_driver_guide.pdf
- BC Stats. (2018). Workforce Profile Report 2018. Online dashboard. Retrieved 2020-05-08, site https://securesurveys.gov.bc.ca/ERAP/workforce-profiles
- Province of British Columbia. (2020). About the Work Environment Survey (WES). Retrieved 2020-05-09, site https://www2.gov.bc.ca/gov/content/data/statistics/government/employee-research/wes/
- Quinton, A., Pearson, A., Nie, F. (2019). BC Stats Capstone Final Report, Quantifying the Responses to Open-Ended Survey Questions. GitHub account of Aaron Quinton. Site: https://github.com/aaronquinton/mds-capstone-bcstats/blob/master/reports/BCStats_Final_Report.pdf
- Jain, S. (2017). Solving Multi-Label Classification problems (Case studies included). Analytics Vidhya. Retrieved 2020-05-05, site https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/