BC Stats Capstone Proposal Report

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

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

2020-05-15

Executive Summary

BC-Stats conducts the Work Environment Survey (WES) on BC Public Service's ministries with the goal of identifying areas for improvement and understanding employee's experiences in the working environment. Currently, the comments to the open-ended questions have been manually encoded into themes and sub-themes. Given a large number of employees across their 26 ministries, hand-labelling comments is expensive and time-consuming. We propose using natural language processing and machine learning classification techniques to automate the labelling of text responses with the goals of improving previously worked on models and gather insight on trends across ministries and given years. These models may be used on their own or to assist human annotators to speed up the labelling process.

Introduction

Since 2006, the BC Public Service has conducted the Work Environment Survey (WES) across their 26 ministries with the goal of understanding their employees' experiences, celebrating their successes, and identifying areas for improvement. The survey consists of ~80 multiple choice questions using a 5 Likert scale and two open-ended questions.

Currently, the BC Stats team has been manually encoding the comments to the open-ended questions into multiple themes and subthemes. This task can be time-consuming and expensive to do manually. Further, BC Stats has not yet looked into exploring trends across the comments for the second question. We have broken down our specific objectives for each question as follows:

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

- To build a model for predicting label(s) for main themes and sub-themes.
- To identify trends across ministries and over years.

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

- To identify labels for theme classification.
  - We will be comparing this with existing labels as current labels are not sufficiently reliable to cover the full scope of comments.
- To build a model for predicting label(s) for themes.

Further, for both questions, we aim to create visualizations for executives to explore trends across the working environments, common word associations, as well as sentimental analysis.

Data Science Techniques

In 2019, the UBC-MDS BC-Stat's Capstone team has addressed the first objective of Question 1 and reached the following results:

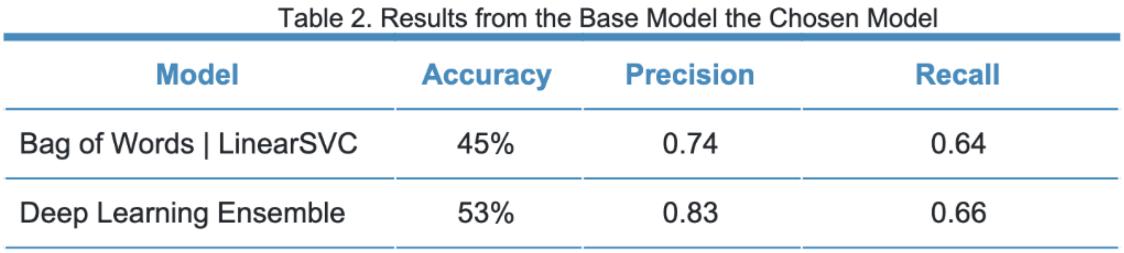


Table 2. Results from the Base Model the Chosen Model

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

Figure 1: Results of Base Model from Final Report of BC Stats Capstone 2019 (Quinton, Pearson and Nie).

These precision and recall results did not meet BC-Stat's required standards for a deployable model and thus, our aim is to improve these results for predicting labels for the main themes for Question 1.

We have around 32,000 labelled observations for the first question and around 6,000 labelled observations for the second question which can be used for addressing our label prediction objectives[1]. Further, we have an additional 9,000 unlabelled comments for the second question.

Question 1:

The BC-Stats team has manually encoded the comments into 13 themes and 63 sub-themes. As each comment may be labelled with more than one label, using multi-label classification algorithms seem like the right choice to create an automated classifier.

The 2019 UBC-MDS Capstone team had proposed Binary Relevance as a base model with a bag-of-words representation. Binary Relevance predicts the labels independently for n-labels in the target variable and then combines them. One drawback of this classifier is that it does not account for label correlations.

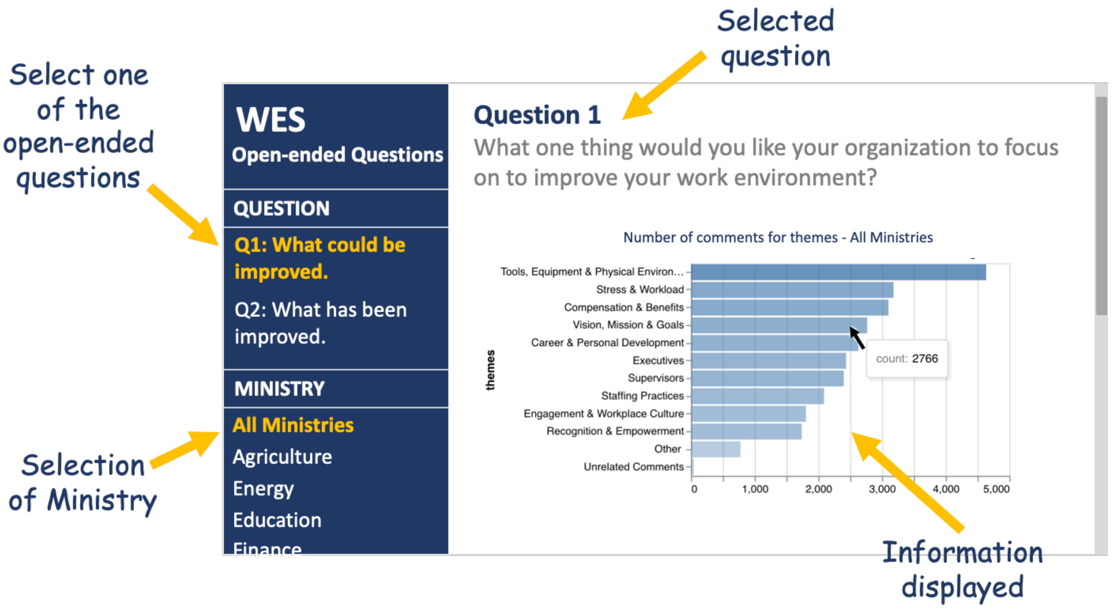
This year, we propose using TF-IDF representation, instead of bag-of-words, along with Classifier Chains as our base model. TF-IDF gives higher weights to important tokens and Classifier Chains preserves the order and occurrence of the labels.

Question 2:

The BC Stats team has encoded the comments for this question into 16 sub-themes but as mentioned, the labels are not sufficiently reliable. Our first approach will be to identify labels using unsupervised learning and manually compare them with the existing labels. Clustering techniques like PCA and Topic Modelling will be implemented for this task. Once the labels are reliable and correctly identified, a multi-label classifier can be trained on this data for automated labelling of comments. These themes may or may not be similar to Question 1.

We expect the final data products to be a data pipeline with the documentation for our models, and a dashboard app that displays the trends across ministries for both qualitative questions.

[1] Labelled data availability: 2013, 2018, 2020 for question 1, and 2018 for question 2.



Select one of the open-ended questions

Selected question

Selection of Ministry

Information displayed

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
0
1,000
2,000
3,000
4,000
5,000
count: 2766

Figure 1: Potential application sketch[2]

Timeline and Evaluation

In the following table, we shared the estimated outline for the Capstone project.

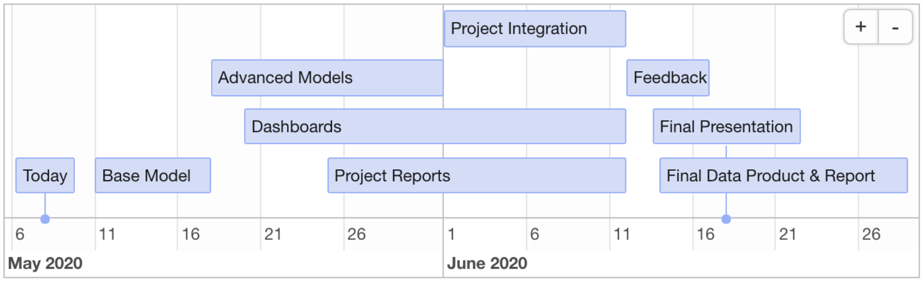


Figure 2: Capstone Project Timeline.

Additionally, we have scheduled weekly meetings on Tuesdays with BC Stats, and Thursdays with our MDS-mentor.

References

- BC Stats. (August 2018). 2018 Work Environment Survey Driver Guide.
- BC Stats. (2018). Workforce Profile Report 2018. Online dashboard. Retrieved 2020-05-08
- Luaces, O., Díez, J., Barranquero, J. et al. (2012). Binary relevance efficacy for multilabel classification. Prog Artif Intell 1, 303–313
- Province of British Columbia. (2020). About the Work Environment Survey (WES). Retrieved 2020-05-09

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

- 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.
- Read J., Pfahringer B., Holmes G., Frank E. (2009) Classifier Chains for Multi-label Classification.
- Wikipedia. (2020, May 3). Tf–idf. In Wikipedia, The Free Encyclopedia. Retrieved 2020-05-15.