BC Stats Capstone Final Report | Quantifying the Responses to Open-Ended Survey Questions

Authors: Aaron Quinton, Ayla Pearson, Fan Nie

Executive summary

The BC Public Service conducts a Work Environment Survey (WES) with the goal of understanding their employees' experience, celebrating their successes, and identifying areas for improvement. We have leveraged data science techniques - particularly in the domain of natural language processing - to automate the labeling of the text responses and to better capture insights from the survey recipients.

Introduction

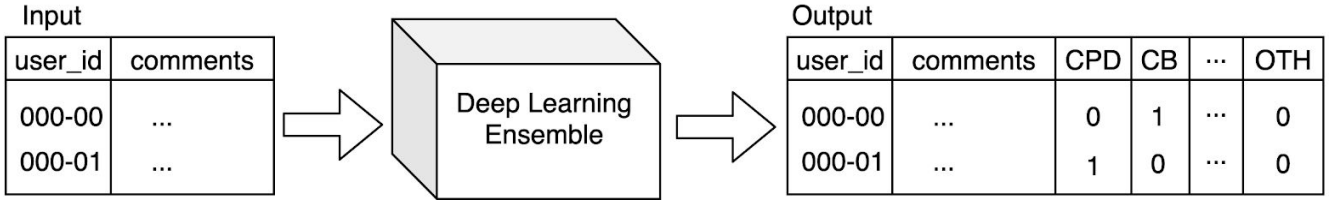
The BC Public Service is committed to understanding the challenges and successes within the workplace. One of the ways this is quantified is through the WES which measures key drivers through qualitative data from the open-ended survey responses and quantitative data from the multiple-choice questions. The survey results in 2018 include over 22,000 respondents. The open ended survey response answers the question:

"What one thing can your organization do to improve your work environment?"

Currently, employee responses to this question are manually coded into 12 themes and 68 sub-themes (see Appendix A for specifics), which is a time-consuming and error-prone process. We propose to use automated classification of themes by training supervised multi-label machine learning classifiers. The current WES analysis also investigates the qualitative and quantitative datasets separately but they have not been investigated to see their similarities and differences. Another point of opportunity is that the WES analysis does not attempt to quantify the underlying emotions within the comments. Given these points, our main objectives are to automate the qualitative labeling and to gain new insights about the open-ended survey responses. These objectives can be broken into three research questions, which are detailed below alongside the proposed solution.

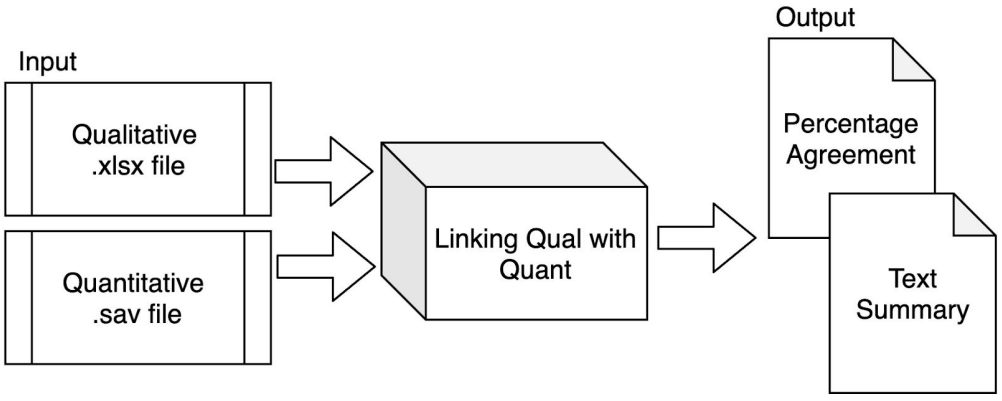
1. **Multi-Label Theme Classification** - Automate the theme classification for the survey responses.

Figure 1. Proposed Classification Solution



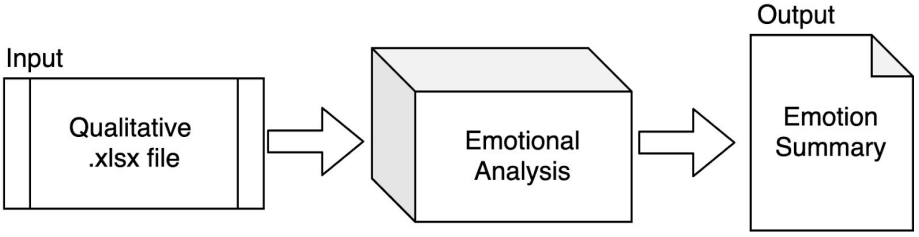
2. **Linking Qualitative to Quantitative** – Does the sentiment of the qualitative responses match the sentiment of the quantitative responses?

Figure 2. Proposed Linking Solution



3. **Emotion Analysis** – How much anger, fear and sadness are present in the qualitative responses?

Figure 3. Proposed Emotion Analysis Solution



Data Science Methods

Multi-Label Theme Classification

We used supervised machine learning methods for multi-label theme classification, as a comment can be labeled with upto 5 themes. In particular, we trained traditional feature-based classifiers with bag-of-words as well as deep-learning classifiers with pre-trained embeddings [1] for multi-label classification.

Bag of Words | LinearSVC

The Bag of Words Model converts the comment data into a sparse matrix indicating the count of words that are in each comment. With this matrix we applied a linear support vector classifier. The advantages of this approach is the simplicity as well as the efficiency in computation. However, it is unable to capture the intricacies and deeper meaning in the text data which limits its ability to achieve a high classification accuracy.

Pre-trained Embeddings | Deep Learning Ensemble

To capture the context of a word as well as it's syntactic and semantic meaning, a more involved method is required. Here enters word embeddings. Word embeddings allow us to represent words as a vector of numbers which have been shown to be useful features for classification [2]. We experimented with several pre-trained embeddings and achieved the best accuracy utilizing the GloVe [3] pre-trained embeddings trained on Wikipedia and common crawl, and the FastText [4] embedding trained on a common crawl. Pre-trained embeddings were a great resource for this project because our data size was limited and it allowed us to leverage more useful vector representations of the words in our comments.

The deep learning ensemble averages predictions of a convolutional neural net model, and three bidirectional GRU models. At the cost of computation time and simplicity, this approach was able to achieve our best accuracies. The architecture for each model is illustrated in Figures 4 and 5. The outputs from the final dense layer represent the prediction probabilities for the 12 themes.

Figure 4. Bidirectional GRU & Conv1d (Trained with GloVe Crawl, GloVe Wiki, and FastText Crawl)

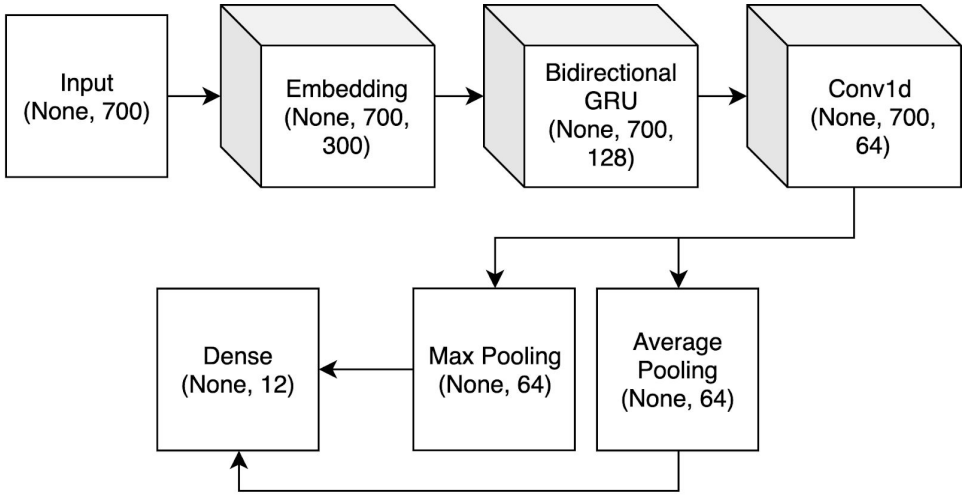
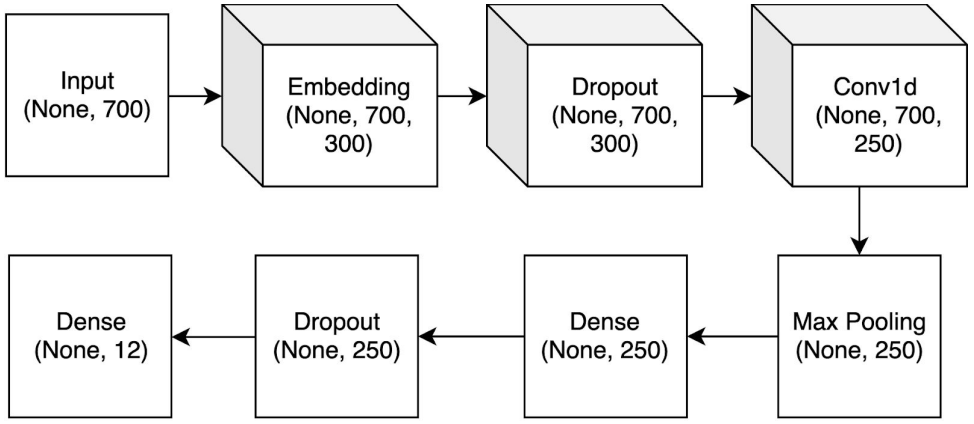


Figure 5. Convolutional Neural Net (Trained with GloVe Wiki)



Linking Qualitative to Quantitative & Text Summarization

We used mixed method research which finds value in integrating qualitative and quantitative datasets. Our aim was to help identify specific areas in the qualitative dataset that are not as well represented in the quantitative dataset. Automated text summarization was used to interpret the results of integrated dataset by reducing the time it would take to manually through the linked comments.

Linking Qualitative to Quantitative

To create a link between the two datasets independently team members coded all of the 80 multiple-choice questions with the sub-theme codes[1]. To validate the linkage between the sub-theme codes and multiple choice question percent agreement was used to understand the inter-rater reliability. Percent agreement was used because there is a low probability of chance agreement due to 68 different sub-themes labels to choose from. The percent agreement between all three raters was 48%, indicating that all three raters gave half the multiple-choice questions the same sub-theme label. Two of the three raters had 80% agreement between them and to determine the correct match when all three didn't agree was the majority, meaning 80% of the links were matched by at least two people and 50% by all three people.

To join the qualitative and quantitative dataset both had to be extensively cleaned and were joined by the multiple-choice sub-theme code linkage that had been created. It was assumed that all the comments had a negative sentiment and were compared to the sentiment of the linked multiple-choice questions. The agreement level of strong, weak or no agreement was calculated for each comment multiple-choice pair based on the relationship shown in table 1. A total of 11,621 unique comments were evaluated.

Table 1. Relationship between the qualitative data (open-ended survey comments) and the quantitative data (multiple-choice responses)

Qualitative Sentiment	Quantitative Sentiment	Agreement Level
Negative	Negative	Strong
Negative	Neutral	Weak
Negative	Positive	No

Text Summarization

To understand the main themes and topics between the different agreement levels a text summarization algorithm was implemented. The advantages of this method is it gives more detailed information and context in comparison to topic modelling.

The first algorithm[2] uses pre-trained word embeddings to get the average sentence embedding, which captures the meaning of the sentence. To determine the similarity of the sentences cosine similarity was calculated and graphed where the nodes are the sentences and the edges are the cosine similarity score. PageRank is applied to the graph and determines the key sentences. The second algorithm is Genism's summarization which uses a variation of TextRank [5].

[1] For more details about the methodology read the linking_methodolody document in the reports section
[2] Our implementation is an adaptation from Prateek Joshi's blog
https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

An improvement to be implemented is creating at least one manual summary and then comparing it to the text summarization algorithms output to determine the quality and tune the length of the summary. This was not done due to time constraints as there are hundreds of comments to read through.

Emotion Analysis

The open-ended survey question is designed to generate responses with a negative sentiment. What is not understood is what emotions motivate the negative sentiment. This analysis looked for the presence of three emotions: anger, fear and sadness. We felt emotion analysis would provide deeper insights then doing sentiment analysis since we were already assuming the comments have a negative sentiment.

To find the emotions in each comment a rule-based matcher was trained using an emotion lexicon[3] which contains words that have a score relating to the intensity of the emotion. The matcher is trained by passing words with specific rules. An example would be to add the rule which contains the word “aggressive” in lower case, then when a document is passed to the matcher if aggressive is present it will produce the locations of the matches. A matcher was trained for anger, fear and sadness and passed the comments. To determine the main emotion present in the comment the scores from each word were added up and the emotion with the highest score was chosen.

One of the issues with using a matcher is it does not understand the context of the words. Joy was removed from the analysis because the words being matched were being used in a negative connotation such as “not better”. An improvement that would create more robust matches would be to manually create custom rules for each word. This improvement could take several days to weeks depending on the level of detail of the rules. Currently it takes a few seconds to create the matcher because all of the rules are the same for every word but when words in the text are not identical to the lexicon like plural they won’t be found.

Data Product and Results

Our data product consists of five components: trained classification models, our codebase to build the models and conduct the analysis, a script to classify new comments, an R pipeline to link the quantitative and qualitative data and summarize the comments, and a report detailing the emotion analysis methodologies.

Multi-Label Theme Classification

To evaluate our multi-label theme classification models we used the following metrics:

Accuracy - the percent of comments that got all 12 labels correct for that comment
Precision - the average proportion of predictions for each theme that are correct
Recall - the average proportion of all the correct labels that were predicted for each theme

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

[3] The emotion lexicon that was used is Saif M. Mohammad NRC Affect Intensity Lexicon v0.5 which contains 6000 words related to anger, fear, sadness and joy http://saifmohammad.com/WebDocs/NRC-AffectIntensity-Lexicon.txt

The predictions of the ensemble can be further adjusted by changing the threshold required to predict. With a probability threshold of 92% the ensemble model is able to predict on 50% of the comments with 67% accuracy and an average precision and recall of 0.96 and 0.69 respectively. The results by theme are tabulated in table 3:

Table 3: Final Model Results by Theme using a 92% Probability Threshold

Metric	CPD	CB	EWC	Exec	FWE	SP	RE	Sup	SW	TEPE	VMG	OTH
Precision	0.96	0.97	1.00	0.95	0.98	0.87	0.85	0.93	0.95	0.97	1.00	1.00
Recall	0.63	0.89	0.37	0.53	0.74	0.59	0.30	0.42	0.57	0.90	0.48	0.13

The results above confirm very high precision on all the themes and average recall depending on the theme. Based on these results, our recommendation for the manual coders is the following:

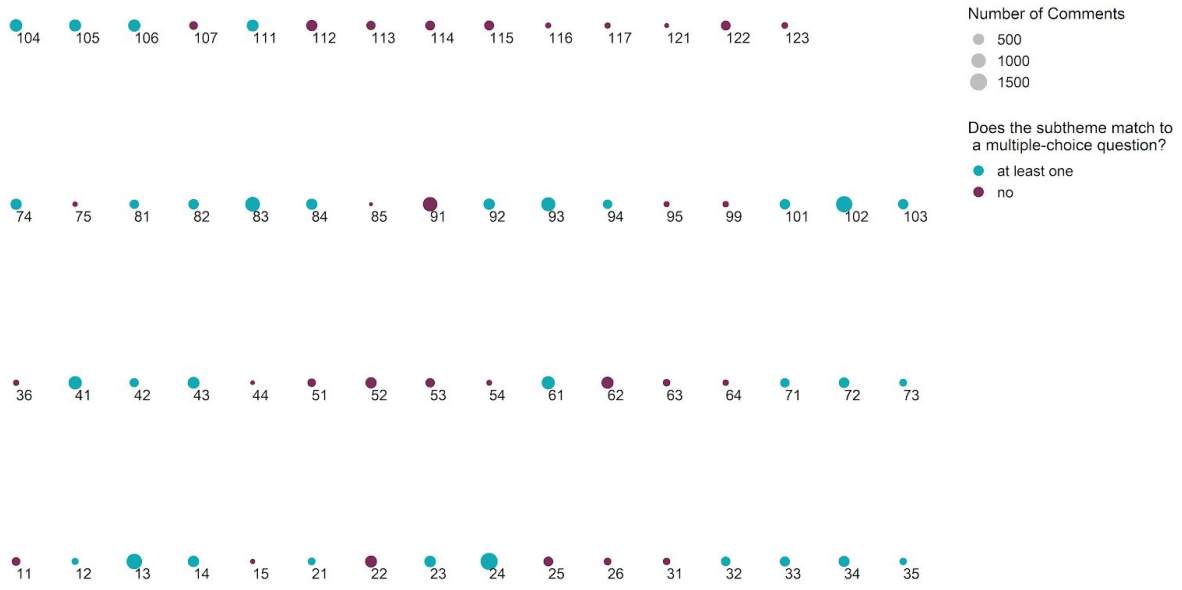
- Filter out all the comments that have no labels (~50%) and put these aside to go through manually
- On the remaining 50% of the comments the coder does not need to confirm the labels that have been predicted and can simply filter to the comments for each theme and assign the appropriate sub theme
- Because of the low recall on some of the themes we recommend the user skim through the comments looking for comments relating to Recognition and Empowerment, Engagement and Work Culture, Vision Mission and Goals, and Supervisors. The model was less confident in predicting these themes so they are prone to false negatives.

Linking Quantitative to Qualitative & Text Summarization

The product that links the qualitative to the quantitative dataset is an R pipeline that can reproduce the analysis. The text summarization is a python module that can create summaries for the open-ended survey comments grouped into themes, sub-themes and agreement levels.

Figure 6 illustrates the linkage between the sub-theme codes and multiple-choice questions, it also highlights the number of comments coded to each sub-theme. From figure 6 we can surmise several observations. For example, Sub-theme 91: “stress and workload - hire more staff” has the most comments within the theme but no corresponding question. Similarly, sub-theme 52: “flexible work environment – flexibility in work locations” has a reasonable number of comments and no corresponding question. Our recommendation from these insights would be to capture more numerical info related to these themes by including corresponding questions in the next survey cycle.

Figure 6. Displays the linkages between the sub-theme codes and multiple-choice questions. The circles represent each sub-theme code and if it matches to a multiple-choice question. The purple colour indicates there is no multiple-choice question linked to the sub-theme (see appendix A for the names of the sub-themes).



From looking at the agreement levels between the linked datasets we recommend focusing on the areas with the lowest levels of agreement like the themes of engagement and workplace culture and vision, mission and goals seen in figure 7. Appendix B shows the number for the agreement levels broken down to the subtheme levels and can be referred for more detail.

Figure 7: Agreement Levels for the Linked dataset Grouped by Theme

Emotion Analysis

The emotion analysis will be provided in a python module that can be used to understand the emotions of the qualitative data. The module is flexible so it is easy to analyze themes or sub-themes of concern.

Since there are 12 themes and 68 sub-themes only a few key results will be discussed. Emotionless in this context refers to comments that did not have any anger, fear or sadness present. It was found that fear was the most common emotion present in all of the comments seen in figure 8. Comments related to

supervisors had more anger, fear and sadness and less emotionless comments than most other themes shown in Figure 9.

An improvement to this product could have been an interactive dashboard to view the results, this would have taken a few weeks to create. The benefit of the current product is fast and simple by allowing you to generate only the plots you are interested in.

Figure 8: Overall Emotions in the WES

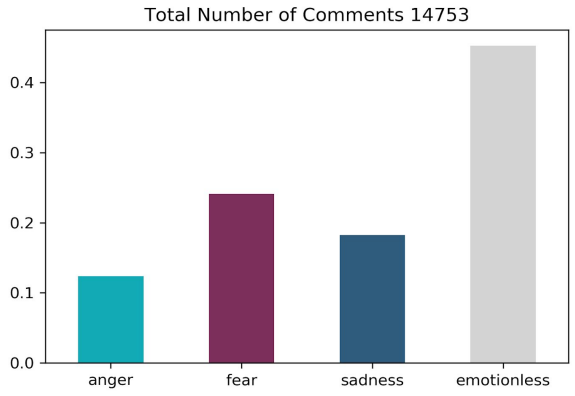
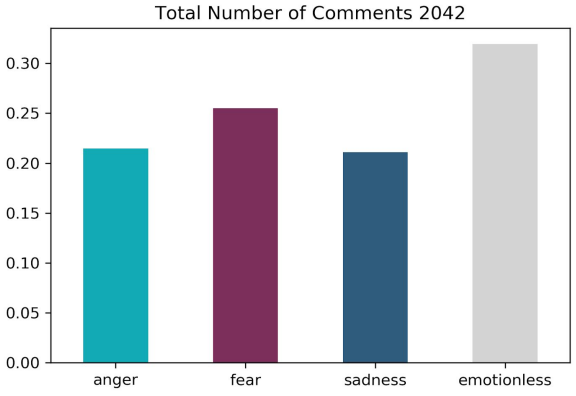


Figure 9: Emotions Related to Supervisors



Conclusions and Recommendations

Classification

The task of automating all 12 themes and 68 sub themes posed quite a challenge. Our solution is not able to remove the manual coders from the process but aims to make the process easier. We are able to confidently predict the main themes on 50% of the data and identify the other 50% for manual labeling.

Our recommendation for future analysis is to extend our model on the sub theme level. This will require more data per sub theme to build a reliable model. Finally we recommend future work to be done utilizing cloud computing technology. With additional computation power you will be able to apply cutting edge natural language processing tools to improve on the existing accuracies. In particular, it is our opinion that BERT will perform very well on this problem.

BERT model

We tried Google BERT which is a state-of-the-art model that was released in October 2018. BERT is known to outperform other models when finely tuned. We ran BERT on the Career and Personal Development theme and it gave promising precision and recall. Unfortunately a single theme took three days to run on a local computer and we estimate it would take 36 days to loop through all of the themes. The time complexity of the model is the main issue. A solution to this issue would be to use cloud computing but the raw comments would need to be uploaded to the cloud server which would not follow privacy policies.

Linking Quantitative to Qualitative & Text Summarization

Over 22,000 employees took the WES in the BC Public Service generating 15,000 written comments categorized into 12 main themes and responses to 80 multiple-choice questions. This generates a lot of information about employee's experiences by finding the relationship between the two datasets it can help BC Stats to focus on certain areas and gain a deeper understanding of the survey design. We were able to identify areas in the comments to focus on like engagement and workplace culture. Automatic text summarization was used to help gain a deeper understanding of the key ideas present in the agreement levels.

Emotion Analysis

The WES aims to understand their employee's experiences and identify areas for improvement. Gaining insights into the general emotion of employees can help the BC Public Service to understand areas for improvements. We found that the most common emotion is fear and makes up over 20% of all the comments, BC Stats want to focus on reducing the factors contributing to employee's fear to improve morale of employees. As this field progresses more emotions will be added to the emotion lexicon and this analysis can be used to understand a broader range of emotions.

References

[1] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems* 26, pages 3111–3119. 2013.

[3] J. Pennington, R. Socher, and C. D. Manning. (2014). GloVe: Global Vectors for Word Representation

[4] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin. Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405, 2017.

[5] Barrios, F., Lopez, F., Argerich, L., Wachenchauzer, R. (2016). Variations of the Similarity Function of TextRank for Automated Summarization. *JAIIO - ASAI*, 65-72.

Appendix A - Comment Theme and Sub-theme Definitions

1.CPD	Career & Personal Development - 11. Improve new employee orientation - 12. Improve performance management process - 13. Improve training and development opportunities - 14. Provide opportunities for career advancement - 15. other
2.CB	Compensation & Benefits - 21. Ensure salary parity across government - 22, Ensure salary parity with other organizations - 23. Improve benefits - 24. Increase salary - 25. Review job classifications and/or reporting levels - 26. other
3.EWC	Engagement & Workplace Culture - 31. Act on engagement initiatives - 32. Address discrimination and harassment - 33. Improve morale and workplace culture - 34. Treat employees and colleagues better - 35. Value diversity - 36. other
4.Exec	Executives - 41. Improve communication between executives and staff - 42. Improve stability and/or change management - 43. Strengthen quality of executive leadership - 44. other
5.FWE	Flexible Work Environment - 51. Improve and/or expand Leading Workplace Strategies (lws) - 52. Increase flexibility in work location - 53. Increase flexibility in work schedule - 54. other
6.SP	Staffing Practices - 61. Ensure hiring and promotions are fair and merit based - 62. Focus on Human Resources planning (recruitment, retention, succession) - 63. Make hiring process more efficient - 64. other
7.RE	Recognition & Empowerment - 71. Enable staff to make decisions - 72. Listen to staff input - 73. Make better use of employees' skills and abilities - 74. Provide more and/or better quality recognition - 75. other

8.Sup	Supervisors - 81. Cultivate effective teamwork and communication within teams - 82. Hold employees accountable for performance - 83. Strengthen quality of supervisory leadership - 84. Improve communication between employees and supervisors - 85. other
9.SW	Stress & Workload - 91. Hire more staff - 92. Improve productivity and efficiency - 93. Review workload expectations - 94. Support a healthy workplace - 95. other
10.TEPE	Tools, Equipment & Physical Environment - 101. Ensure safety and security of employees - 102. Improve facilities (e.g. office space, noise levels, air quality, etc.) - 103. Provide better supplies and equipment (e.g., office supplies, field instruments, printers, etc.) - 104. Provide better office furniture (e.g. desks, chairs, etc.) - 105. Provide better computer-based hardware (e.g., desktops, laptops, etc.) - 106. Upgrade/improve software - 107. other
11.VMG	Vision, Mission & Goals - 111. Assess plans, priorities and strategies for the organization - 112. Improve collaboration between work units or regions - 113. Improve program and/or policy implementation - 114. Pay attention to the public interest and service delivery - 115. Review funding or budget allocation to program - 116. Reduce political influence - 117. other
12.Oth	Other - 121. Other related comments - 122. Positive Comments - 123. Survey Feedback

Appendix B - Supporting Figures

Table B1. Linking Quantitative to Qualitative Sub-theme Agreement Levels Table

Theme	Sub-Theme Code	Total Number	Strong Agreement (%)	Weak Agreement (%)	No Agreement (%)
Career & Personal Development	12	103	50.49	12.62	36.89
	13	1202	47.92	25.37	26.71
	14	502	47.21	22.31	30.48
Compensation & Benefits	21	146	86.99	9.59	3.42
	23	541	76.71	13.12	10.17
	24	1554	87.45	7.98	4.57
Engagement & Workplace Culture	32	295	63.73	10.85	25.42
	33	357	8.12	14.29	77.59
	34	459	55.12	20.04	24.84
	35	123	31.71	24.39	43.9
Executives	41	785	63.69	21.02	15.29
	42	268	51.87	19.03	29.1
	43	586	67.75	18.09	14.16
Staffing Practices	61	769	66.32	16.25	17.43
Recognition & Empowerment	71	292	34.93	20.55	44.52
	72	434	57.6	22.81	19.59
	73	149	16.11	20.81	63.09
	74	476	56.09	18.7	25.21
Supervisors	81	294	23.13	21.43	55.44
	82	407	63.39	17.44	19.16
	83	1051	52.9	16.46	30.64
	84	491	39.71	21.79	38.49

Stress & Workload	92	510	43.53	25.69	30.78
	93	929	69.32	18.08	12.59
	94	281	52.67	30.25	17.08
Tools, Equipment & Physical Environment	101	406	64.53	16.26	19.21
	102	1348	60.61	21.07	18.32
	103	382	51.31	30.89	17.8
	104	674	55.04	24.78	20.18
	105	611	63.83	22.59	13.58
	106	641	57.88	22.31	19.81
Vision, Mission & Goals	111	587	42.93	29.13	27.94