

BC Stats

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

Mentor: Varada Kolhatkar

Team: Fan Nie, Ayla Pearson, Aaron Quinton

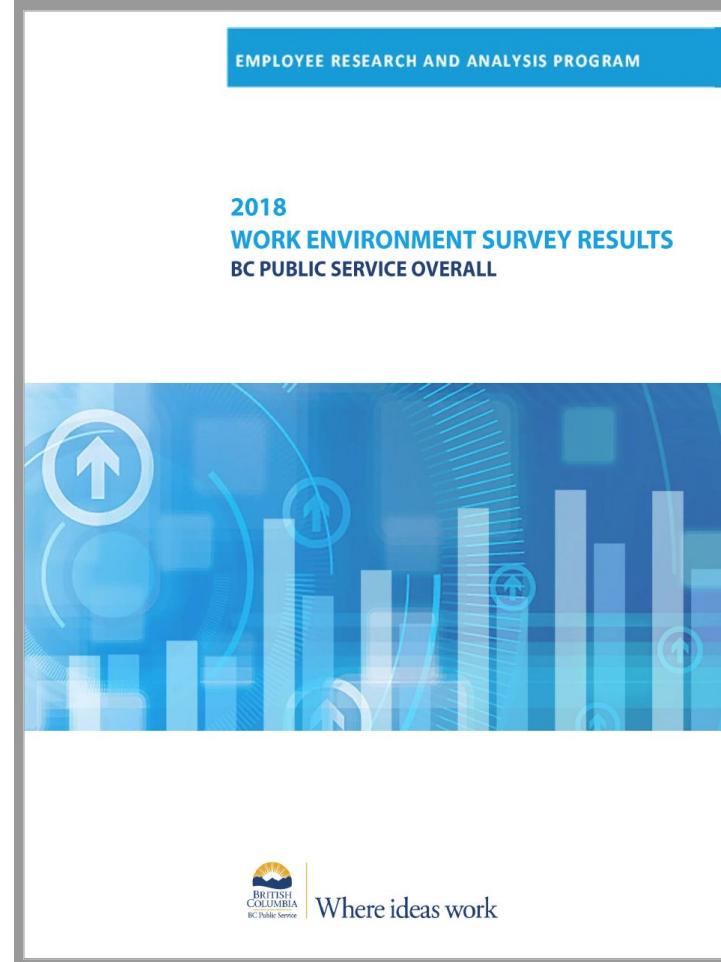
Introduction

Background

- Work Environment Survey (WES)

Primary Goals:

- Understand employee experiences
- Celebrate successes
- Identify improvements



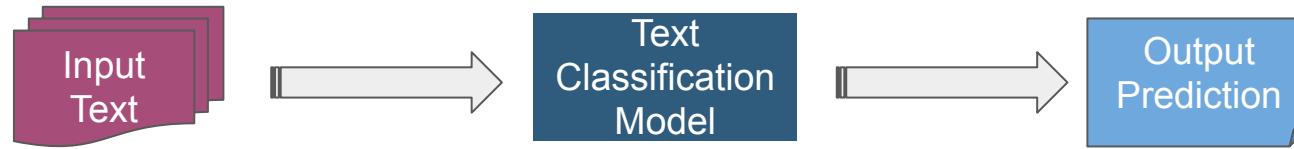
Data

- ❖ There are over 15,000 respondents across 26 ministries in the 2018 Survey.
- ❖ All of the Survey data can be categorized in two:
 - ★ Qualitative data (comments/open ended responses)
 - Open ended question:
“What one thing would you like your organization to focus on to improve your work environment?”
 - Comment example:
“Give us up-to-date equipment and software (not 2010 versions in 2018).”
 - Theme: Tools, Equipment & Physical Environment*
Sub-theme: Provide better computer-based hardware
 - ★ Quantitative data (Approximately 80 5-point scale questions)
 - Example:
“The quality of training and development I have received is satisfactory: [1-5]”

Research Questions:

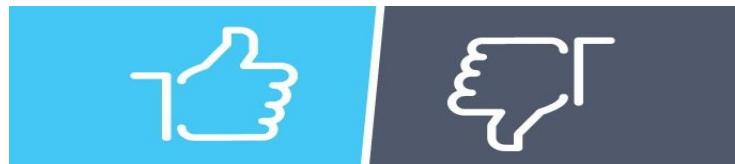
1. Text Classification:

Automate the written questions labelling to the themes and sub-themes

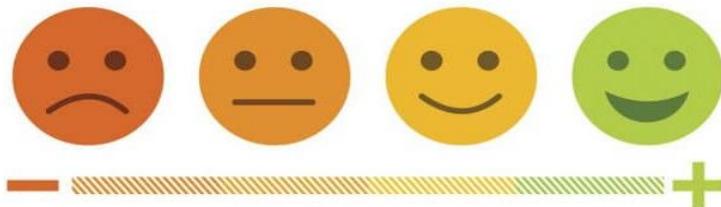


2. Linking Quantitative to Qualitative:

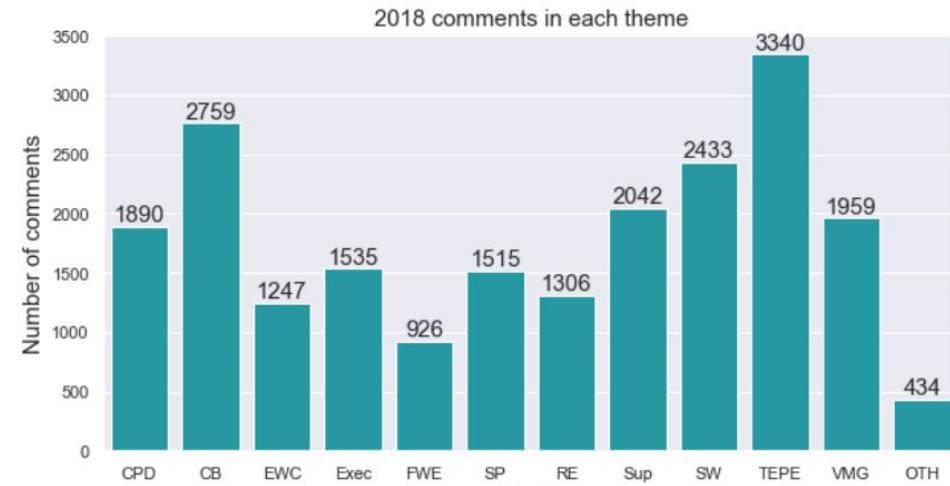
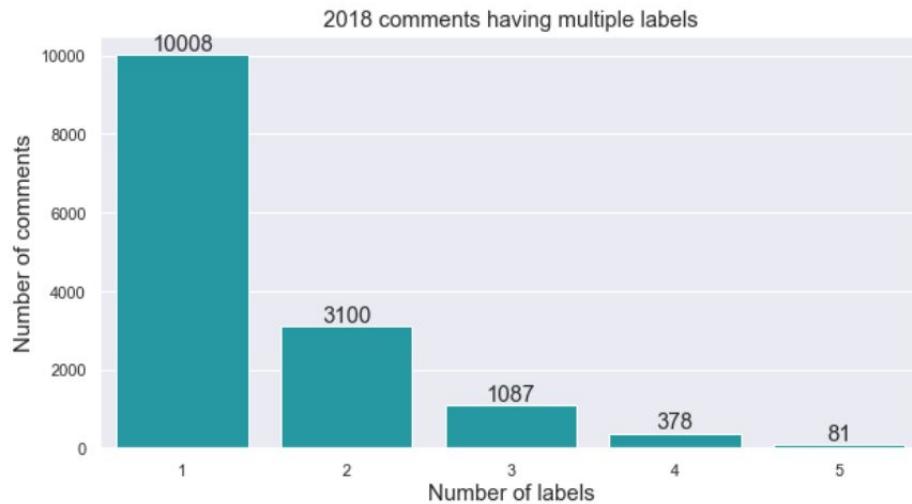
How well does the sentiment of the qualitative responses agree with the quantitative responses?



3. Discovery from Text Analysis

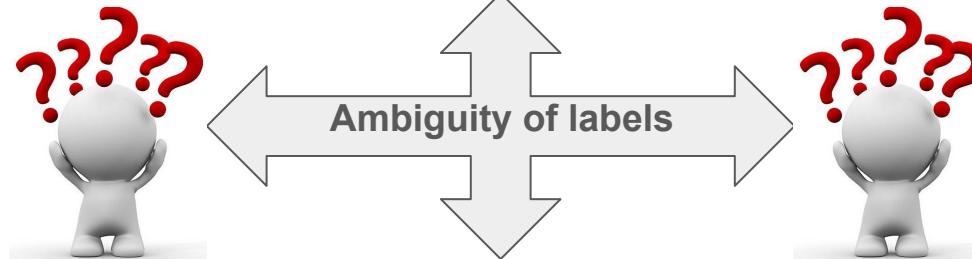


The challenges we have been facing: Multi-label Text Classification



12 labels (themes) in total and individual comments can be coded up to 5 labels

Imbalanced data on each theme



Text Classification Methodology

Data
Preparation

Pre-processing Text and
Feature Engineering

Modeling

Results and
Visualization

Flow Chart Legend

Prerequisite File

Required files that are input into the pipeline. These are provided by the client or downloaded online

Script

A Python or R script used to transform, model, or visualize the data

Generated File

A data file or model generated by a script and often used downstream in the analysis

Text Classification Methodology

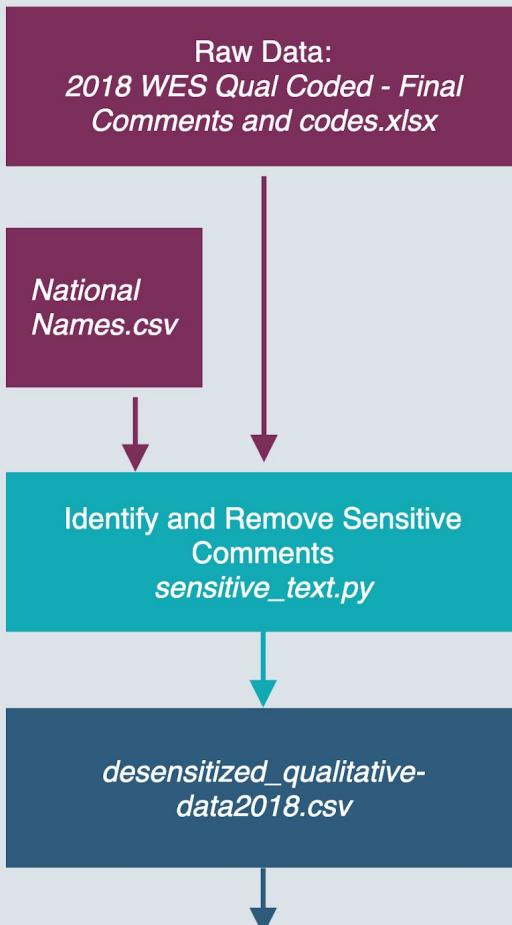
Data
Preparation

Pre-processing Text and
Feature Engineering

Modeling

Results and
Visualization

Data Preparation



Raw Data xlsx file

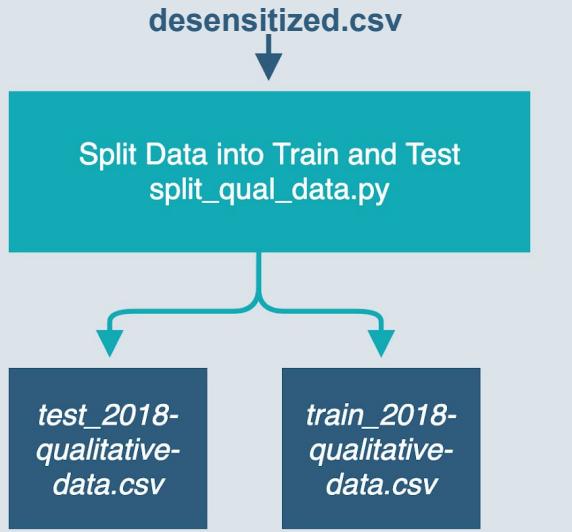
_telkey	2018 Comment	Code 1	...	Code 5	CPD	...
188273-537556	Permit more flexible working arrangements...	53	...		0	...

- Identify Sensitive Comments using Named Entity Recognition (NER) to look for Person names
- Cross Reference that with csv of baby names published by the US Data.gov
- ~200 of 12,000 removed from dataset

Example Comment that got flagged:

“To quote Tom Cruise and Cuba Gooding Jr. in the film Jerry Maguire 'show me the money!!!!”

Data Preparation



One of the Golden Rules of Machine Learning - Don't touch the test data!

As such, we randomly split the data into 90% Train, and 10% Test

Text Classification Methodology

Data Preparation

Raw Data:
2018 WES Qual Coded - Final
Comments and codes.xlsx

National
Names.csv

Identify and Remove Sensitive
Comments
sensitive_text.py

desensitized_qualitative-
data2018.csv

Split Data into Train and Test
split_qual_data.py

test_2018-
qualitative-
data.csv

train_2018-
qualitative-
data.csv

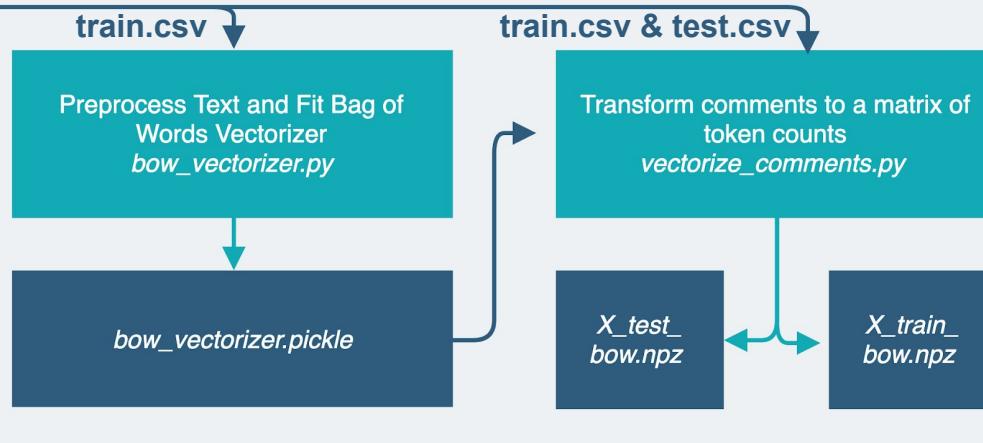
Pre-processing Text and
Feature Engineering

Modeling

Results and
Visualization

Pre-processing Text and Feature Engineering

Bag of Words approach for Linear Classifier



Example Bag of Words (BOW) data

busy	compensation	flexible work	...
0	1	0	...
2	0	0	...

Preprocessing:

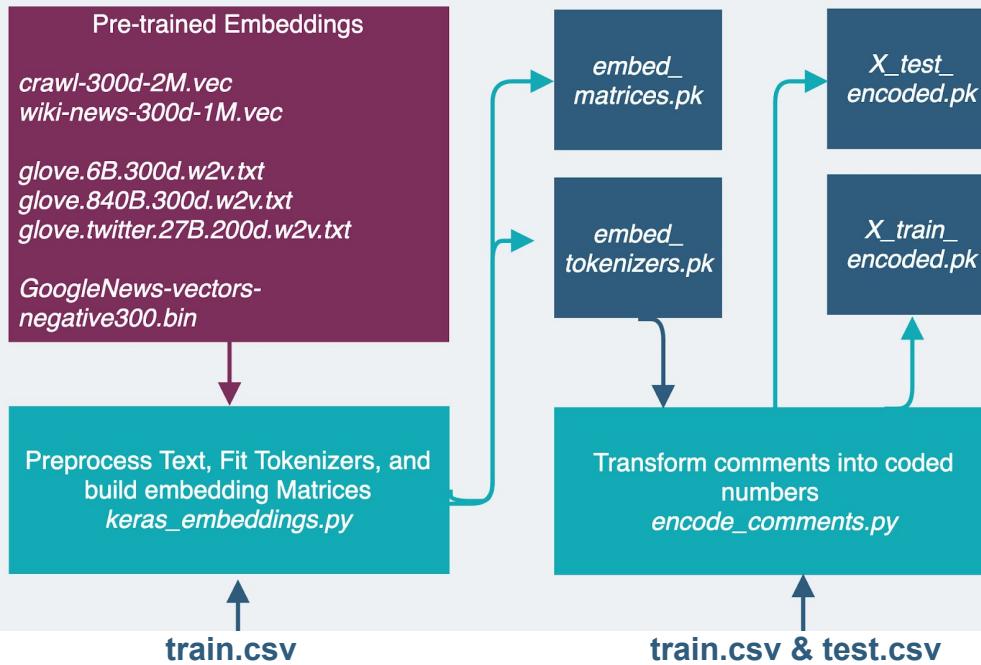
- Removed stop words, punct., lowercase, etc.
- Marginal improvements compared to no pre-processing

Bag of Words (BOW) Transform:

- BOW Model used as our Baseline Classifier
- Considered Count Vectorizer with one to five n-grams
- ~ 30,000 features

Pre-processing Text and Feature Engineering

Word Embeddings for Deep learning Models



4705 60 112 35 800

Ex. Comment: "Permit more flexible working arrangements"

Ex. Tokenized Comment: [0, 0, 0, ..., 0, 4705, 60, 112, 35, 800]

Preprocessing:

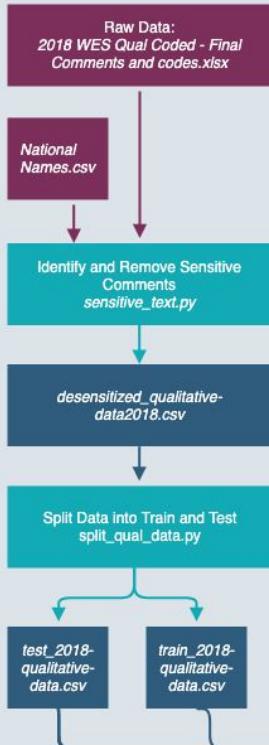
- Minimal pre-processing to capture as much intricacies as possible
- Maximize vocab coverage for each embedding
- Achieved 95 - 98% text coverage

Pre-trained Embeddings and data transformation:

- Explored several embeddings
- Tokenizer maps vocab to indices
- Embed Matrix maps the indices to the embeddings

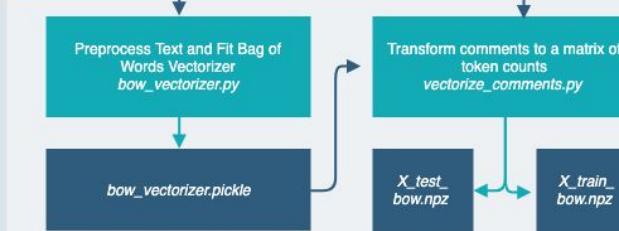
Text Classification Methodology

Data Preparation

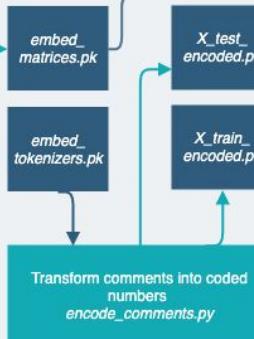
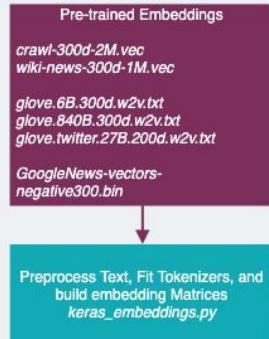


Pre-processing Text and Feature Engineering

Bag of Words approach for Linear Classifier



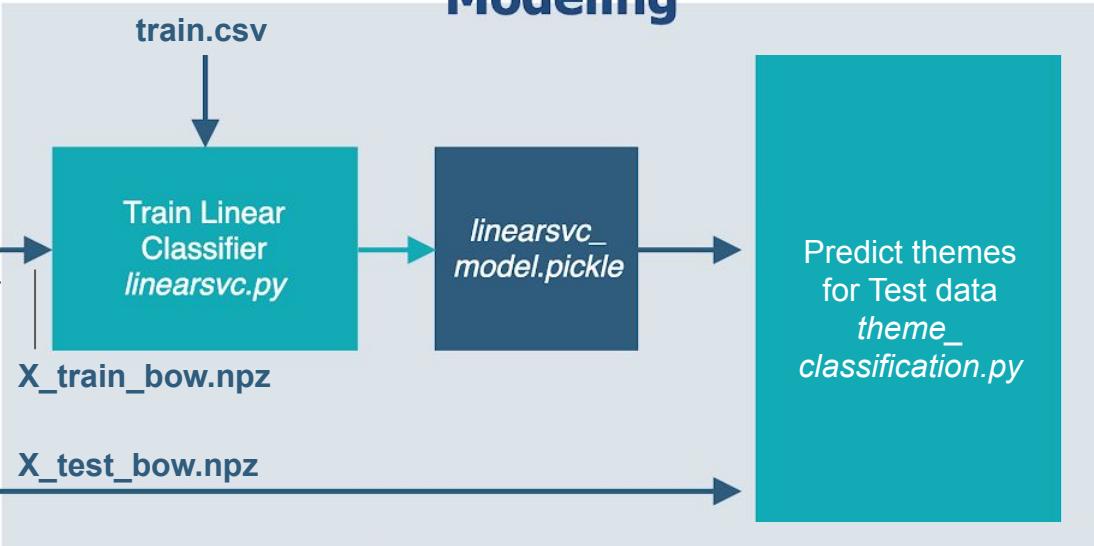
Word Embeddings for Deep learning Models



Modeling

Results and Visualization

Modeling



BOW | LinearSVC:

- One vs All Approach
- train/valid split (not shown in img)
- Grid Search used for Hyper Parameter Optimization:
Marginal Improvements

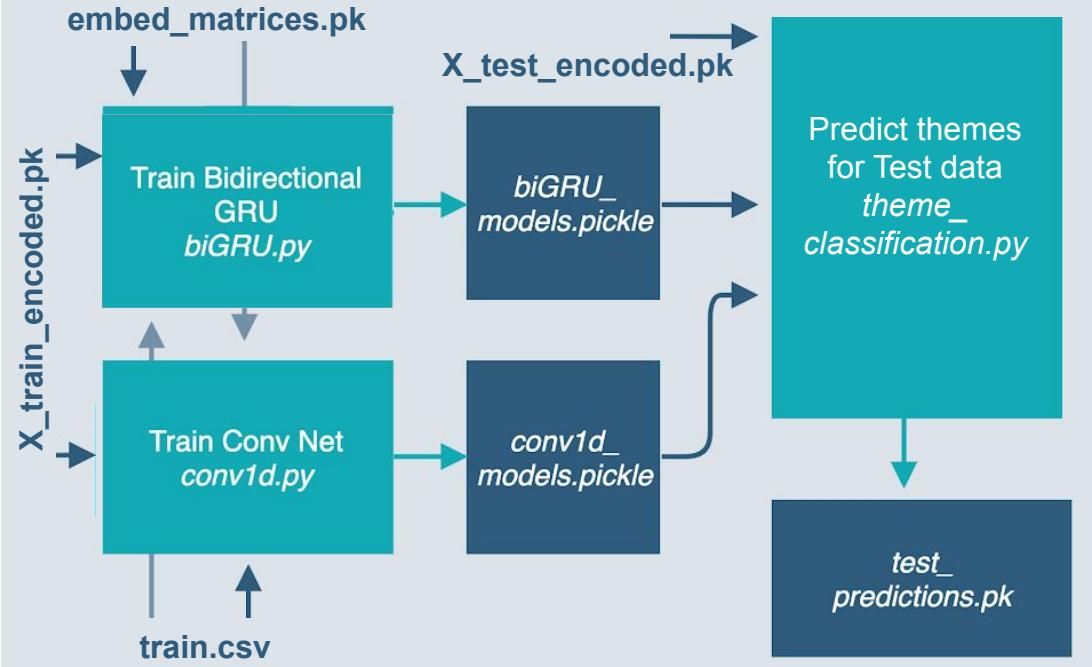
Example Comment Prediction:

Model	CPD	CB	EWC	Exec	FWE	SP	RE	Sup	SW	TEPE	VMG	OTH
BOW LinearSVC	0	0	0	0	0	0	0	0	0	1	0	0

Predictive BOW features

CPD	CB	EWC	Exec	FWE	SP
training courses mentoring growth advancement	wages wage pay compensation salary	favoritism bullying diversity harassment morale	communication executives executive change - management authenticity	lws teleworking telecommuting telework westshore	succession retention recruiting successional vacancies
RE	Sup	SW	TEPE	VMG	OTH
recognition empowerment micro- management autonomy acknowledgment	accountability respect - executive management supervision communication - improved	workload lean bureaucracy caseloads workloads	ergonomics safety icm technology ergonomic	stewardship budget integration funding politics	text satisfied perfect happy jorpp

Modeling



Conv1d & biGRU

- train/valid split (not shown in img)
- Trained each model with each embedding
- Experimenting with Neural Net Architecture made marginal differences
- Ensembled
 - Conv1d | glove_wiki
 - biGRU | glove_wiki
 - biGRU | glove_crawl
 - biGRU | fasttext_crawl

Example Comment Prediction:

Model	CPD	CB	EWC	Exec	FWE	SP	RE	Sup	SW	TEPE	VMG	OTH
Conv1d glove_wiki	0.23	0.04	0.15	0.02	0.34	0.08	0.12	0.07	0.4	0.65	0.2	0.01 ₁₇

Text Classification Methodology

Data Preparation

Raw Data:
2018 WES Qual Coded - Final
Comments and codes.xlsx

National
Names.csv

Identify and Remove Sensitive
Comments
sensitive_text.py

desensitized_qualitative-
data2018.csv

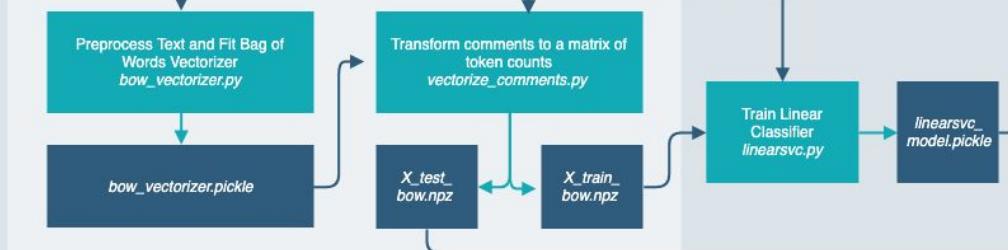
Split Data into Train and Test
split_qual_data.py

test_2018-
qualitative-
data.csv

train_2018-
qualitative-
data.csv

Pre-processing Text and Feature Engineering

Bag of Words approach for Linear Classifier



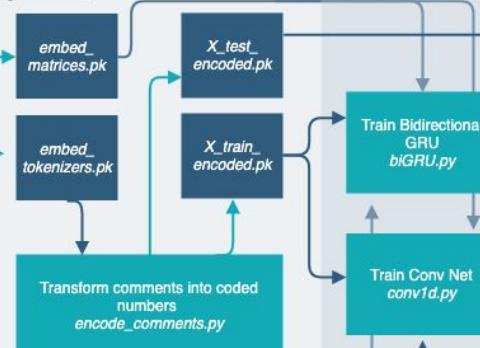
Word Embeddings for Deep learning Models

Pre-trained Embeddings
crawl-300d-2M.vec
wiki-news-300d-1M.vec

glove.6B.300d.w2v.txt
glove.840B.300d.w2v.txt
glove.twitter.27B.200d.w2v.txt

GoogleNews-vectors-negative300.bin

Preprocess Text, Fit Tokenizers, and
build embedding Matrices
keras_embeddings.py



Modeling

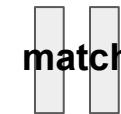
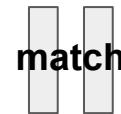


Results and Visualization

How do we measure the success of a model:

- Accuracy on individual label/theme vs Overall Accuracy on the 12 labels as a whole

	Career Development	Compensation	...	Vision, Mission & Goals	Other
True Labels	1	0		1	0



	Career Development	Compensation	...	Vision, Mission & Goals	Other
Predicted Labels	1	1		0	0

Overall Accuracy

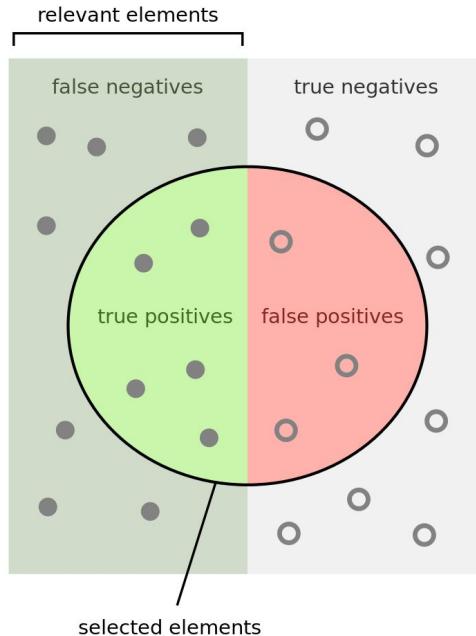


Accuracy on theme level



How do we measure the success of a model:

- Precision and Recall



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

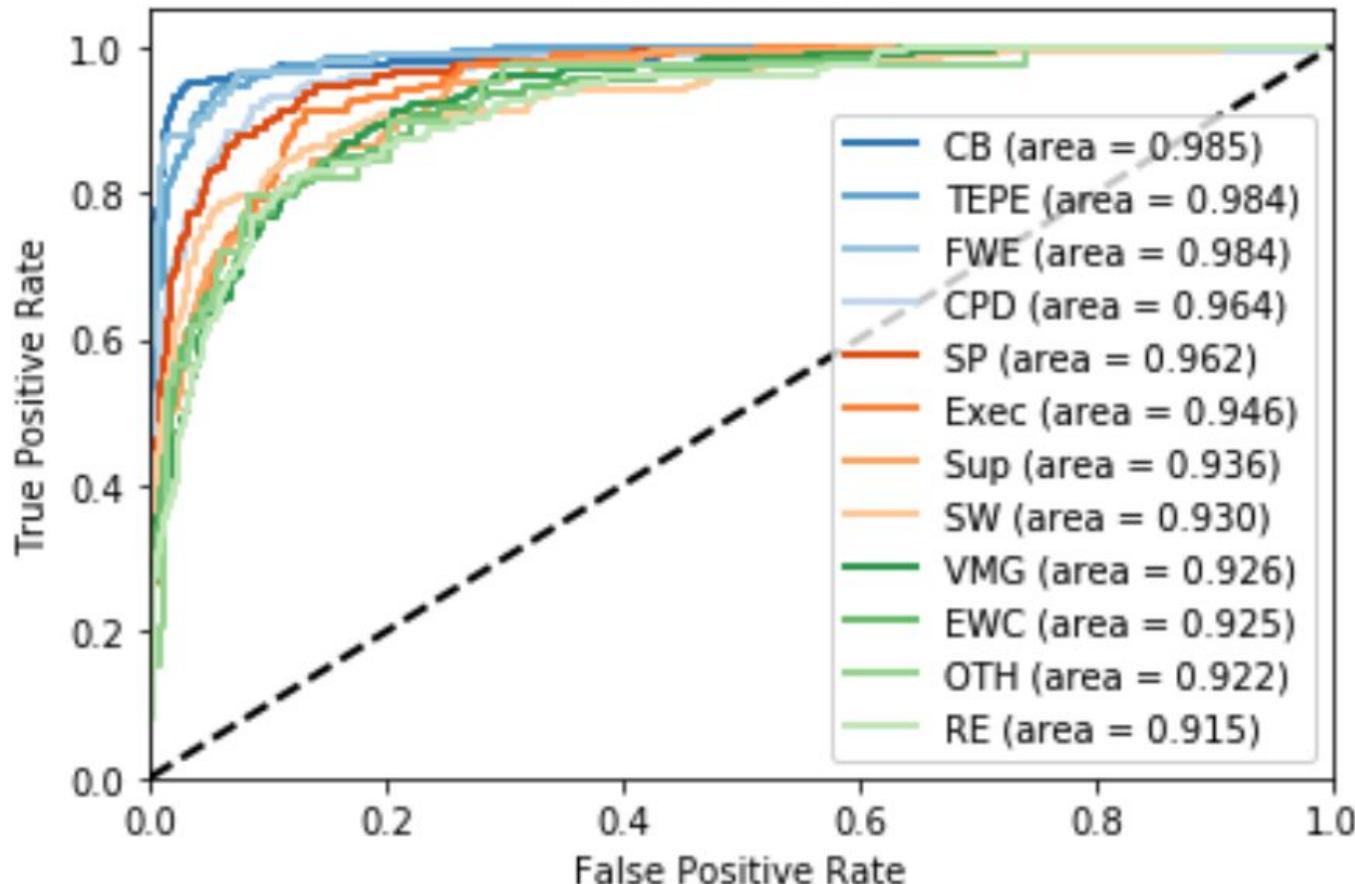
$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

- ROC Curve -- is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- Micro Average -- a weighted average of the precision and recall

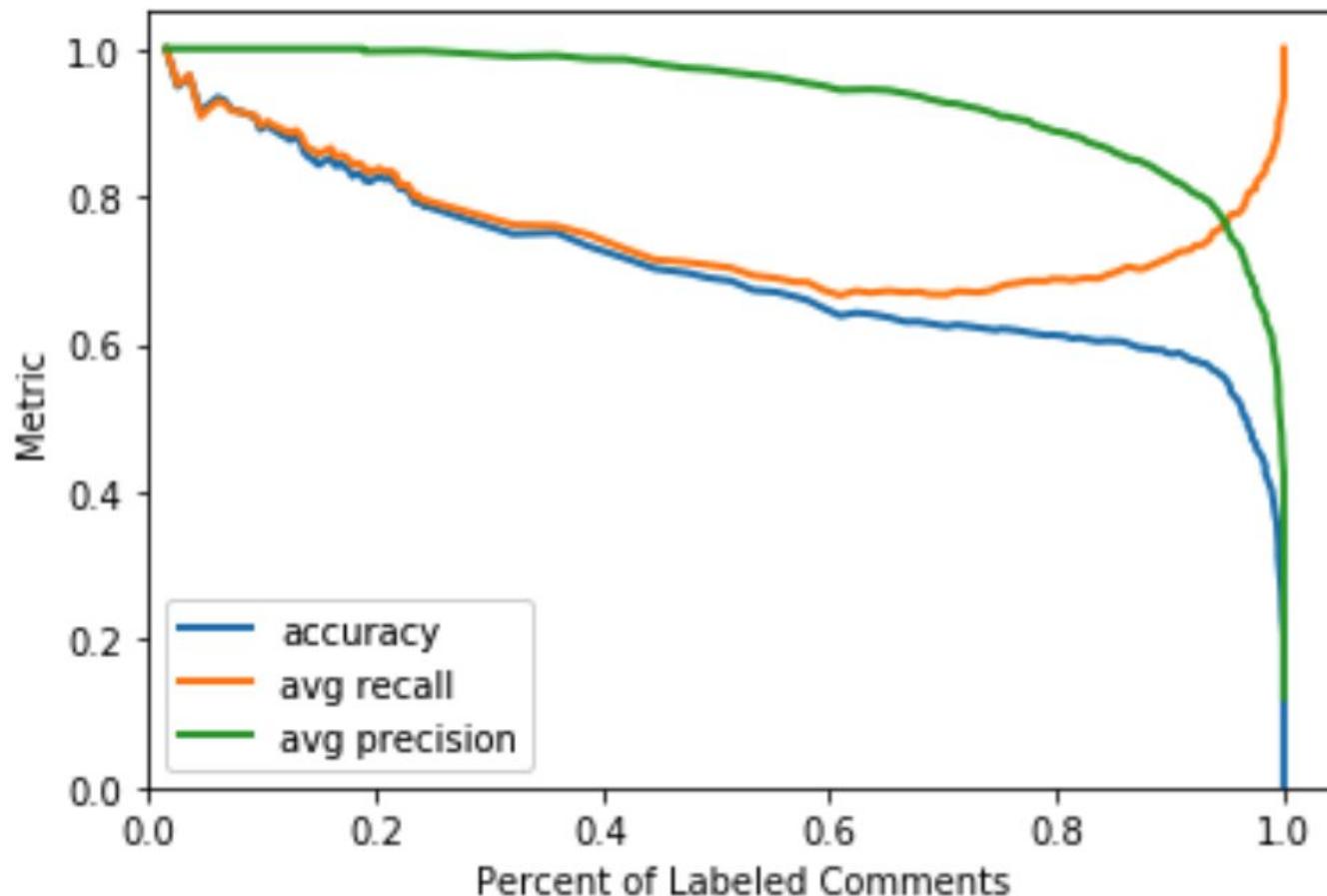
Theme Classification - Selected Model Results

Model	Accuracy	Precision	Recall
BOW LinearSVC	45%	74%	64%
Conv1d GloVe Wiki	51%	80%	64%
biGRU GloVe Wiki	50%	79%	66%
biGRU fasttext Crawl	51%	80%	66%
biGRU GloVe Crawl	50%	78%	66%
Ensemble	53%	83%	66%

Ensemble - ROC curve by Theme



Ensemble - Adjusting the Probability Threshold



Methodologies that did not work on our dataset:

- Term Frequency-Inverse Document Frequency (TF-IDF)
- Pre-processing
- Increasing the data (Doubling the data)
- Average Word Vectors per comment
- Intricate Stacking
- Feature selection - adding meta data / m.c questions

Google BERT

- Was born in October 2018 at Google AI
- BERT is state-of-the-art NLP pre-trained Model
- Out of box, on a subset of our data, BERT has given the pretty good precision and recall compared to our other models
- Computational and privacy challenges
- On our local computer this model took 3 days to do on a subset of our data (for one theme)



Relating Open Ended Survey Questions to Multiple-Choice Survey Questions

The computers are so slow
and big source of frustration

The computer based tools help me to excel

**Strong
Disagree**

**Strongly
Agree**

1

2

3

△

5

- Strength results by agreement between types of survey questions
 - Help BC Stats focus on comments not well captured in the mc questions

Matching Sub-themes & Multiple-choice Questions



Raters	Percent Agreement
All	0.48

Insights

Create a multiple choice question related to subtheme 91 “Hire more staff”

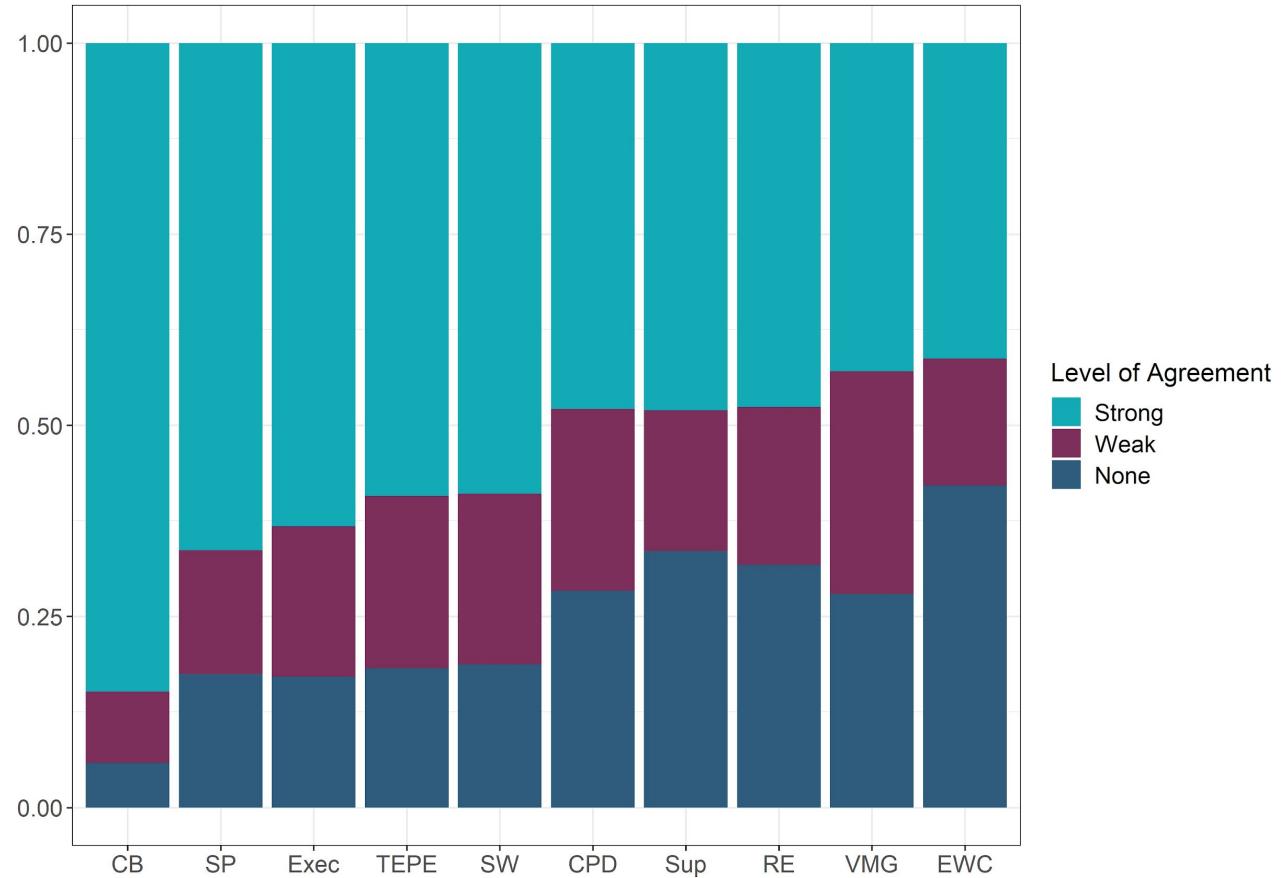
Methodology

Challenges

- There was no mapping between sub-themes & multiple-choice questions
- Not all subthemes matched to a question & vice versa
- Comments could contain more than one topic
- Positive comments weren't coded to a topic

Comments Sentiment	Multiple-choice Sentiment	Level of Agreement
negative	negative	strong
negative	neutral	weak
negative	positive	no

Results



Text Summarization



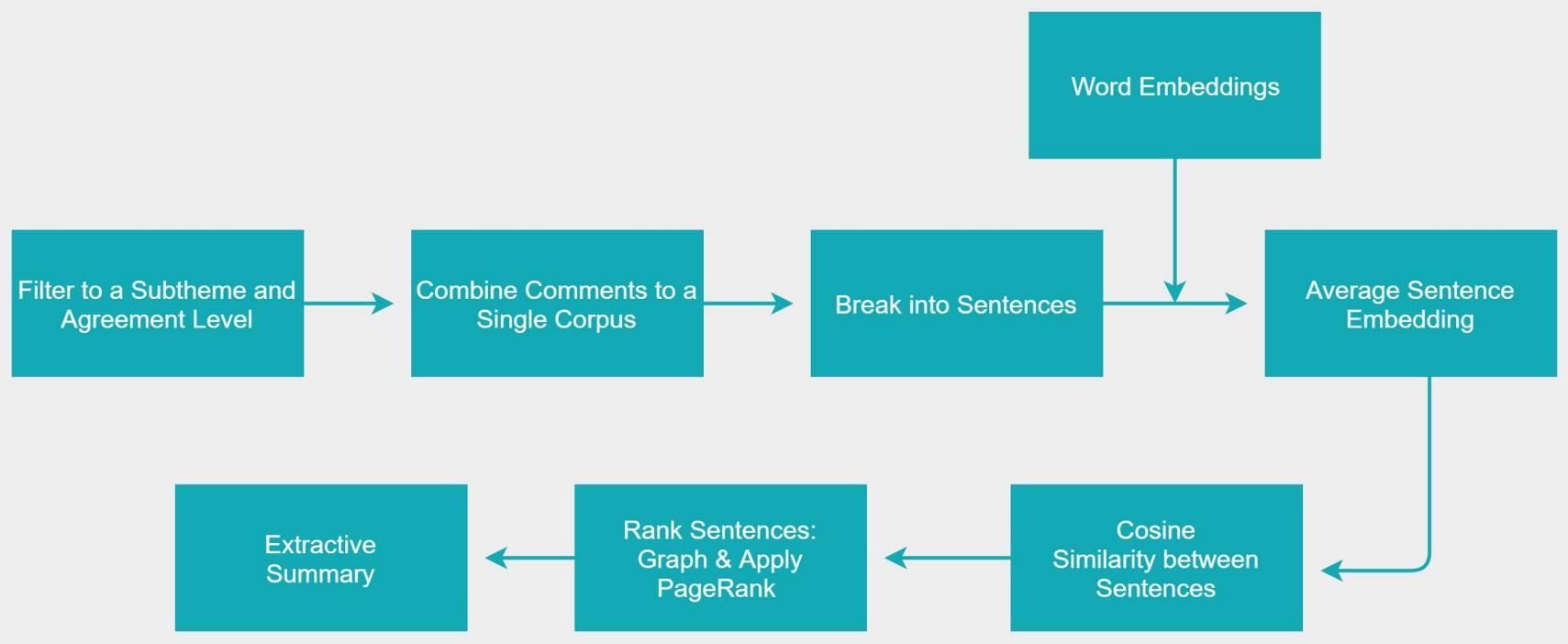
NBA News Articles

Generate
Summary with 4
Sentences



1. This came after the Toronto Raptors clinched their first title with 114 110 win in Game 6 The Finals against the Golden State Warriors at Oracle Arena.
2. Game 6 ended the series while history might suggest the healthier team instead the better team won this title the Raptors did what champions do They pounced on the opportunity sealed the deal.
3. Discouraged irritated by the Raptors past playoff failures that followed successful regular seasons Ujiri traded star guard fan favorite DeMar Derozan for then devalued star Leonard who went rogue on the San Antonio Spurs.
4. The Raptors beat the Golden State Warriors 114 110 in Game 6 the Finals on Thursday win the series 4 2 with commanding performance on the road in Oracle Arenas last game dethrone the leagues reigning dynasty win the Raptors first title.

Text Summarization



Emotion of a Comment

Theme	Sub-theme Description	Sub-theme Code	Anger	Fear	Sad
Executive	Improve Stability	42	0.0	$0.198 + 0.359 = 0.557$	0.0

Comment

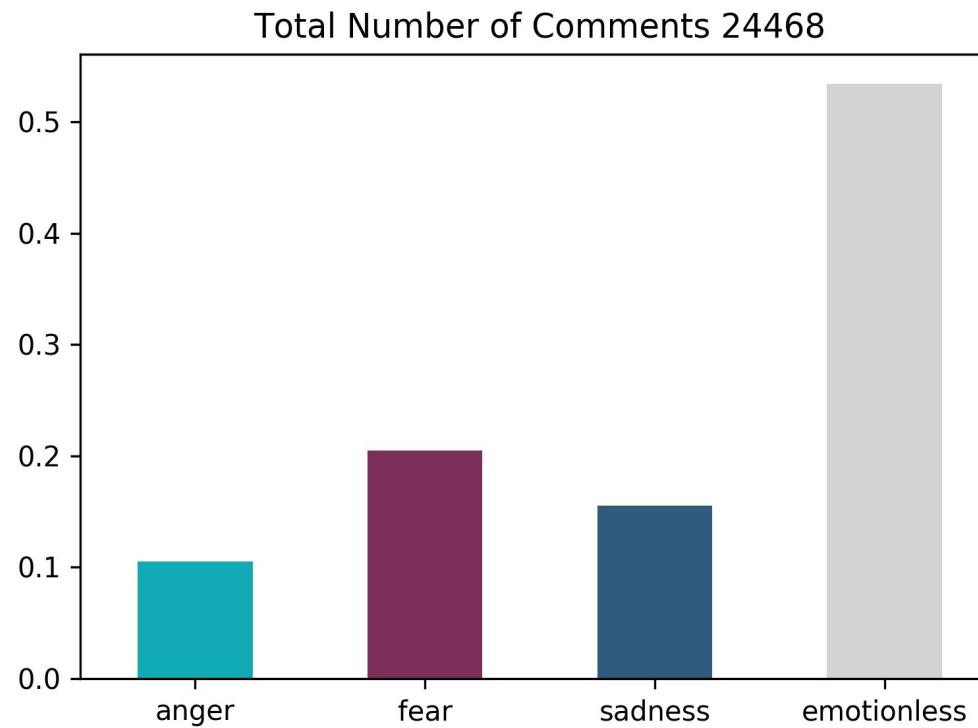
Fear score:
0.198

Fear score:
0.359

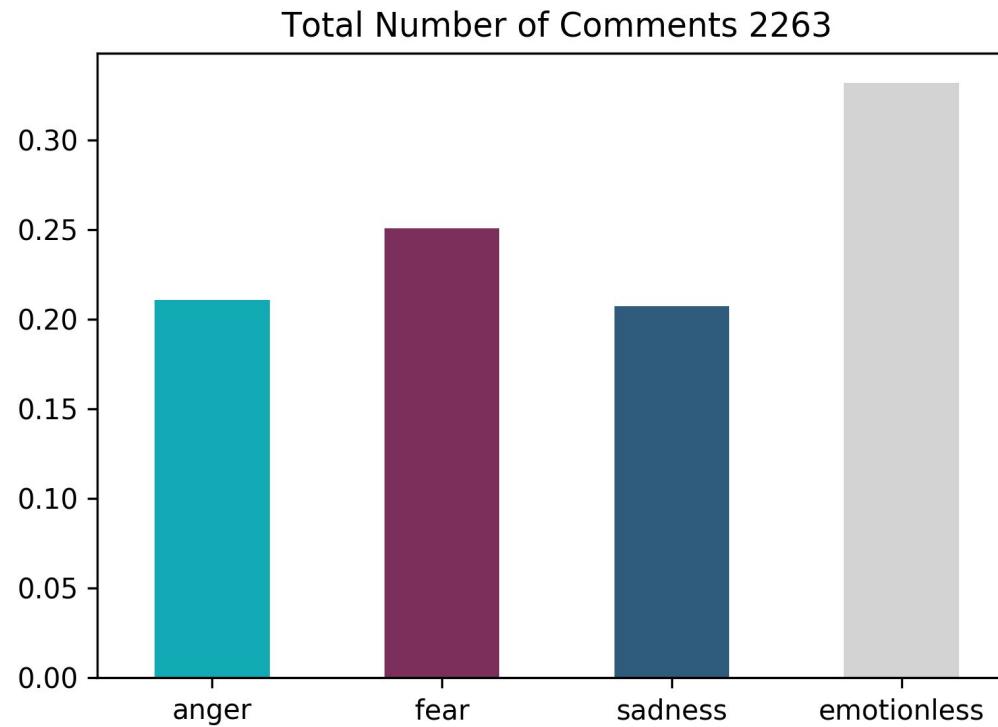
with such a big change-over of priorities and values in government in the past year, it has been somewhat confusing to be a new employee for my organization. some of the changes in organizational structure don't seem to make sense for our line of work.

Credit to The NRC Affect Intensity Lexicon v0.5 by Saif Mohammad

Overall Emotion



Emotions about Supervisors



Data Science & Text

Supervised

- Multi-label classification
 - Bag of Words
 - TfIdf
 - Preprocessing
 - Word Embeddings
 - CNN
 - LSTM
 - BERT

Discovery Analysis

- Mixed methods
 - Text Summarization
 - Emotion analysis



The background of the slide features a abstract blue design. It includes several white circular icons containing upward-pointing arrows of varying sizes. There are also several vertical bars of different heights, some with white outlines and others solid blue. A faint grid pattern is visible across the background.

Thank you

Themes

CPD	Career & Personal Development
CB	Compensation & Benefits
EWC	Engagement & Workplace Culture
Exec	Executives
FWE	Flexible Work Environment
SP	Staffing Practices

RE	Recognition & Empowerment
Sup	Supervisors
SW	Stress & Workload
TEPE	Tools, Equipment & Physical Environment
VMG	Vision, Mission & Goals
Oth	Other

Sub-themes

Executives	41	Improve communication between executives and staff	
	42	Improve stability and/or change management	
	43	Strengthen quality of executive leadership	
	44	Executives - other	
	⋮	⋮	⋮
Stress & Workload	91	Hire more staff	
	92	Improve productivity and efficiency	
	93	Review workload expectations	
	94	Support a healthy workplace	
	95	Stress & workload - other	