Practical NLP

Final Presentation

Agenda

1. Introduction
2. Data Preparation
3. Sentiment Analysis
4. Topic Modeling
5. Executive Report
6. Reflection
7. References

Introduction

Introduction

1. Main Goals
    - Explore and implement practical NLP methods towards resolution of a real-world use case problem (employee survey analysis).

Introduction

1. Main Goals
    - Explore and implement practical NLP methods towards resolution of a real-world use case problem (employee survey analysis).

2. Project Phases
    - S1: Explore and prototype available approaches.
        - Decompose the task and organize a sensible group structure.
    - S2: Integrate microservices into a coherent product.
        - Informative, group-specific presentations.

Introduction

1. Data
    a. Employee survey data from a large, international company.

Introduction

1. Data
    a. Employee survey data from a large, international company.
    b. Open comments.

Introduction

1. Data
    a. Employee survey data from a large, international company.
    b. Open comments.
    c. Two questions:
        i. What is working well?
        ii. What can be improved?

Introduction

1. Data
    a. Employee survey data from a large, international company.
    b. Open comments.
    c. Two questions:
        i. What is working well?
        ii. What can be improved?
    d. Two departments:
        i. Large department.
        ii. Small department.

Introduction

1. Problem Statement
    a. Report on employee satisfaction.

Introduction

1. Problem Statement
    a. Report on employee satisfaction.

2. Solution
    a. Model topics in dataset.
    b. Sentiment analysis.

Introduction

1. Problem Statement
    a. Report on employee satisfaction.

2. Solution
    a. Model topics in dataset.
    b. Sentiment analysis.

THE CUSTOMER IS THE CENTER.
we are usually seeking customer satisfaction
The CIF training is really a good chance for everyone
Training programs are good for team building and product knowledge (ie CIF program).
PLS CONTINUE THE ZUMBA SESSION.
[company] is a good company

Introduction

1. Problem Statement
   a. Report on employee satisfaction.
2. Solution
   a. Model topics in dataset.
   b. Sentiment analysis.
3. Methods
   a. Data preparation.
   b. Unsupervised document clustering.
   c. Sentiment classification.

THE CUSTOMER IS THE CENTER.
we are usually seeking customer satisfaction
The CIF training is really a good chance for everyone
Training programs are good for team building and product knowledge (ie CIF program).
PLS CONTINUE THE ZUMBA SESSION.
[company] is a good company

Data Preparation

Data Preparation

Aim

Clean and consolidate data so that it can be used for further processing.

Data Preprocessing Techniques

1) Removing punctuations and stopwords.
2) Replacing special characters and single letter characters with spaces.
3) Lemmatization and Stemming of the Words

Named Entity Recognizer:
A Multiple Labelled Classification

Objective

- To have a model to identify entities from the data provided to us.
- To have a classification model with pre-defined labels
- A model functional for multiple classifications labels

Work Done

- We create a basic sequential model of Recurrent Neural Network Layers.
- We train on all the Labels in one model. (Geographical Locations, Organization, Persons, Events and Natural Phenomenon)
- The model predicts the probability for each label for a particular token.
- The label with the highest probability is considered to be the classified label for that token.

Spelling Correction

Problem:

Out-of-box solutions usually use dictionaries to detect errors → unacceptable amount of False Positives (e.g. in cases of abbreviations).

Solution:

Develop spelling correction that does not automatically consider unknown words to be errors.

Spelling Correction

How we do things differently:

Instead of checking every word against a dictionary, we generate an “error dictionary” based on the entire text.

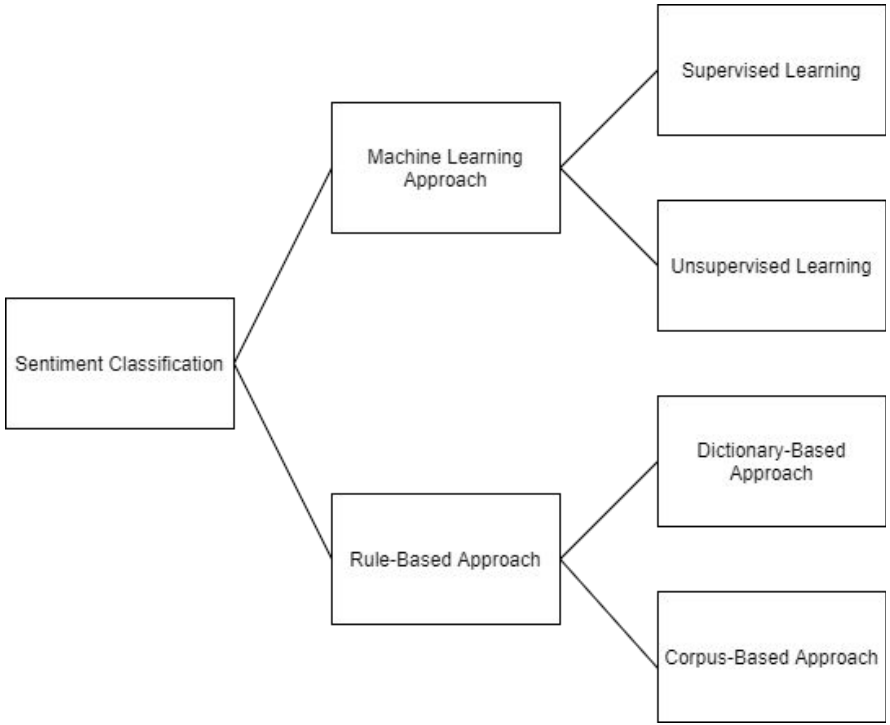
The Process:

1. We generate FastText embeddings on the given text.
2. We query in descending order of frequency (in the given text) for neighbors.
3. We check these neighbors against a list of criteria whether they are errors.
4. If conditions are met, add error to the dict and consider the queried word the original, “correct” word.

Sentiment Analysis

Problem statement

- Automatically identify the polarity of a comment in a **supervised manner**
- Classification task: positive, negative, neutral
- Criteria for choosing the approach: speed, scalability, noisy data
- Most common approaches are deep neural networks or SVMs



Explored approaches

Bert (Bidirectional Encoder Representations from Transformers)

- Obtains state-of-the-art results on many NLP tasks and outperforms other models on SST-2 and SST-5 datasets
- Massive pre-trained models that can be fine-tuned on a smaller datasets

LSTM (Long Short-Term Memory)

- Accessible and relatively fast neural network architecture
- Works well on sequential data due to its ability to store information for longer periods of time
- Scalable

Explored approaches

Bert (Bidirectional Encoder Representations from Transformers)

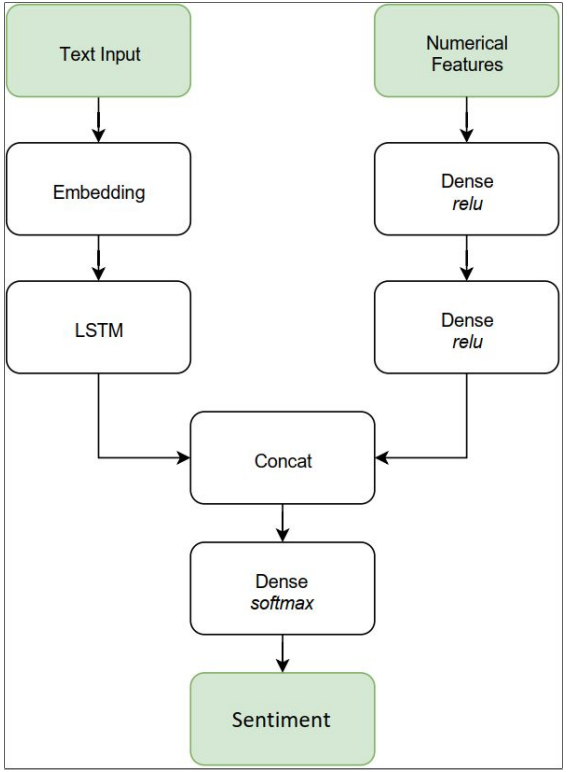
- Obtains state-of-the-art results on many NLP tasks and outperforms other models on SST-2 and SST-5 datasets
- Massive pre-trained models that can be fine-tuned on a smaller datasets
- Computationally expensive

LSTM (Long Short-Term Memory)

- Accessible and relatively fast neural network architecture
- Works well on sequential data due to its ability to store information for longer periods of time
- Scalable

LSTM Architecture

- LSTMs work best with sequential data
- In order to use non-sequential features (punctuation, spelling-based features), we adopt a **hybrid LSTM approach**
- Left side input: Sequential information
- Right side input: Numerical features



Word Embeddings

Twitter US Airline Sentiment

- from Crowdflower's Data for Everyone library
- contains about 15000 tweets about problems of each major US airline
- manually labeled by contributors: negative, neutral, positive

Pretrained Embeddings

- Word Embeddings trained using Word2Vec
- Trained on 590 million English Tweets

The Dataset - Annotation

[Q2] Please tell us what needs to be improved.

[Q1] Please tell us what is working well.

- Two questions, highly directional
- Answers lean strongly on the context of the questions
- We manually labeled 600 questions (3 annotators)

	Q1	**Q2**	**total**
Negative Sentiment	16	210	226
Neutral Sentiment	103	86	189
Positive Sentiment	180	4	184

- **60.02%** of answers in **Q1** have positive sentiment
- **70%** of answers in Q2 have negative sentiment

Training

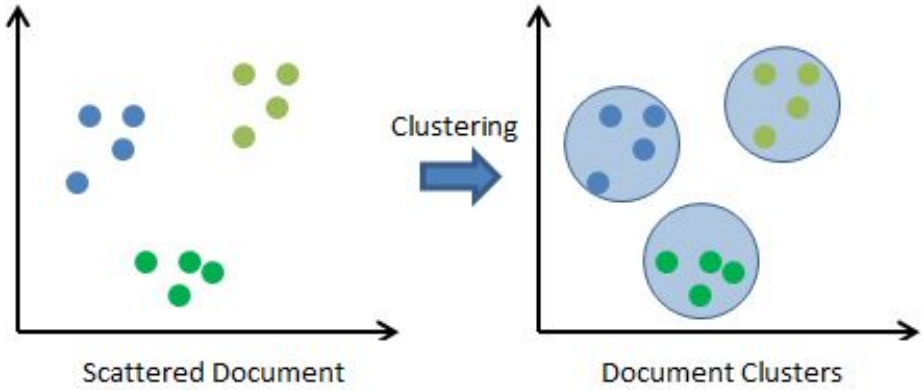
- Due to the nature of the data, one model would not generalize
- Train 2 models, one for each question
- Worth noting that the labeled training dataset was quite small and uncertain: the annotators chose the same labels in 73% of cases, with an Cohen’s kappa score of 0.34

Accuracy
Model question 1 = 0.64 Model question 2 = 0.78 Average = 0.71

Topic Modeling

Topic Modeling

1. Vectorization
2. Dimensionality reduction
3. Cluster analysis
4. Topic labeling and representative sentences
5. Summary



Vectorization

- obtaining word vectors for each token in the dataset
  - using pretrained word embedding models (e.g. Word2Vec, FastText)

Vectorization

- word vectors for each token in the dataset

→ **But:** we want to cluster text not tokens!

  - *Goal*: uniform data structure which preserves the semantic structure of original texts' concatenated embeddings

  - *Solution*: simple average of the token embeddings

Vectorization

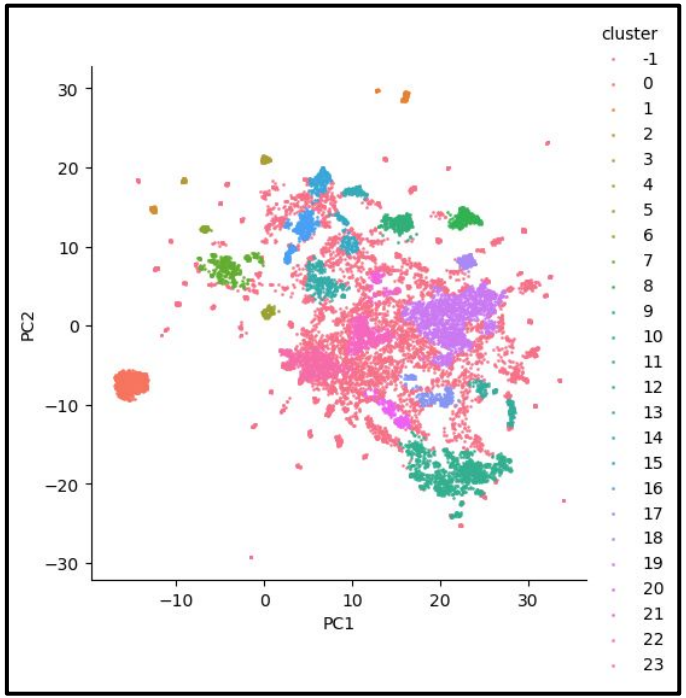
- word vectors for each token in the dataset

→ **But:** we want to cluster text not token!

  - *Goal*: uniform data structure which preserves the semantic structure of original texts' concatenated embeddings

  - *Solution*: simple average of the token embeddings

  → 'washes out' the semantic topography



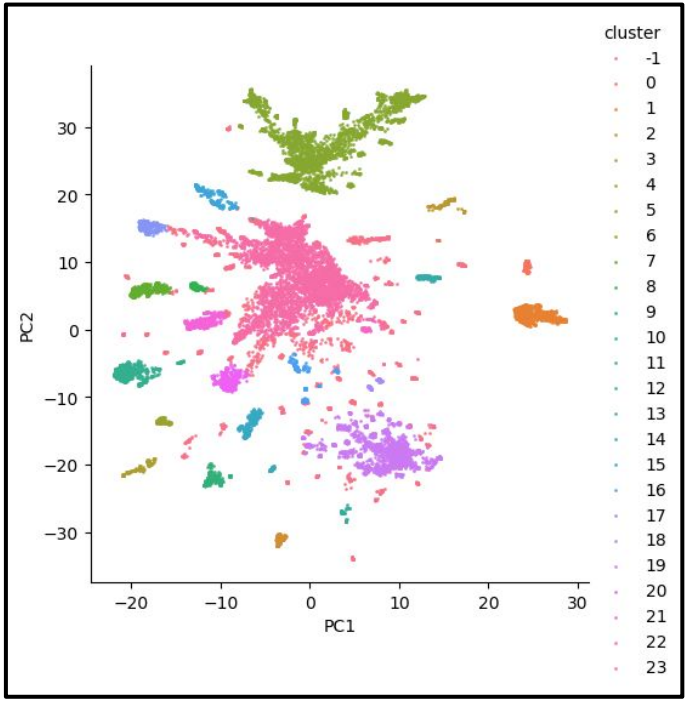
Vectorization

- Final approach:
  - smoothed tfidf-weighted average token embeddings

Vectorization

- Final approach:
  - smoothed tfidf-weighted average token embeddings

→ sample clustering with this approach suggests improved separation of document clusters



Dimensionality Reduction

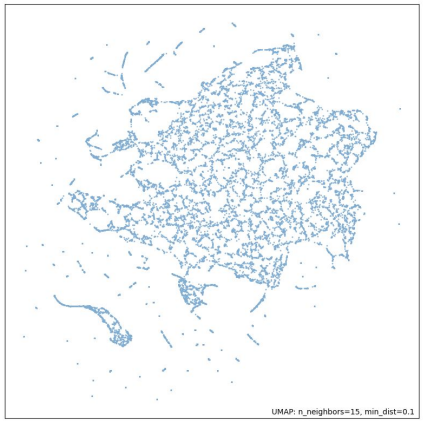
- discovering non-linear, non-local relationship in data
- reducing noise by reducing the dimensions
- easier to apply simple learning algorithms to smaller subset

Dimensionality Reduction

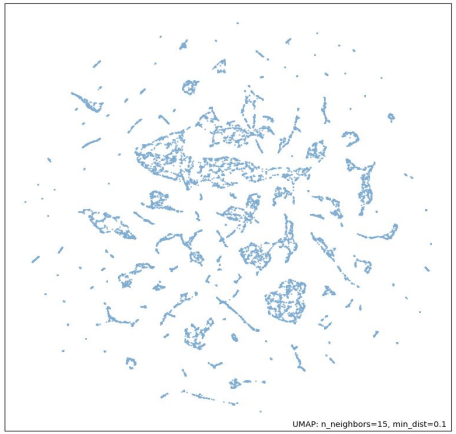
- discovering non-linear, non-local relationship in data
- reducing noise by reducing the dimensions
- easier to apply simple learning algorithms to smaller subset

- Our approach
  - Using PCA and UMAP in combination

dimensionality red only with UMAP (without PCA)

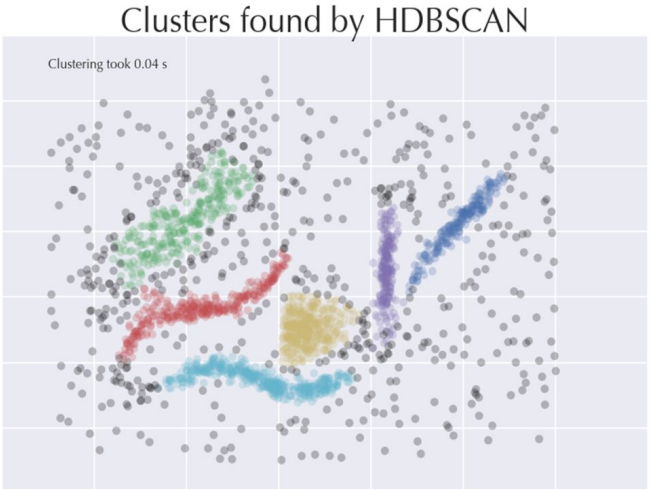


dimensionality red with UMAP and PCA



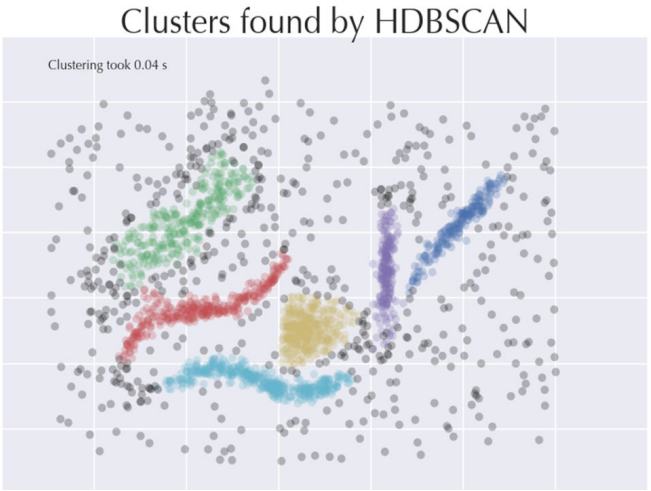
Clustering

- Cluster comments by semantic similarity
- With right amount of clusters, meaningful topics should emerge
  BUT: number of topics not known!



Clustering

- Cluster comments by semantic similarity
- With right amount of clusters, meaningful topics should emerge
  BUT: number of topics not known!



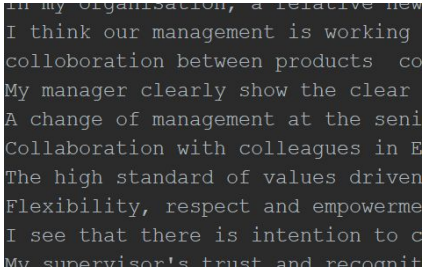
- We use HDBSCAN (Hierarchical Density Based Clustering)
  - Best for noisy and high dimensional data
  - Works well when shape and density of clusters are undefined
- Fine-tune hyperparameters using Silhouette Score as metric for quality of clusters

Topic Labeling

- For clusters to be interpretable topic label and/or representative sentences needed
- Difficult problem:
  - Unsupervised task
  - What does a good solution look like?
  - Several equally good solutions possible
  - Good Labels often contain abstract concepts (e.g. sustainability)

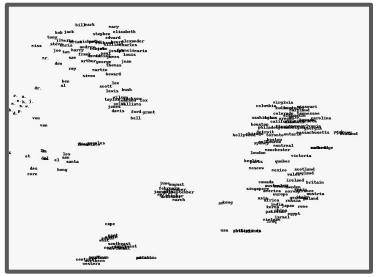
Topic Labeling

- For clusters to be interpretable topic label and/or representative sentences needed
- Difficult problem:
  - Unsupervised task
  - What does a good solution look like?
  - Several equally good solutions possible
  - Good Labels often contain abstract concepts (e.g. sustainability)
- Approach for representative sentence:
  - Find nearest neighbours of mean sentence embedding of cluster in vector space
- Approach for topic labels:
  - TF-IDF of all comments in cluster → important words for each topic
  - Use important words and calculate keywords with nearest neighbour with GenSim library
  - Final selection based on keywords and representative sentences



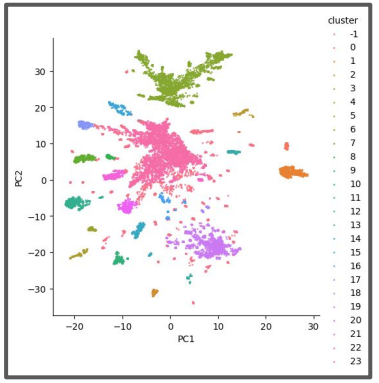
Raw comments

Tokenization



Vectorized comments

Dim. reduction



Clustered comments

Topic Modeling Pipeline Summary

Tf-idf, Mean embeddings

Topic labels

```
cluster  0 : [('working', 74618.234375),
cluster  1 : [('companies', 101210.4687
cluster  2 : [('working', 70848.03125),
cluster  3 : [('service-and', 38884.820
cluster  4 : [('organization', 74787.96
cluster  5 : [('compensation', 45519.13
cluster  6 : [('managers', 95331.039062
cluster  7 : [('personnel', 79832.75),
cluster  8 : [('services', 56049.777343
cluster  9 : [('home', 39433.83984375),
cluster  10 : [('enterprise', 60520.234
```

Most similar

Topic x:
Keywords: [...]
Representative Sentences: [...]

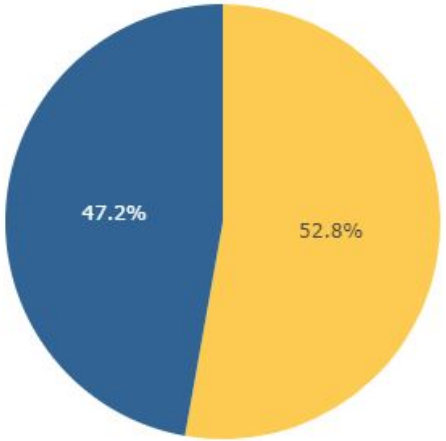
Summary

Output

Executive Report

Quantitative Disparity of Answers

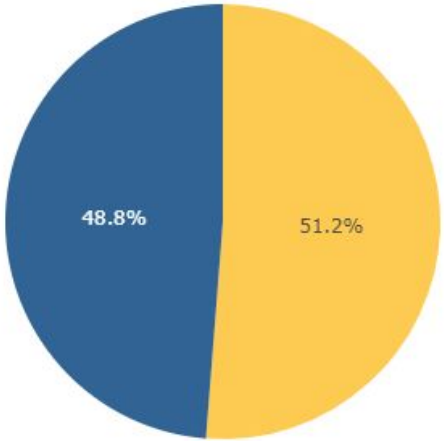
Distribution of Answers in the Small Department

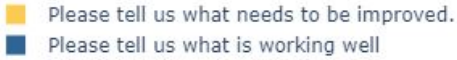


In absolute numbers:

Answers Q1: 575
Answers Q2: 643

Distribution of Answers in the Large Department





In absolute numbers:

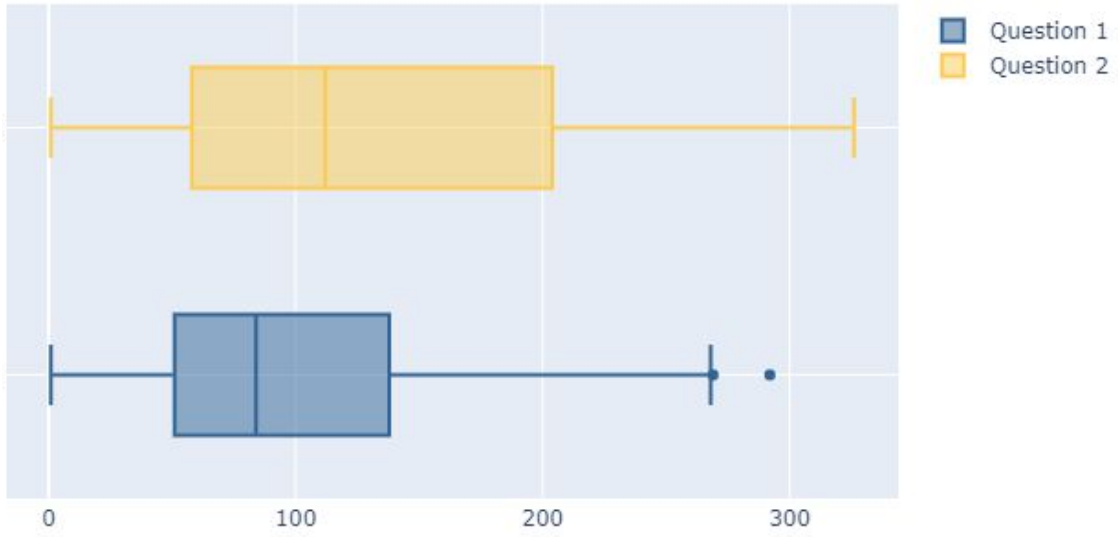
Answers Q1: 8,023
Answers Q2: 8,431

Employee Engagement (measured in comment length)

Between Questions

Please tell us what could be improved.

Please tell us what is working well.

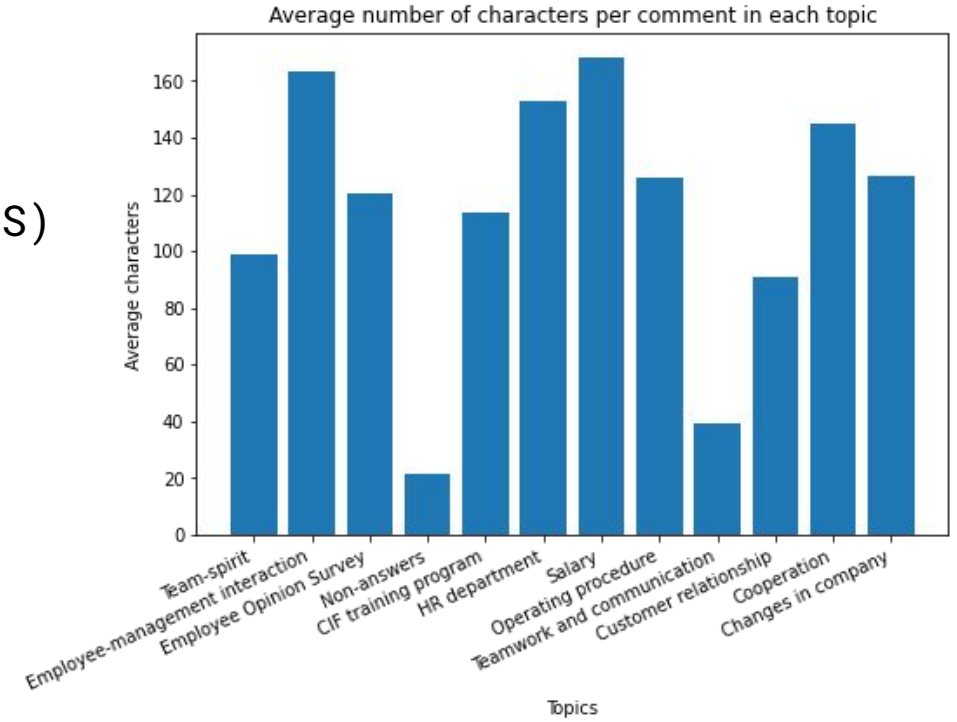


Topics and Engagement

- Team-Spirit
- Employee-Management Interaction
- Employee Opinion Survey (EOS)
- 'Non-Answers'
- CIF Training Program
- HR Department
- Salary
- Operating Procedures
- Teamwork and Communication
- Customer Relations
- Cooperation

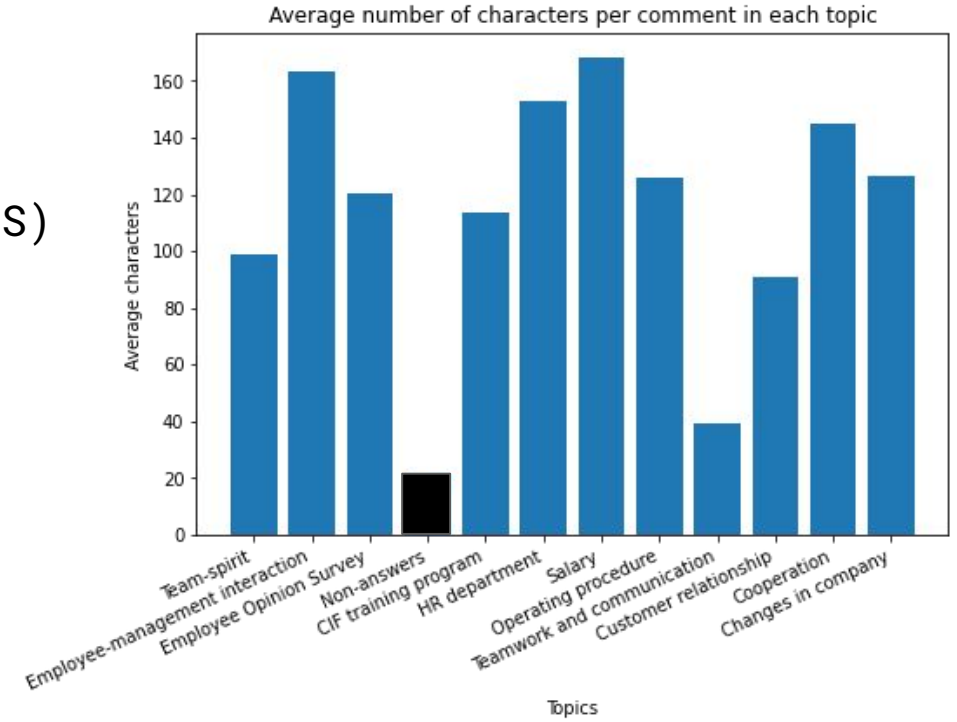
Topics and Engagement

- Team-Spirit
- Employee-Management Interaction
- Employee Opinion Survey (EOS)
- 'Non-Answers'
- CIF Training Program
- HR Department
- Salary
- Operating Procedures
- Teamwork and Communication
- Customer Relations
- Cooperation



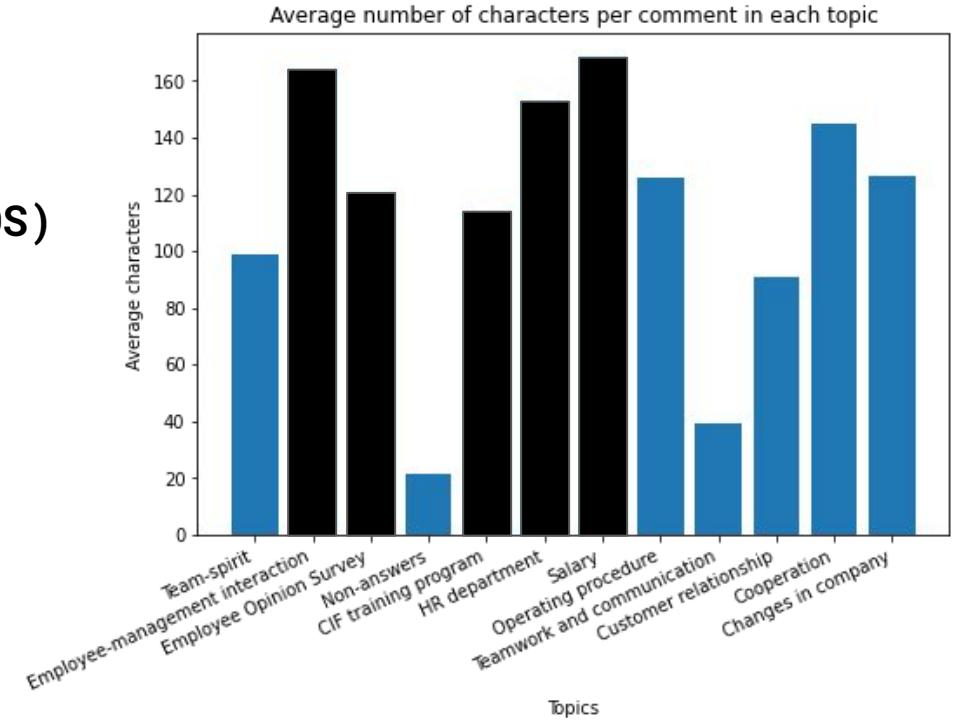
Topics and Engagement

- Team-Spirit
- Employee-Management Interaction
- Employee Opinion Survey (EOS)
- **'Non-Answers'**
- CIF Training Program
- HR Department
- Salary
- Operating Procedures
- Teamwork and Communication
- Customer Relations
- Cooperation

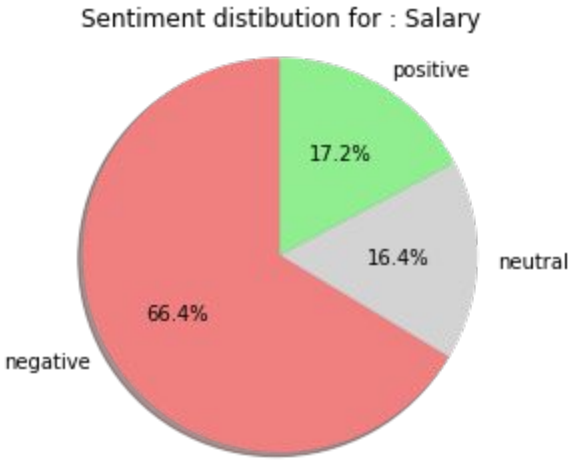


Topics and Engagement

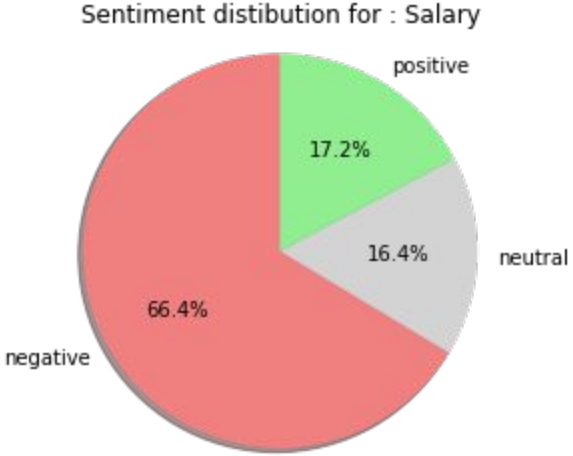
- Team-Spirit
- **Employee-Management Interaction**
- **Employee Opinion Survey (EOS)**
- 'Non-Answers'
- **CIF Training Program**
- **HR Department**
- **Salary**
- Operating Procedures
- Teamwork and Communication
- Customer Relations
- Cooperation



Salary

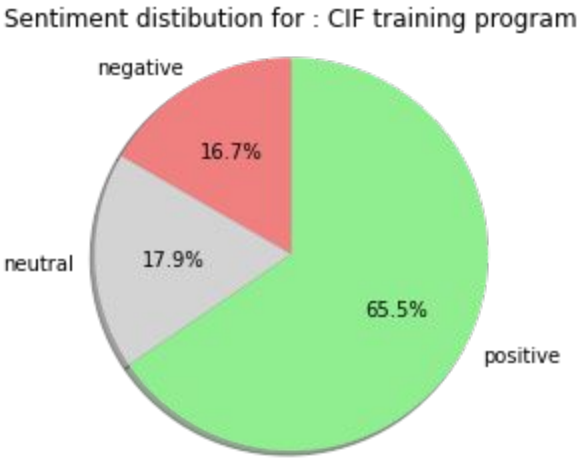

Sentiment distibution for : Salary
positive
17.2%
16.4%
neutral
66.4%
negative

Salary

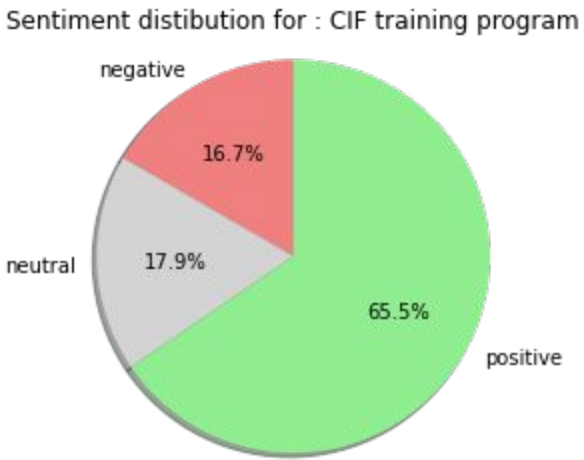

Sentiment distibution for : Salary
positive
17.2%
16.4%
neutral
66.4%
negative


Salary
need
less
salary
work
employee
company
time
team
working
people
much
supervisor
training
compared
paid
good
provide
companies
Review
cost
make
leadership
communication
management
improved
manager
staff
customer
department
service
focus
business
getting
number
must
better
etc
increment
position
reports
many
system
support
increase
issue
problem
new
one
facility
KPI
meeting
sales
level
pay
clear
email
take
quality
senior

CIF Training Program

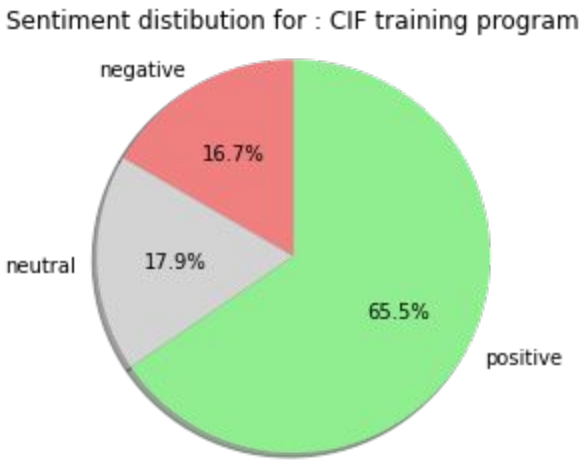


CIF Training Program



- “I feel that the CIF program and its different modules are a big investment and that its results will be favorable to [company]. I believe that this types of programs are much needed to change the culture and achieve the results [company] needs.”

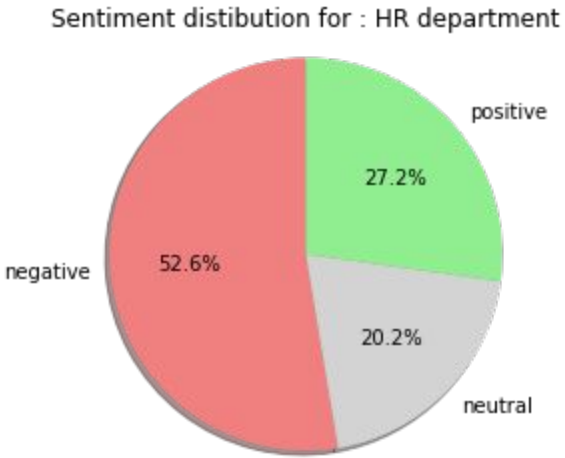
CIF Training Program



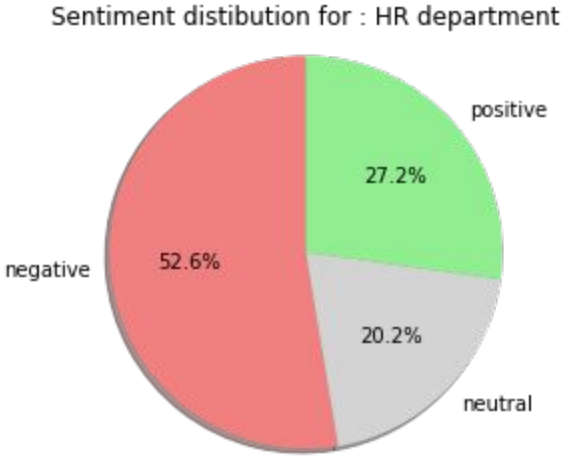
- “I feel that the CIF program and its different modules are a big investment and that its results will be favorable to [company]. I believe that this types of programs are much needed to change the culture and achieve the results [company] needs.”

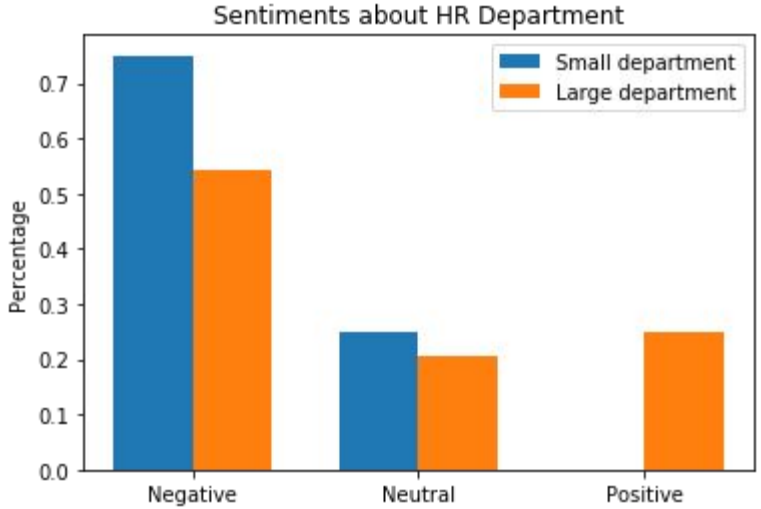
- “CIF training was an eye opener. If we felt that we were working for a good company, the perception changed to we are working for the worlds greatest company. The fact that we can leverage on the network for support and speed is awesome.”

HR department

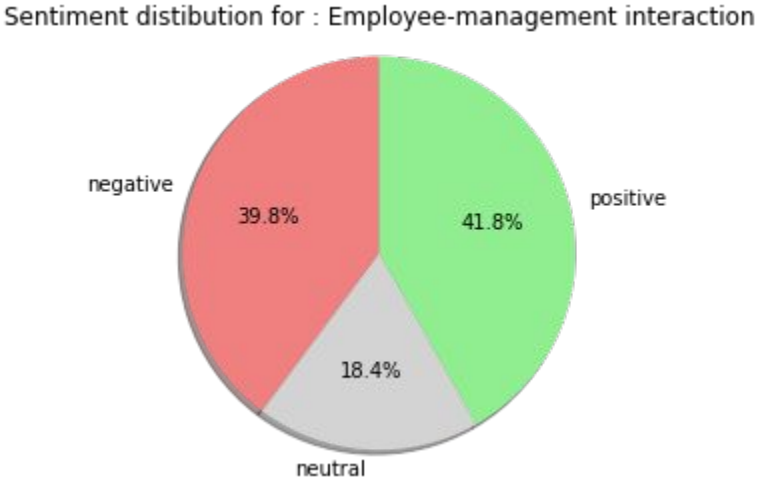


HR Department

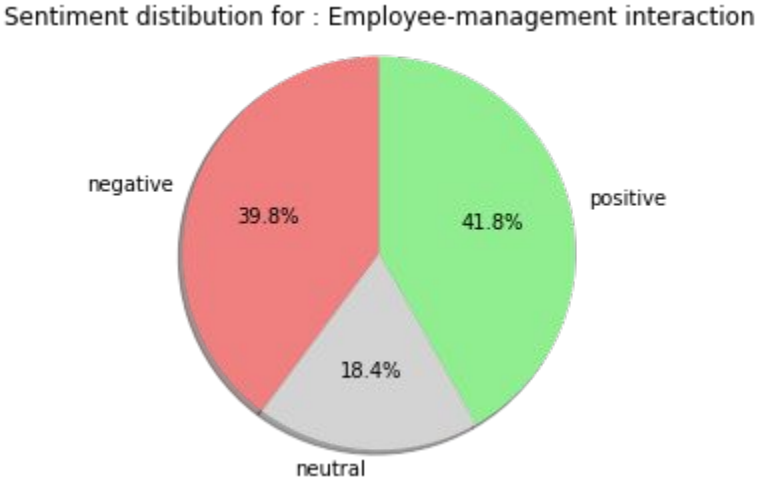




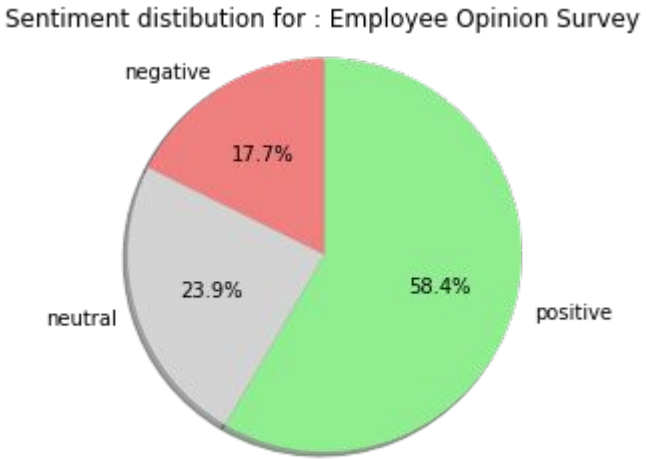
Employee-Management Interaction



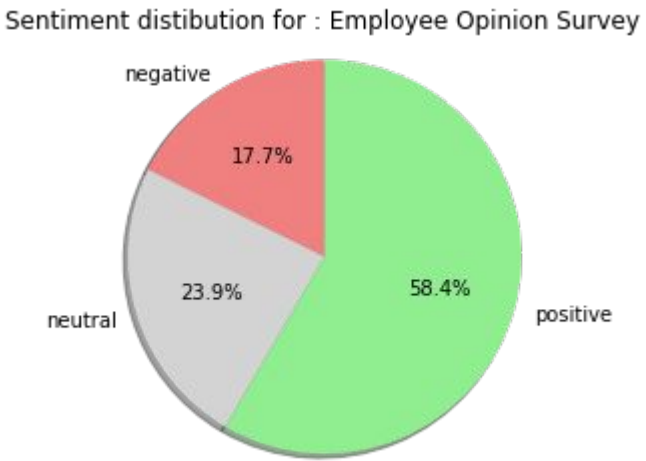
Employee-Management Interaction



Last Year’s Employee Opinion Survey

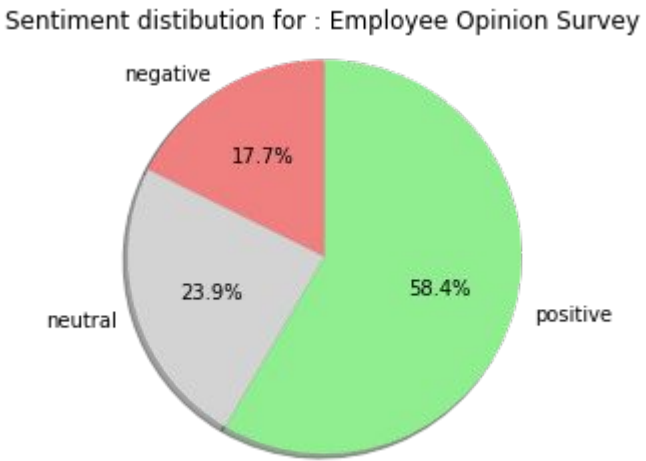


Last Year’s Employee Opinion Survey



- “Based on the last EOS, we have seen many positive changes within the department in terms of employee engagement and also the team bonding has ben [sic] improved.”

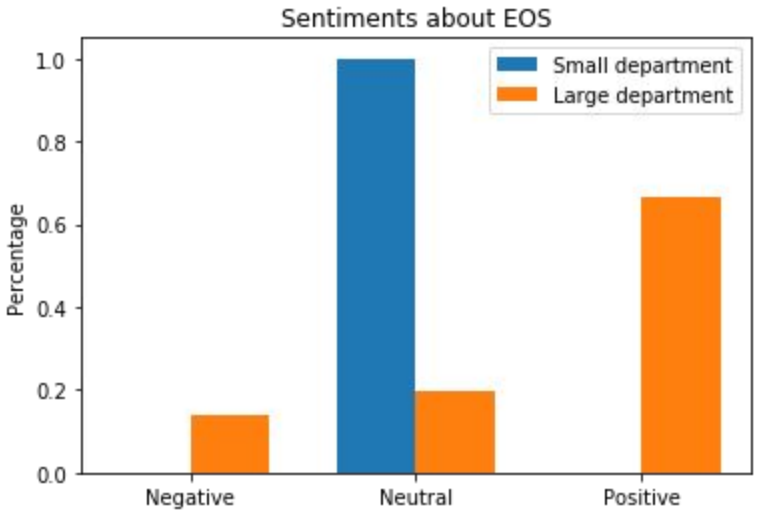
Last Year’s Employee Opinion Survey



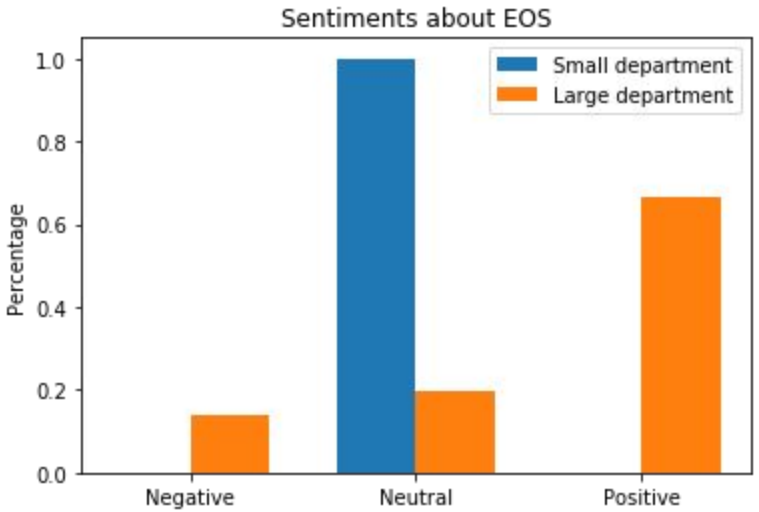
- “Based on the last EOS, we have seen many positive changes within the department in terms of employee engagement and also the team bonding has ben [sic] improved.”

- “I think that the results of the last EOS were addressed with proper attention and changes have already happened because of them. When problems arise the employees/ senior leadership generally handle them effectively and in a timely manner.”

Correlation of Sentiment and Future Engagement?



Correlation of Sentiment and Future Engagement?



Reflection

Implementation

- In some cases the approach works well in grouping comments of similar semantic import.
- Inconsistent cluster 'quality', duplicated topics.
  - More attention to preprocessing and data segmentation.
- Imprecise topic labels.
  - Leverage semantic networks to identify more nuanced relationships between words.
- Lack of metric which evaluates clustering in an intuitive way
- In real world use-case, labeled data for training of sentiment classifier might be available

Project

- (Self-)Organization
  - Task-definition
  - Time-management
  - Structures for knowledge storing

Project

- (Self-)Organization
  - Task-definition
  - Time-management
  - Structures for knowledge storing

- Group Dynamics
  - Varying experience backgrounds
  - Different roles in project
  - Integration of new people for second semester

Project

- (Self-)Organization
  - Task-definition
  - Time-management
  - Structures for knowledge storing
- Group Dynamics
  - Varying experience backgrounds
  - Different roles in project
  - Integration of new people for second semester
- Report in form of a website

Discussion