# Report

This deliverable contains an analysis of a dataset with employee complaints from an organization. Data was downloaded from Kaggle. Firstly, the available data presented. Secondly, an exploratory data analysis was provided to detect patterns and trends across the dataset. Afterward, taking the hypothesis that the Genre as information to represent employee complaints is kind of generic and vague, I suggest a framework to extract more fine-grained topics. I believe that by leveraging the produced topics the HR team can obtain more insights. Hence understand better and in a more accurate way employees' issues. The following framework describes the procedure to identify the fine-grained topics, Figure 1.

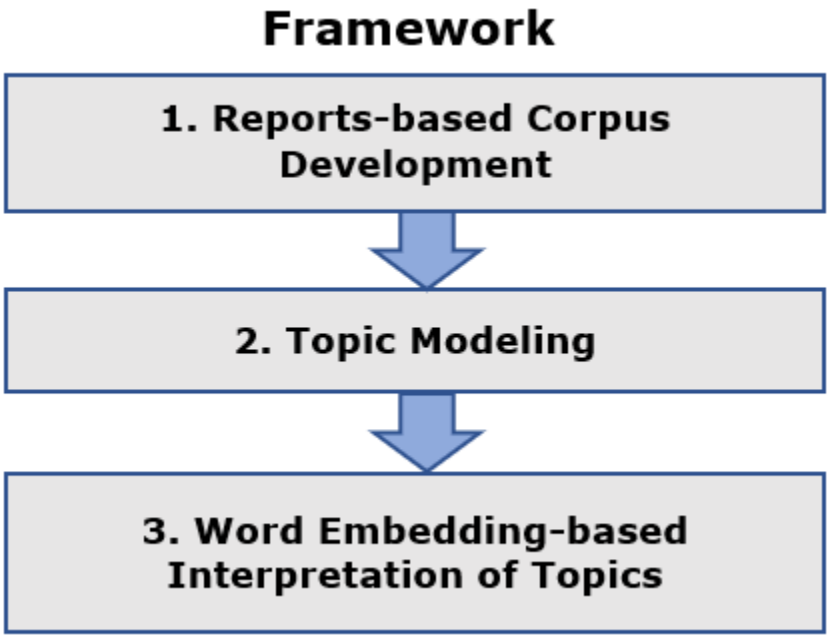



*Figure 1 Proposed Framework for Fine-grained topics*

# Data

## Report

I never receive clear instructions for my tasks.

## Genre:

Communication Issues

Workload and stress

Management Lifestyle

Compensation and benefits

Career Development

Workplace Environment

Lack of training and development

## Employee Role:

Intern

Junior

Senior

Manager

**Age:** employee's age

**Gender:** Female/Male

# Exploratory Data Analysis

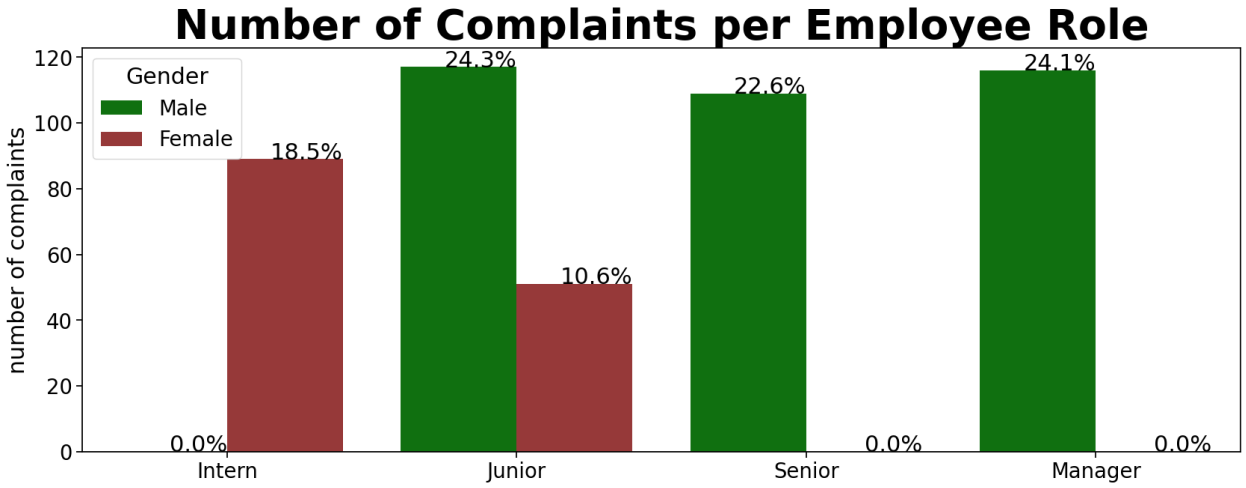



*Figure 2 Number of complaints per Role and gender*

Considering Figure 2 we note that there is a trend showcasing that males complain more than females. Also, we observe that females occur only in Intern and Junior roles. Finally, males represent approximately 71% of employees and 29% which are females.

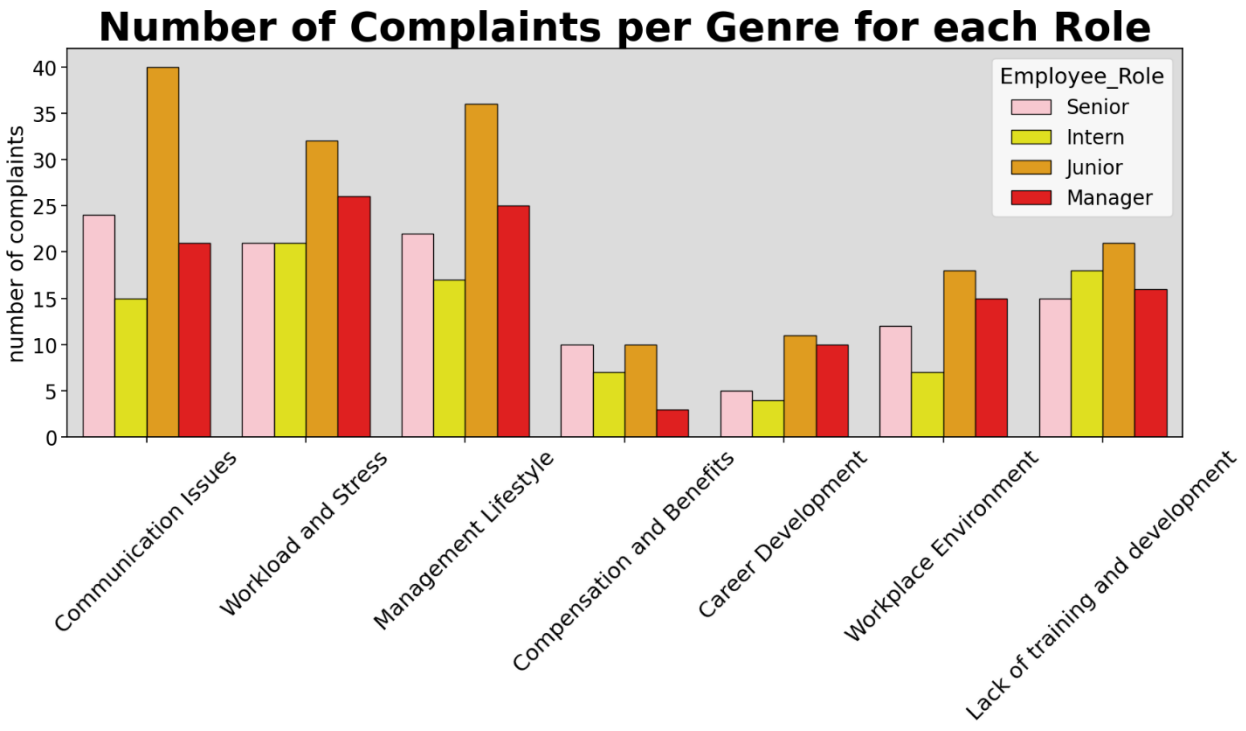



*Figure 3 Number of complaints per Genre and Role*

From Figure 3 we observe that communication issues and workload and stress genre represent most employees. Moreover, for all categories, Juniors complain more than for the other roles. Also, Compensation and Benefits and Career development collect the less listings. Finally, seems that a pattern occurs considering how the complaints are distributed, for the first 3 genres over the roles.

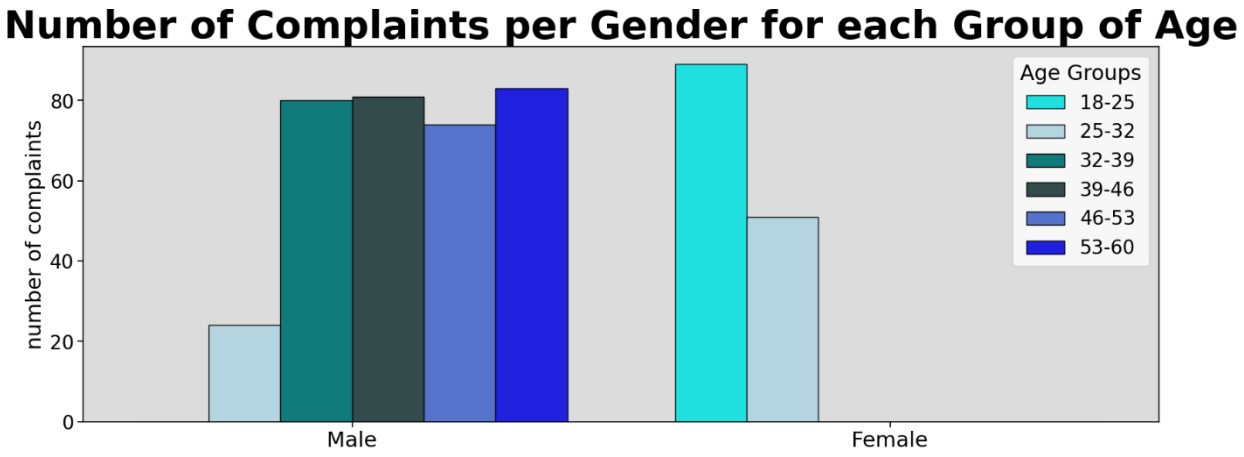


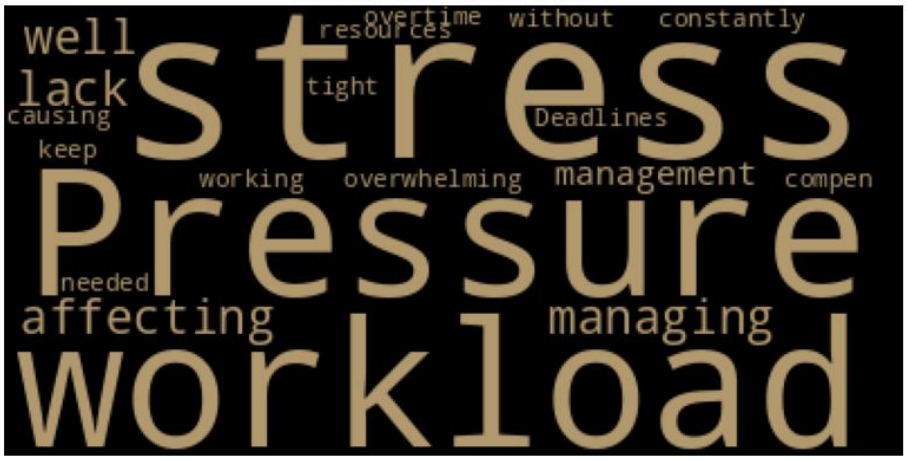

*Figure 4 Number of complaints per gender for each age group*

For the construction of Figure 4 I created a new feature named Age groups. I did that to examine in a more effective way the age as a factor. So, the graph above showcases that females included only the first age group which is rational because from Figure 1 we saw that only females work as interns. Also, it is interesting that from 32 to 60 ages only males occurred.

## Top 20 words based on their frecuency per Genre

Management Lifestyle


Workload and Stress




Lack of training and development


*Figure 5 Wordcloud of Genre of complaints*

Figure 5 is a word cloud graph and through that, we can note per Genre of complaints the top 20 most frequent words. All of the above Genres seem that they describe well the top words. For instance, Compensation and Benefits have as most frequent words the benefits, raises, bonuses, etc.

We move on the proposed framework. We collect all the reports to create the corpus and apply topic modeling.

## Preprocessing of the dataset for LDA

First, for the given corpus, we retrieved the reports and removed English stopwords based on stopwords('englishl'). Then we remove punctuation (e.g. period, comma, apostrophe, quotation, question, exclamation, brackets, braces, parenthesis, dash, hyphen, ellipsis, colon, semicolon), apply lemmatization, keep the tokens contained in at least five documents and convert them to lowercase. Furthermore, we keep only the first 70k most frequent tokens, and keep tokens contained in no more than 0.5 of the total corpus size. We chose the bag of words representation for each document-reports. Therefore, we prepare a corpus with words from reports and a dictionary containing each word's frequency. Those two inputs are necessary to start the topic modeling process.

## Tuning Number of Topics

In this part, we fine-tune the number of topics for our model, using the implementation by Gensim. The number of topics is the most crucial parameter to consider. We want to set the respective parameter with a value in which we achieve the minimum coherence score. So, as the tuning process runs for each number of topics, a coherence score results for every topic. Then we took the average coherence score over the topics and chose the model with the respective value. We set the values for the number of topics to test from 2 to 60 with step 2. In Figure6, we observe the curve of average coherence score over the number of topics. We used the elbow method to fit the best parameter value in our model. First, we fine-tune the number of topics considering the average coherence measure and obtain the number of topics leveraging the elbow method. Coherence measure scores a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements help distinguish between semantically interpretable topics and topics that are artifacts of statistical inference.

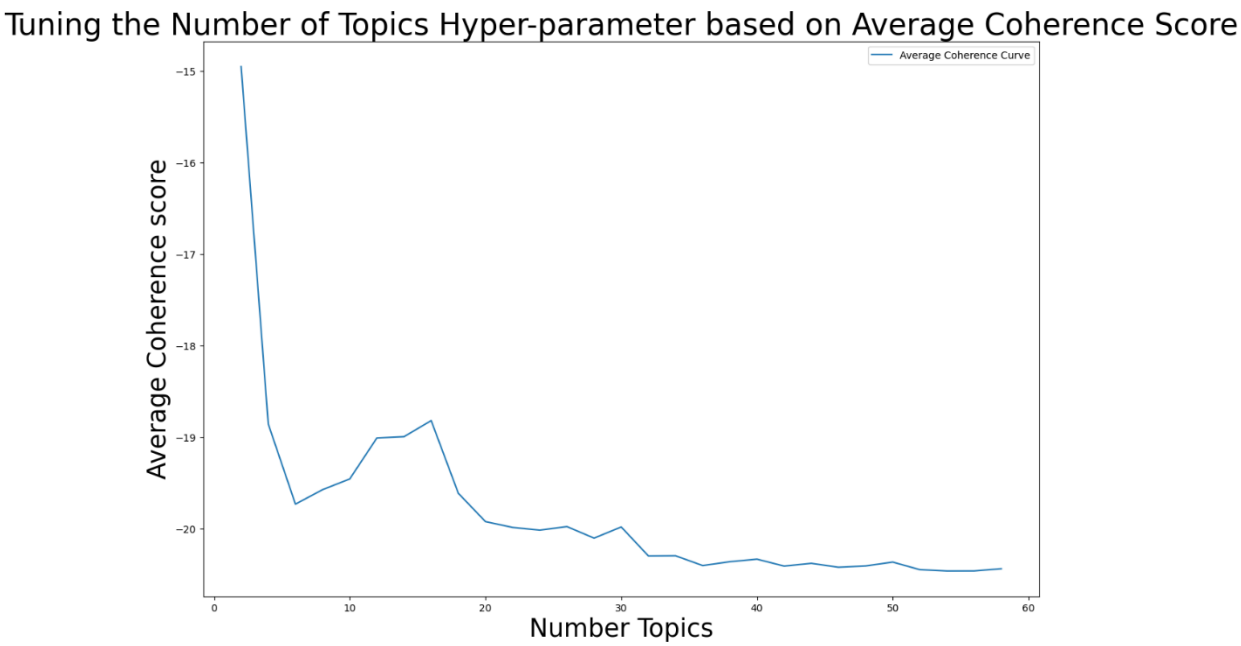



*Figure 6 Tuning number of topics relied on avg coherence and apply elbow method to retrieve best value*

# Model Visualization of LDA

At this point, we are ready to move on to the next step for the suggested Framework, in Figure 1. We apply topic modeling on the Report-based corpus and with the best hyper-parameter retrieved from tuning.

The next step is to examine the produced topics and the associated keywords. Figure 7 showcases a snapshot from the interactive plot. We execute the model by setting the number of topics equal to six. Each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent that topic is. A good topic model will have fairly big, non-overlapping bubbles scattered throughout the chart instead of clustered in one quadrant. A model with too many topics will typically have many overlaps, small-sized bubbles clustered in one region of the chart. Finally, moving the cursor over one of the bubbles, the words and bars on the right-hand side will update. These words are the salient keywords that form the selected topic.

## Schematic of LDA using pyLDAvis

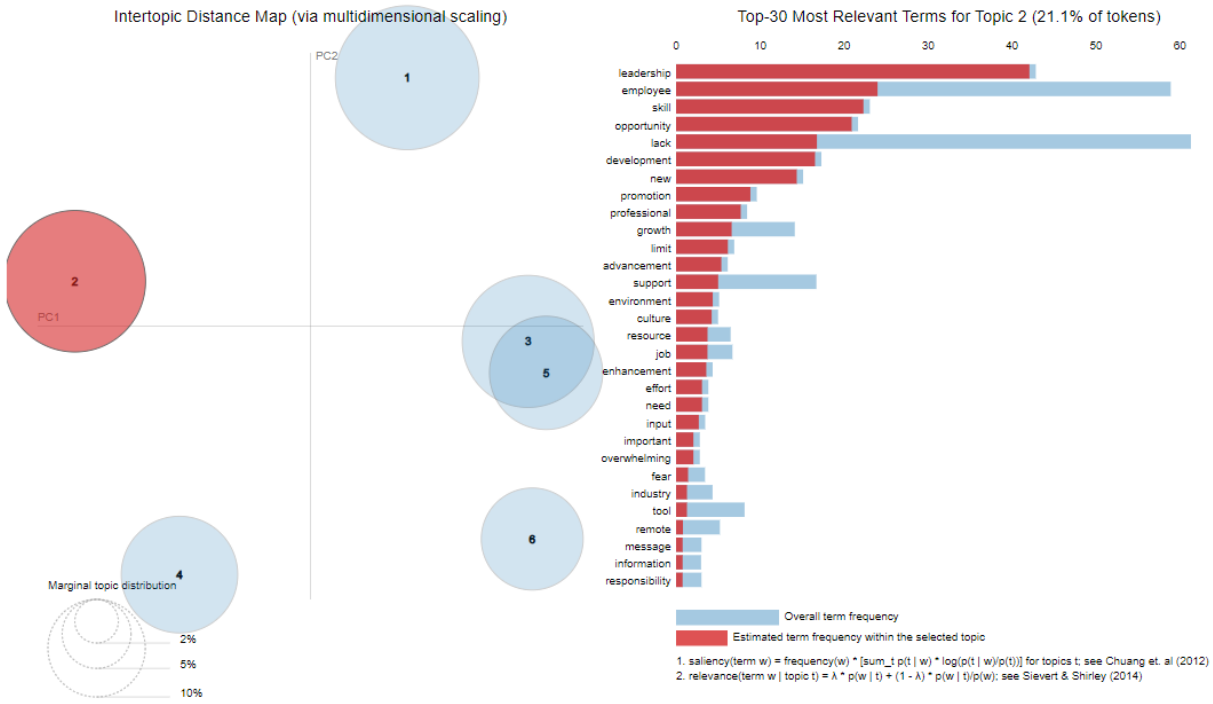



*Figure 7 Topics Visualization using pyLDAvis from Gensim*

We observe that topics 3 and 5 overlap with others, because they share similar words.

# Strategy

Moving on to the next and final step, Word Embedding-based Interpretation of Topics, we present the strategy below.

For the following steps we rely on the H.Gonen et al. (2021)] approach for the analysis. We build a word embedding space produced by the employee reports.

We build the word embedding spaces in order to hold the semantic properties. And the goal is to generate a representative word for each topic, by calculating the centroid over the top words of each topic.

The next step is to calculate the centroid for each topic and qualify a representative word respectively. The centroid for each topic is calculated by the top ten words of each topic. We select the most probable word closer to the centroid's coordinates. Then we fetch the ten closest words-neighbors for our centroids-representatives and observe them at the word embedding space.

More intuitively, we apply the following steps:

- **i.** We build word embedding spaces produced by the corpus from reports
- **ii.** We calculate the centroid for each topic and qualify a representative word, respectively.
- **iii.** We fetched the ten closest words-neighbors for our centroids-representatives and depicted them.

# Word Embeddings Space

**(Step-i)** To build the word embedding spaces, we use Word2vec implemented by Gensim. For all our experiments, we set the vector dimension to 100, the window size to 5, and the minimum number of occurrences of a word to 3. The rest of the hyperparameters are set to their default value.

# Centroids-Representatives

At this stage of analysis, we calculate the centroid for each topic to qualify a representative word for each topic **(step-ii)**. Hence, ***Table 1*** showcases the final representatives for each topic. Moreover, these representative words represent the six topics from the previous analysis. Therefore, for Topic 1, we calculate as a centroid word "Leadership", so this word becomes the representative for Topic 1. Respectively the "workload" represents Topic 2, and so forth. It is worth reminding that the calculation of the centroid returns coordinates over the word embedding space. Then we fetch the most similar word according to the cosine similarity, closest to the centroid coordinates.

| Topics | Centroid word |
|---|---|
| Topic 1 | Leadership |
| Topic 2 | Workload |
| Topic 3 | Benefit |
| Topic 4 | Supervisor |
| Topic 5 | Feedback |

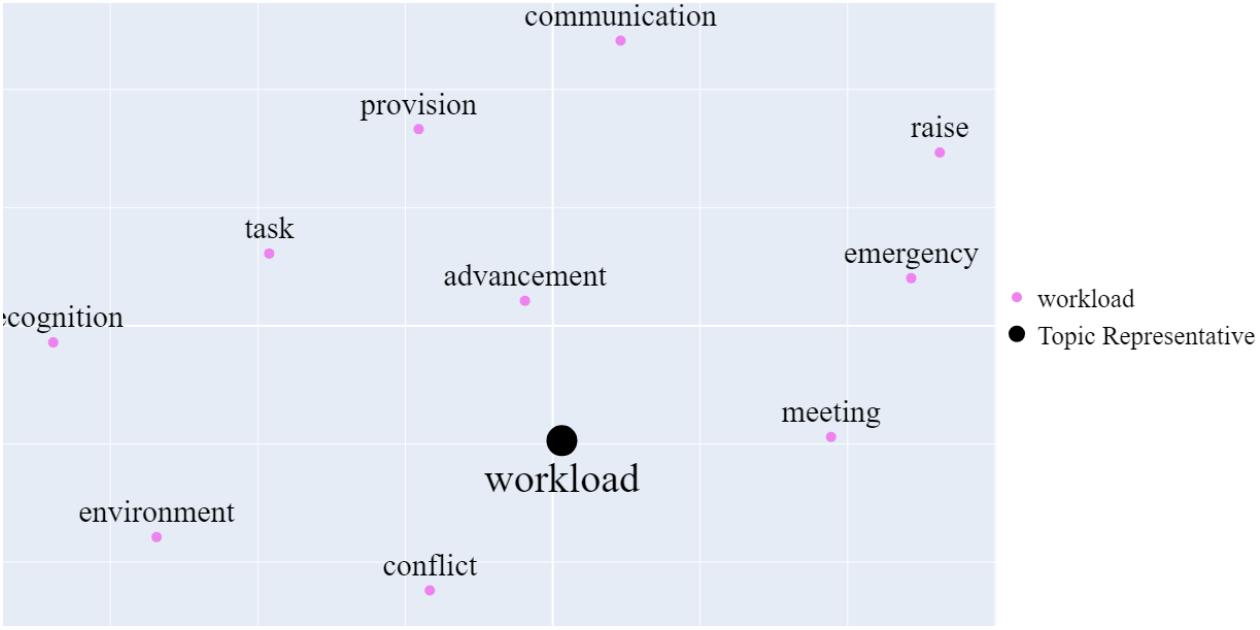
*Table 1 Centroid-Representatives for each Topic*

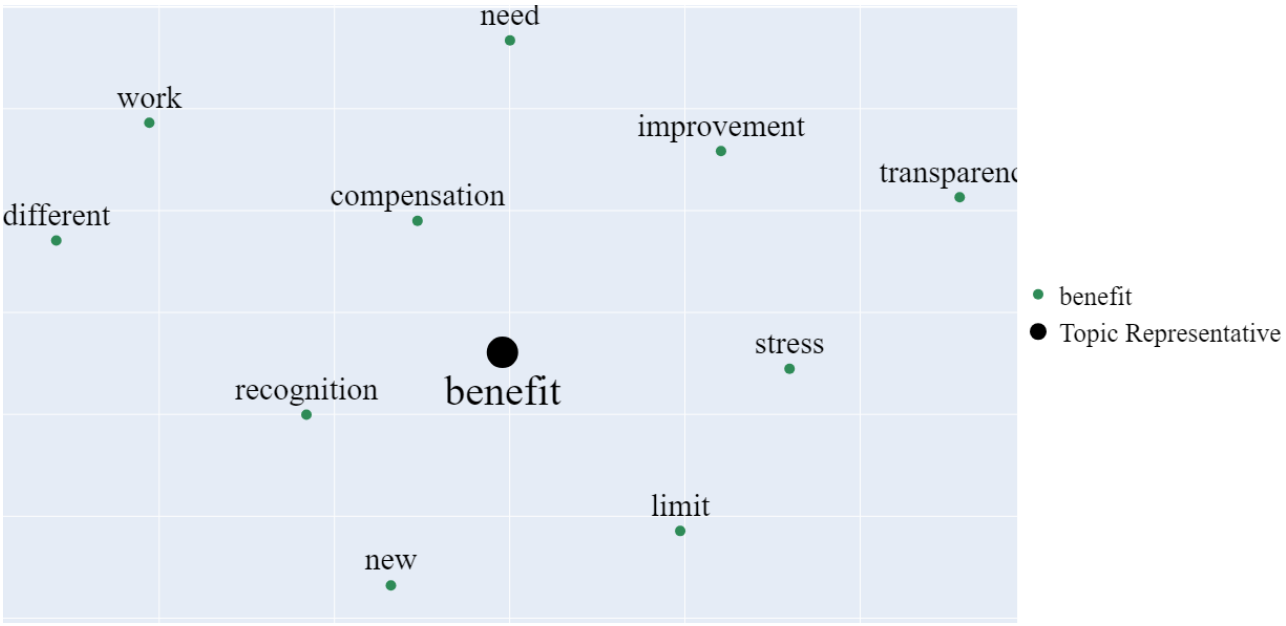
Considering the table above we retrieved 5 cendroids while we had 6 topics. That means that the 2 topics had similar content hence they got merged.

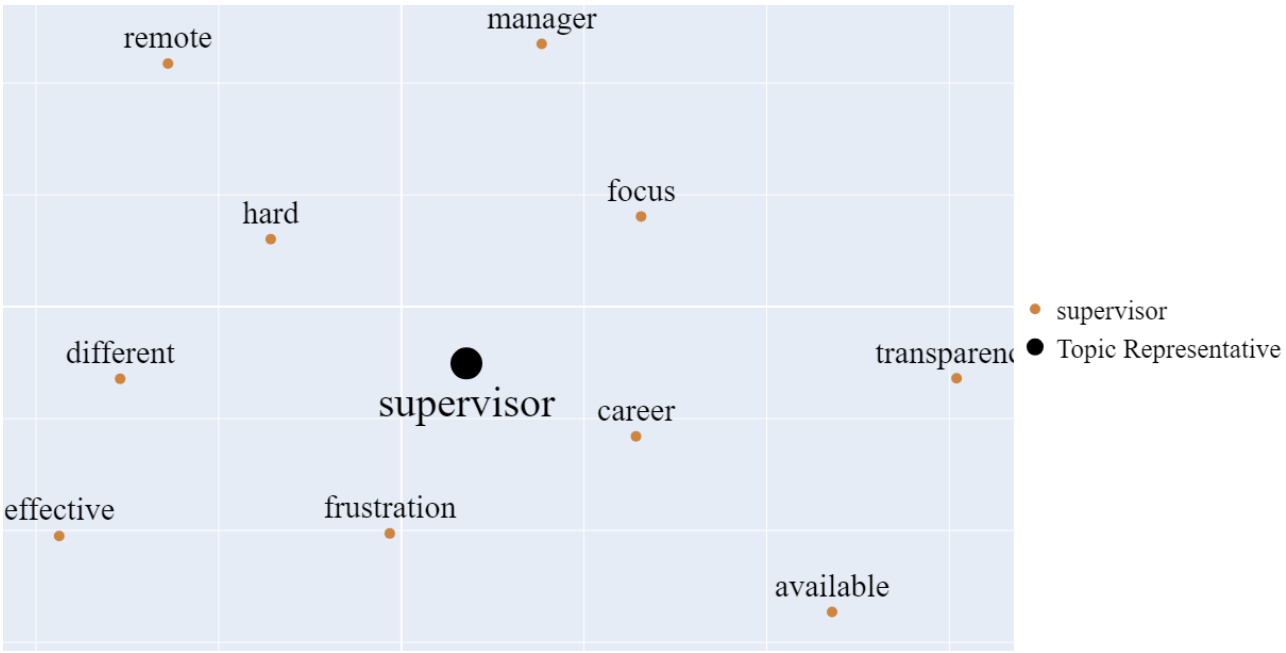
# Topic Representatives Concerning Word Neighbors

**(step-iii)** In order to deeply examine and understand the topics of representatives and their neighbors we provide the depiction below Figure 8. We note the words with the biggest size and dot as representatives for each topic. The dot size is bigger to make it more apparent to the reader that this word represents each topic. The ten nearest neighbors are the words that accompany the representatives.


rumor
leader
available
promotion
communicat
overwhelming
advancement
leadership
task
avenue
stress
leadership
Topic Representative


communication
provision
raise
task
emergency
advancement
ecognition
meeting
workload
environment
conflict
workload
Topic Representative


need
work
improvement
transparenc
different
compensation
stress
recognition
benefit
limit
new
benefit
Topic Representative


remote
manager
focus
hard
different
transparenc
supervisor
career
effective
frustration
available
supervisor
Topic Representative

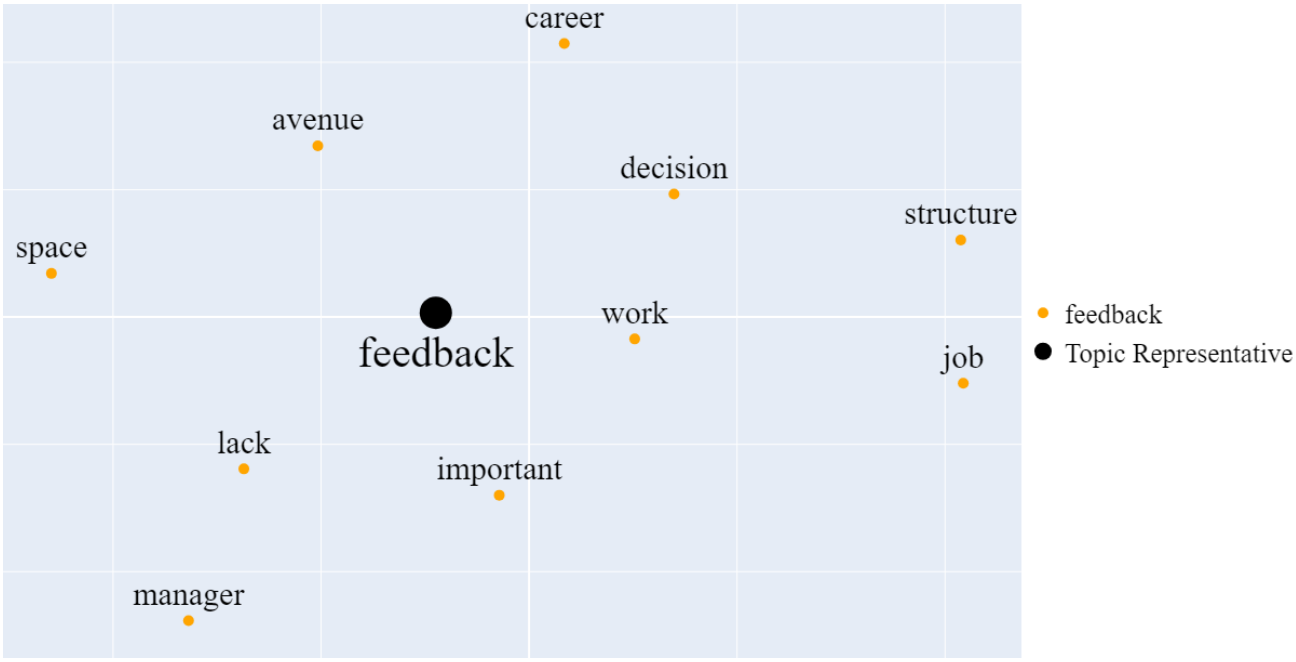



*Figure 8 Topic Representative words with their 10 closest nearest word-neighbors*

From the above graph, we understand that employees complain for the following reasons.

From the first topic representative, we realize that the communication with leadership and the tasks assigned by leadership are reasons for complaining. Also, the fact that employees aren't promoted also identifies the above.

Now considering the workload topic representative, meetings that probably overwhelm the employees, and also the fact that there is not recognition for the amount of workload leads to complaints.

To conclude for the last topic representative, feedback words like lack, manager, and structure indicate that employees face problems with these factors in their work environment. Similar analysis could developed on topic representatives 3 and 4.

# Results

Taking the above results into account, I believe we the produced topics have high interpretation, and considering a scenario in which we would have no idea about the Genre of complaints, imagine how useful the above framework could be. It is evident that the evaluation between the initial Genre and the produced topics was evaluated qualitatively and that is why the resulting topics were estimated as quality and representative.