Evaluating Text Similarity Techniques for Matching Personal Employee Objectives

Mohab Ghanem, 11.09.2020 – Guided Research Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

1) Problem Statement
2) Solution Approach
3) Deliverables
4) Evaluation

Outline

1) **Problem Statement**
2) Solution Approach
3) Deliverables
4) Evaluation

Problem Statement

MERCK is a multinational company with many offices worldwide.

The company employs thousands of employees spread all over the globe.

Each employee writes a yearly objective.

The HR department wants to connect employees with similar objectives in order to cooperate.

Problem Statement

⚠️ This process happens every year.

⚠️ With thousands of employees.

✅ The solution is to **automate** the process of **finding work partners with similar objectives.**

Outline

1) Problem Statement
2) **Solution Approach**
3) Deliverables
4) Evaluation

Outline

1) Problem Statement

2) **Solution Approach**

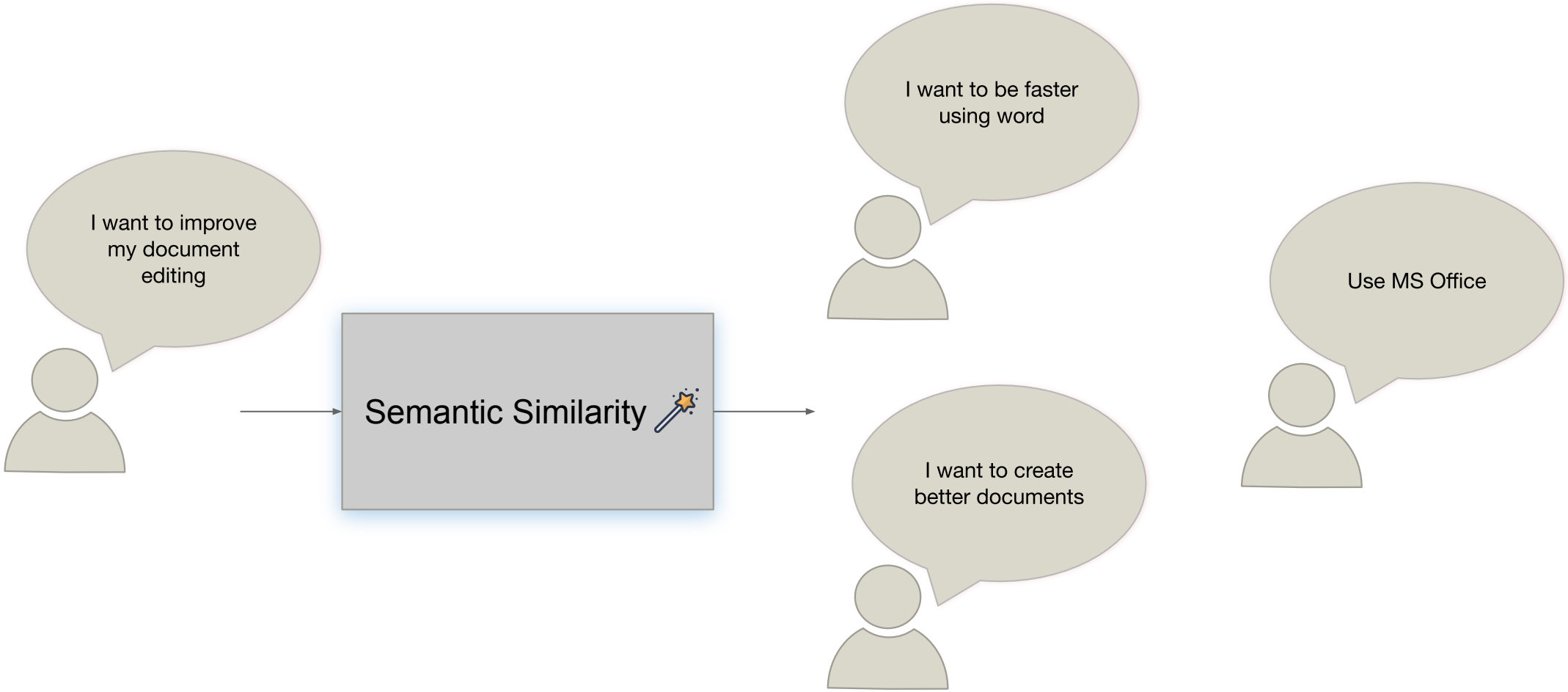
Idea Models Data

3) Deliverables

4) Evaluation

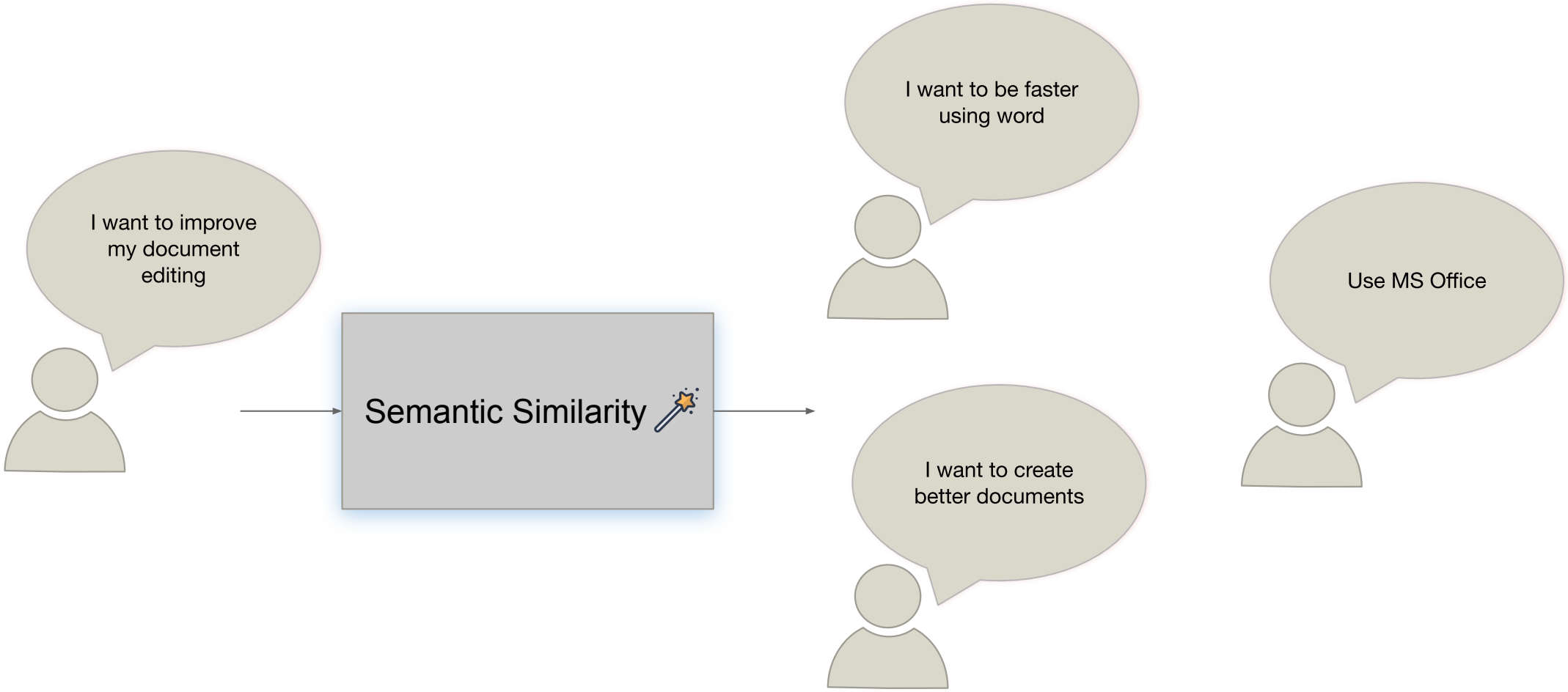
Solution Approach - Idea

Use **semantic similarity** of objectives to find the most similar N employees.



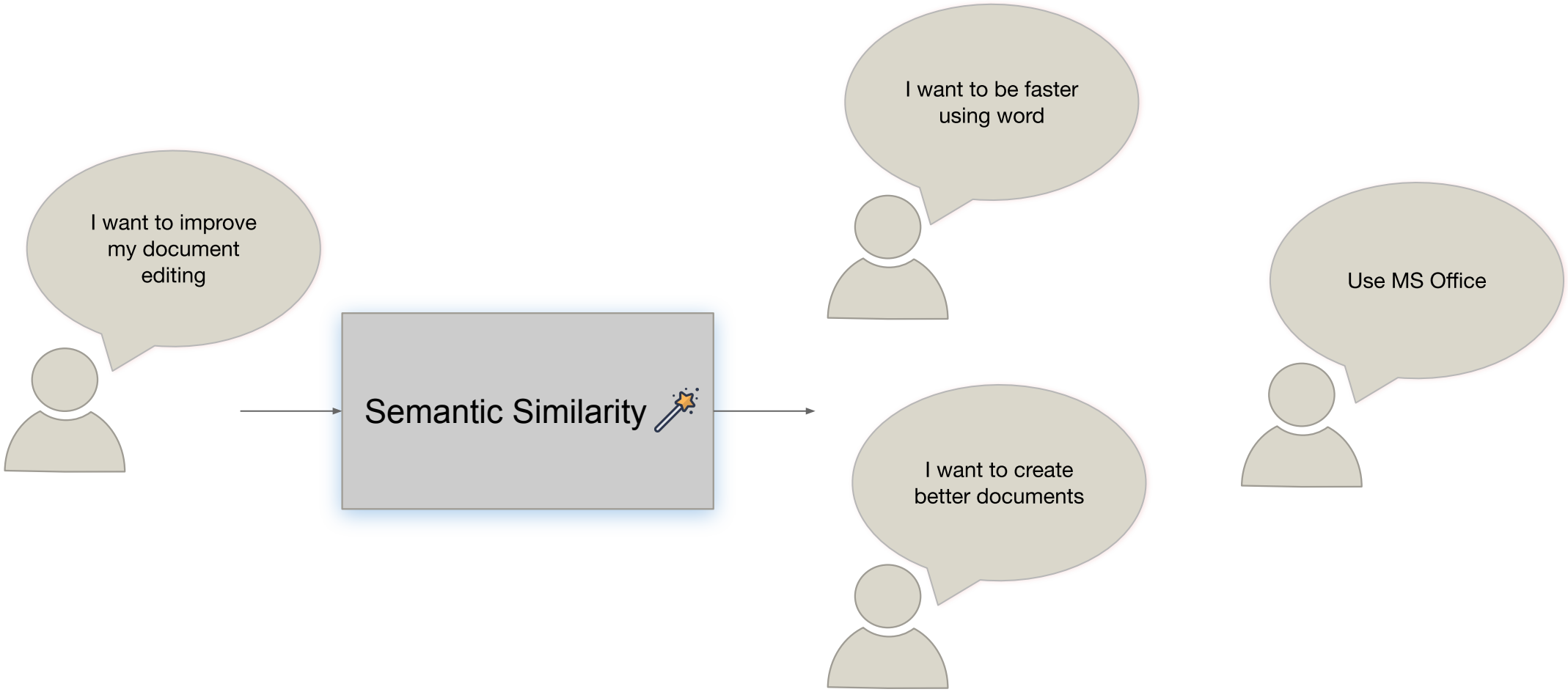
Solution Approach - Idea

How to estimate **semantic similarity**?



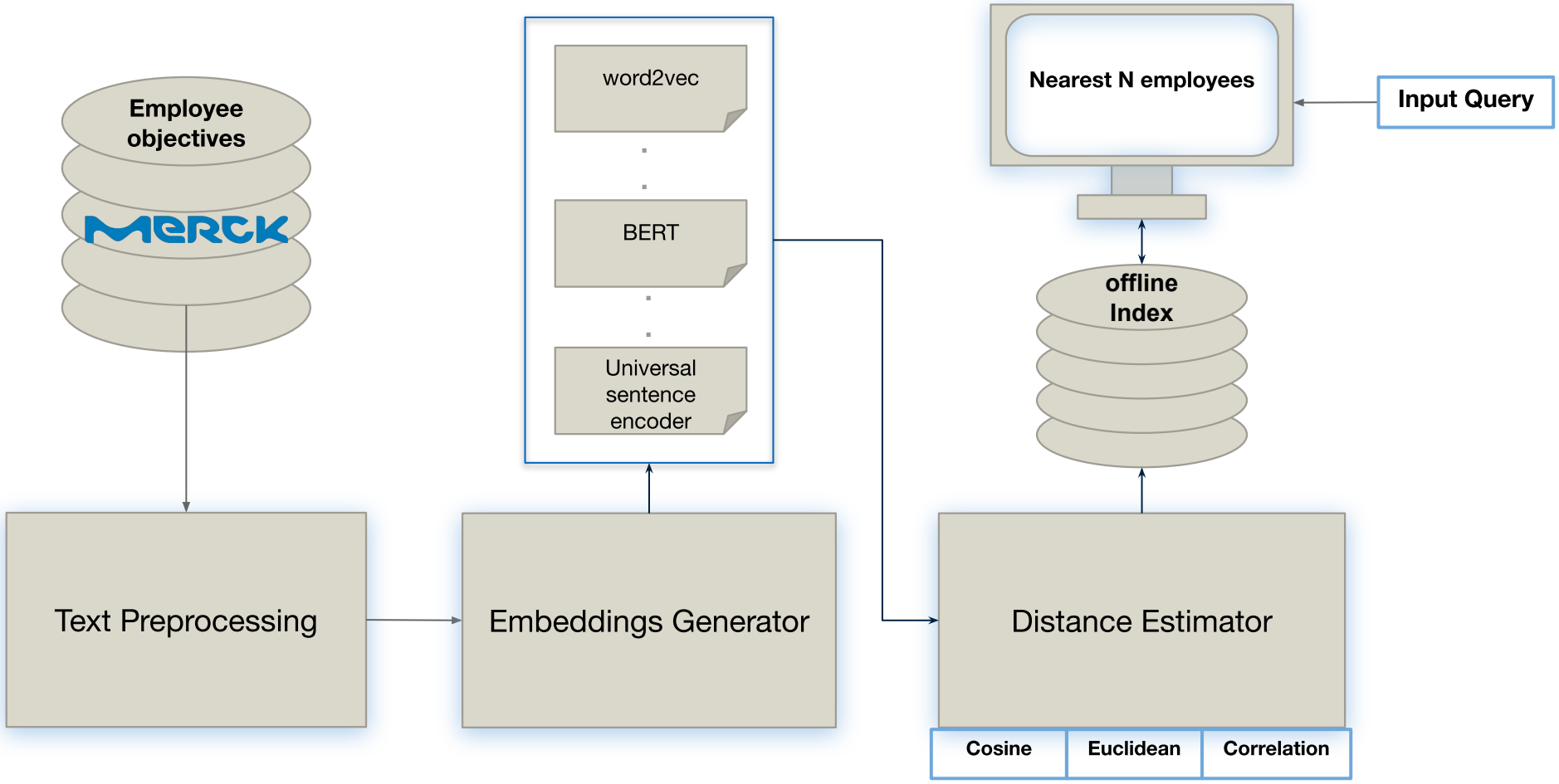
Solution Approach - Idea

How to estimate **semantic similarity**? **Pre-trained Word Embeddings**



Solution Approach - Idea

General Program flow, from data to user interface



Outline

1) Problem Statement

2) **Solution Approach**

❗ Idea ❗ **Models** ❗ Data

3) Deliverables

4) Evaluation

Solution Approach - Models

Text embeddings:

⚠️ **Word Based:**

- Word2vec
- Elmo
- BERT
- ...

⚠️ **Sentence Based:**

- Universal Sentence Encoder
- Sentence Transformers

Solution Approach - Models

Text embeddings:

⚠️ **Word Based:**

- Word2vec
- Elmo
- BERT
- ...

⚠️ **Sentence Based:**

- Universal Sentence Encoder
- Sentence Transformers

⚙️ Average to get sentence embeddings ⚙️

Solution Approach - Models

Text embeddings:

⚠️ **Word Based:**

- Word2vec
- Elmo
- BERT
- ...

⚠️ **Sentence Based:**

- Universal Sentence Encoder
- Sentence Transformers

⚙️ Average to get sentence embeddings ⚙️

✅ Out of the box sentence embeddings ✅

Solution Approach - Models

Text embeddings:

⚠️ **Word Based:**

- Word2vec
- Elmo
- BERT
- ...

⚙️ Average to get sentence embeddings ⚙️

⚠️ **Sentence Based:**

- Universal Sentence Encoder

⭐ Sentence Transformers ⭐

✅ Out of the box sentence embeddings ✅

Solution Approach - Models

How to calculate similarity between the generated embeddings?

Solution Approach - Models

How to calculate similarity between the generated embeddings?

Distance Metrics:

- Cosine Distance.
- Euclidean Distance.
- Correlation.
- City Block Distance.
- ...

Solution Approach - Models

⚠️ How to calculate similarity between the generated embeddings?

✅ Distance Metrics:

⭐ Cosine Distance. ⭐

- Euclidean Distance.
- Correlation.
- City Block Distance.
- ...

Outline

1) Problem Statement

2) **Solution Approach**

❗ Idea ❗ Models ❗ **Data**

3) Deliverables

4) Evaluation

Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K**

  records for historical data from 2018 and 2019

Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K**

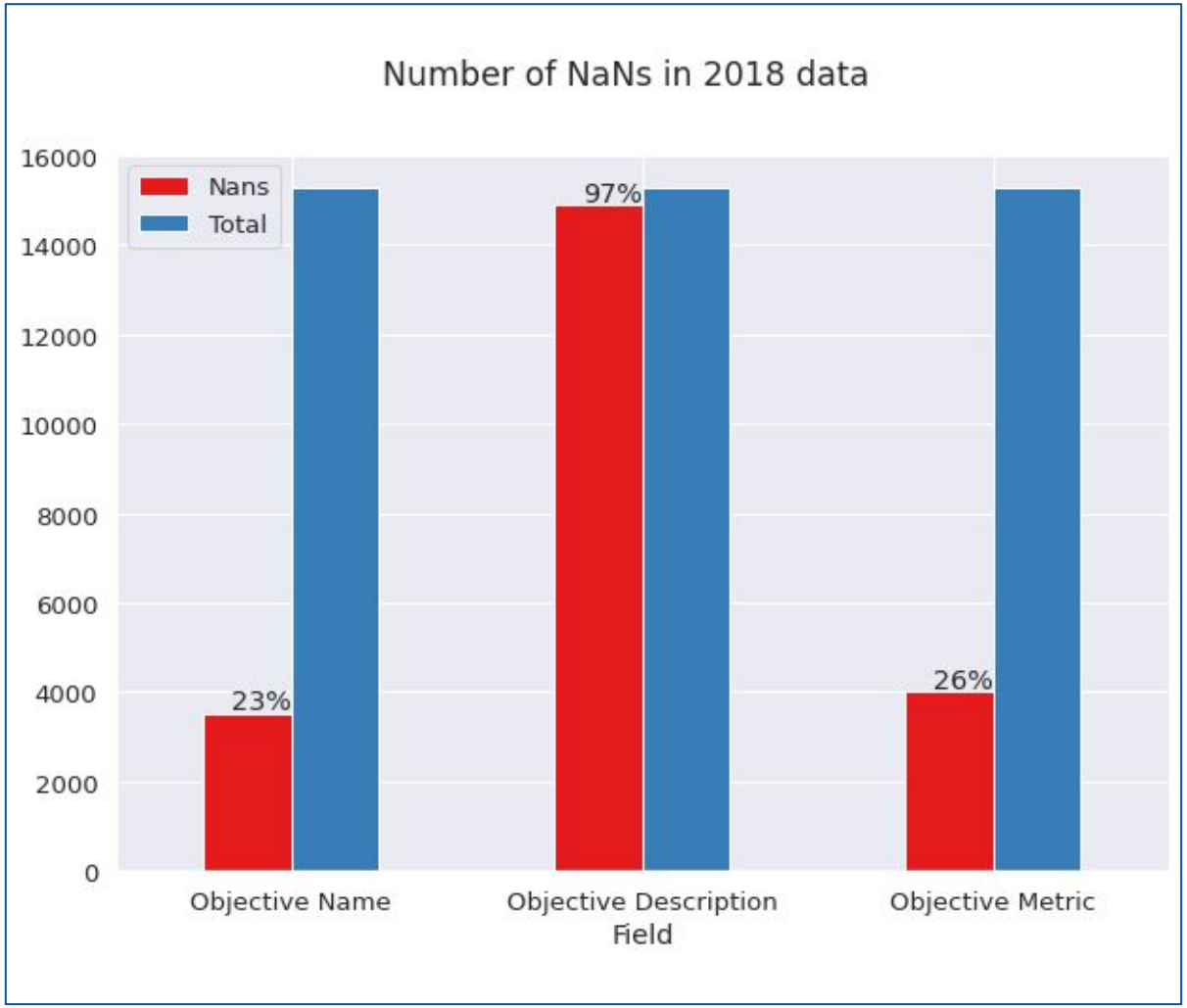
  records for historical data from 2018 and 2019

User ID	Global Key	Functional Area	Location	Objective Name	Objective Description	Objective Comment	Objective Metric	Form Template Name
*100002	HR	Human Resources	Burlington	Lead Change and Comms workstream. Develop and drive change & comms plan for EC globally			Completion of all milestones as per defined project timeline & roadmap. Satisfied stakeholder feedback at end of the project/YEE	2019 Performance Management

Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns

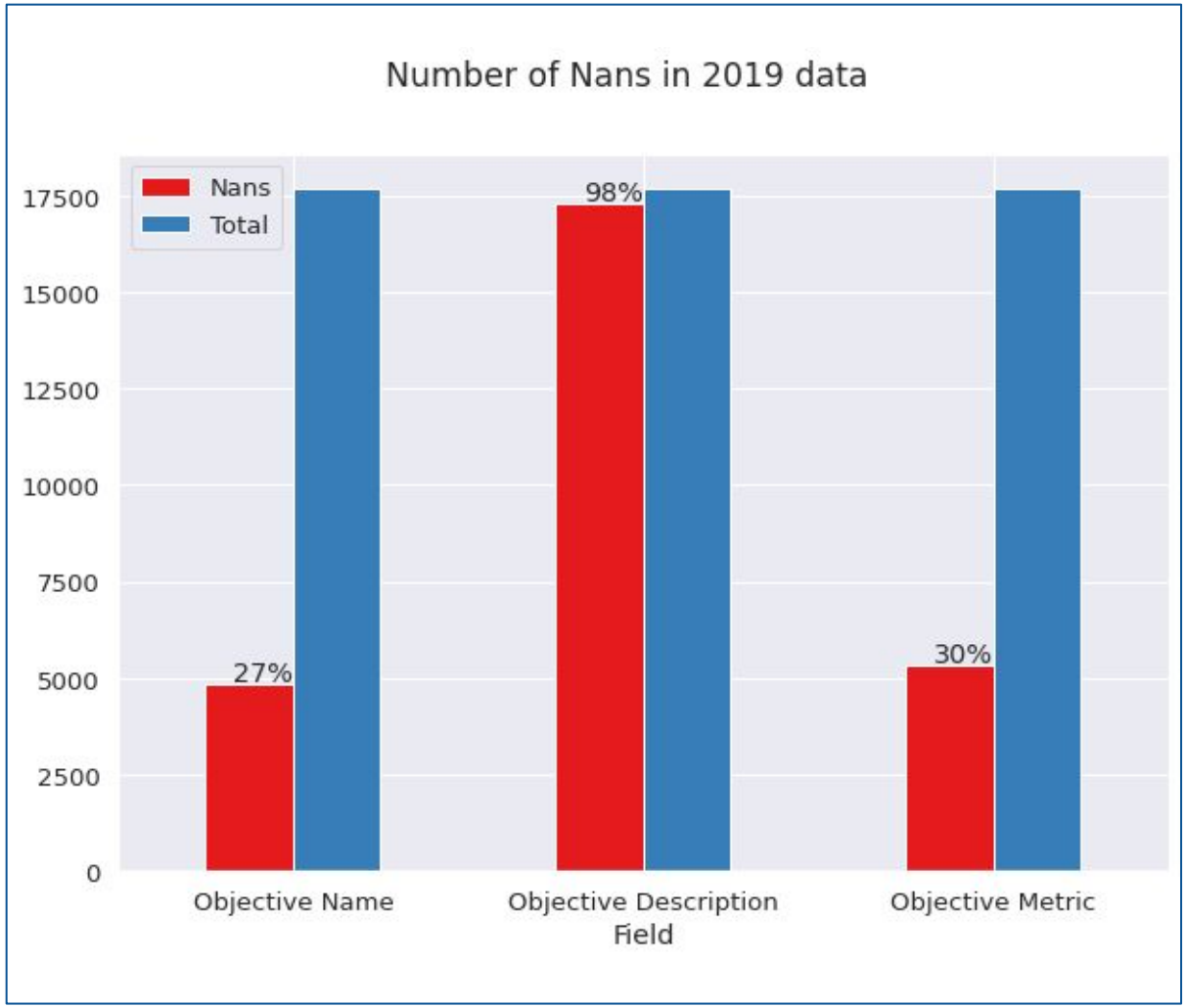


Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019

⚠ Objectives were spread across multiple columns



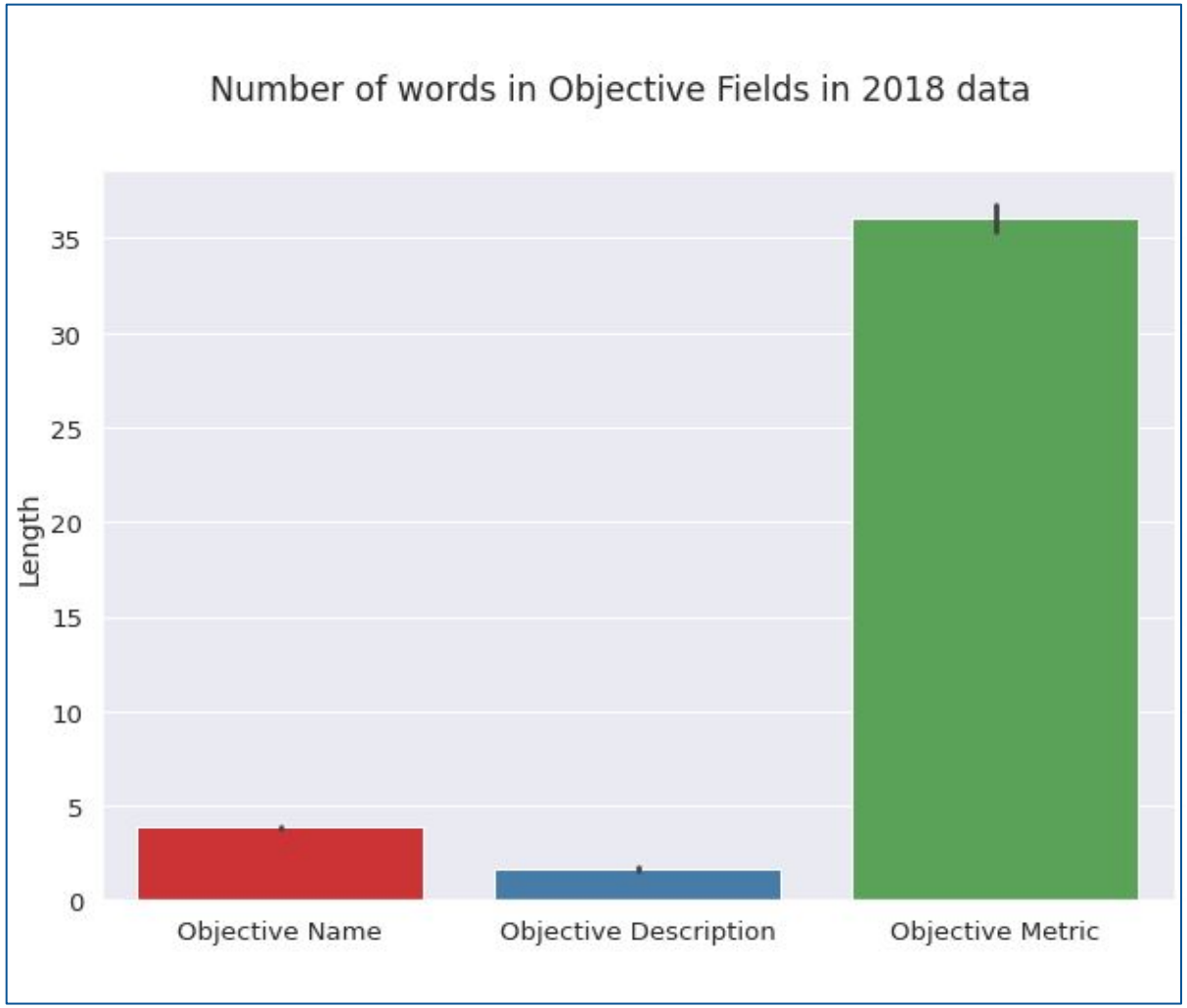
Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019

⚠ Objectives were spread across multiple columns

⚠ Length of objectives varies per column

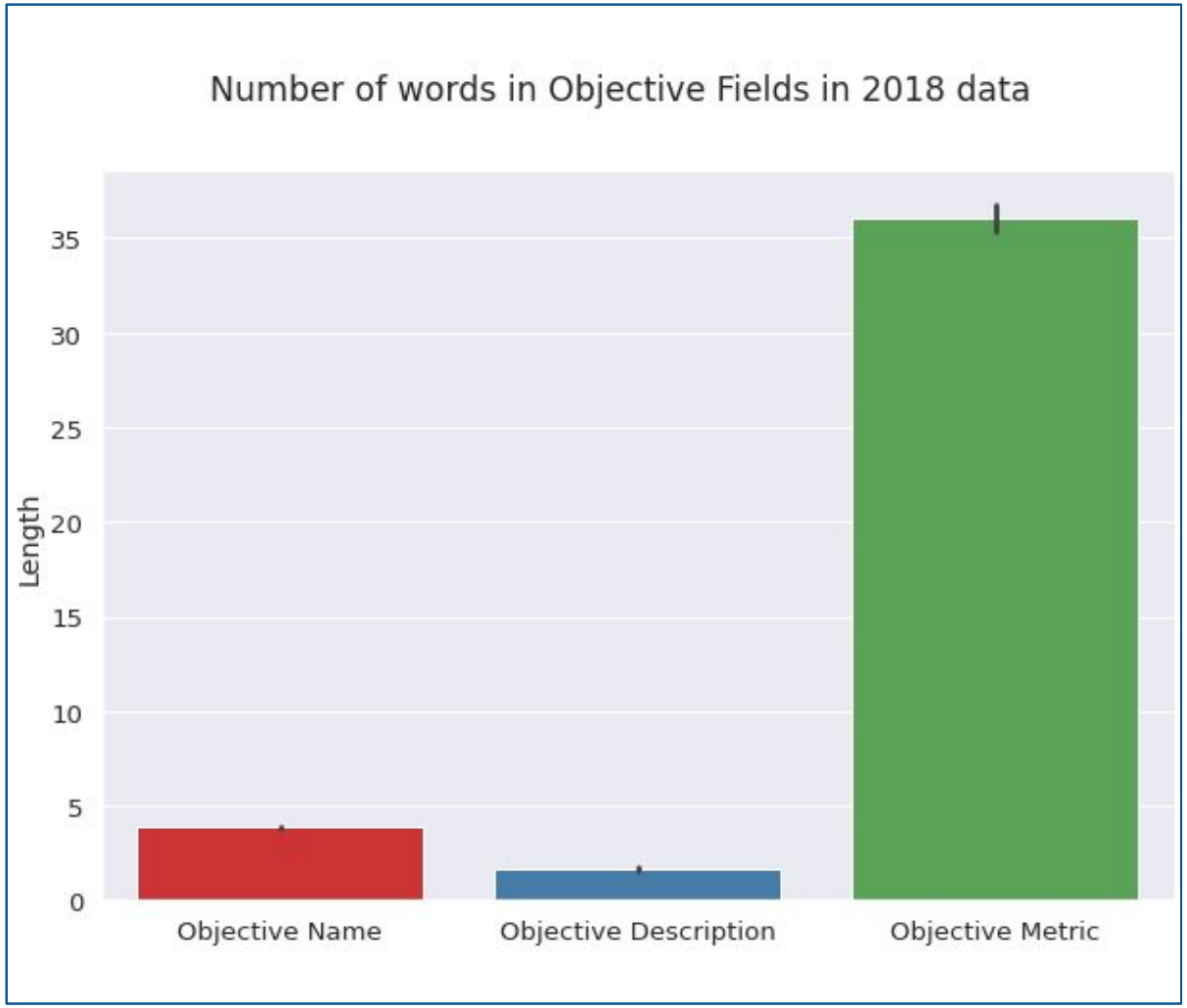


Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns
- Length of objectives varies per column

Combine into one column



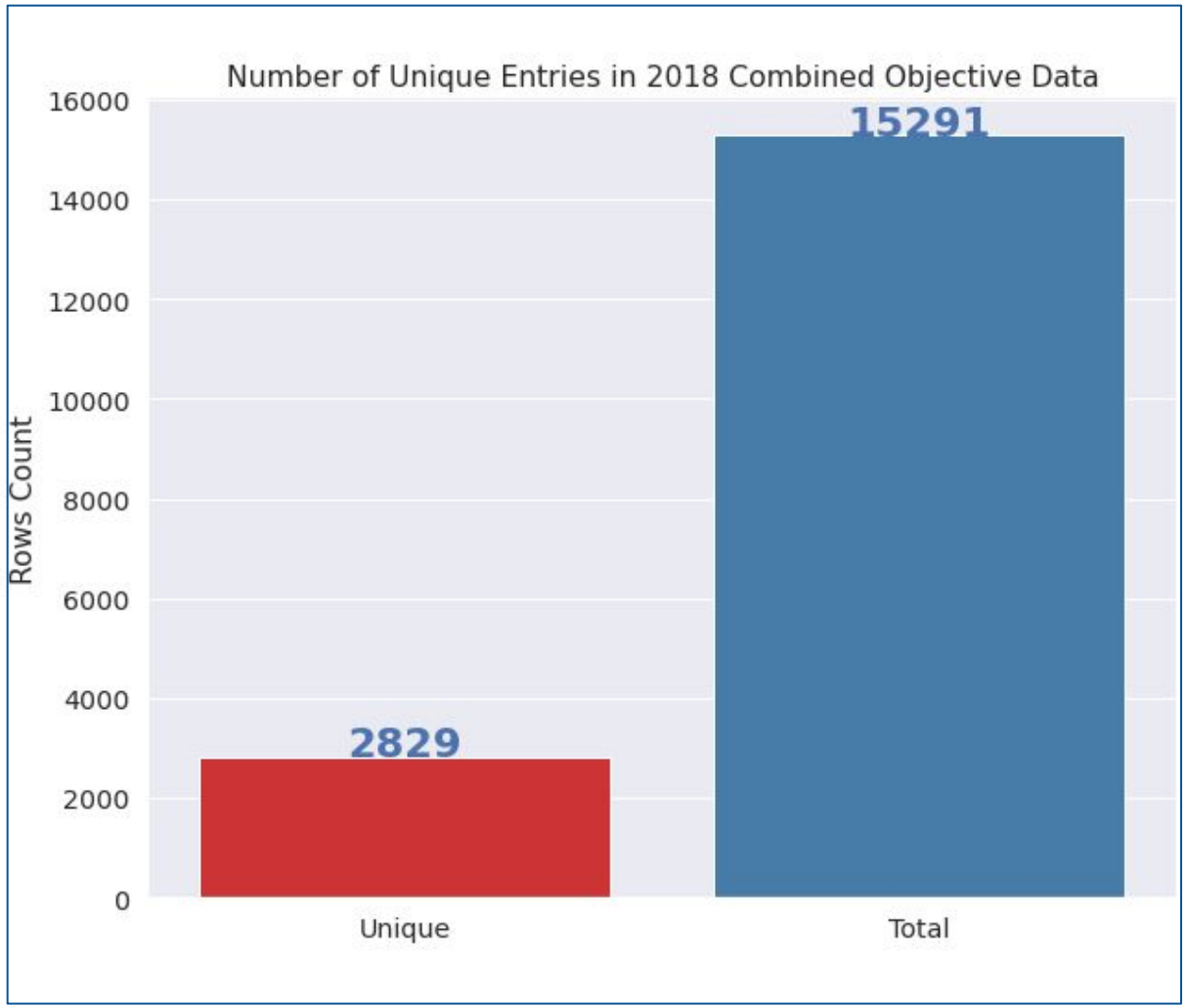
Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns
- Length of objectives varies per column

Combine into one column

Data contains too much duplication



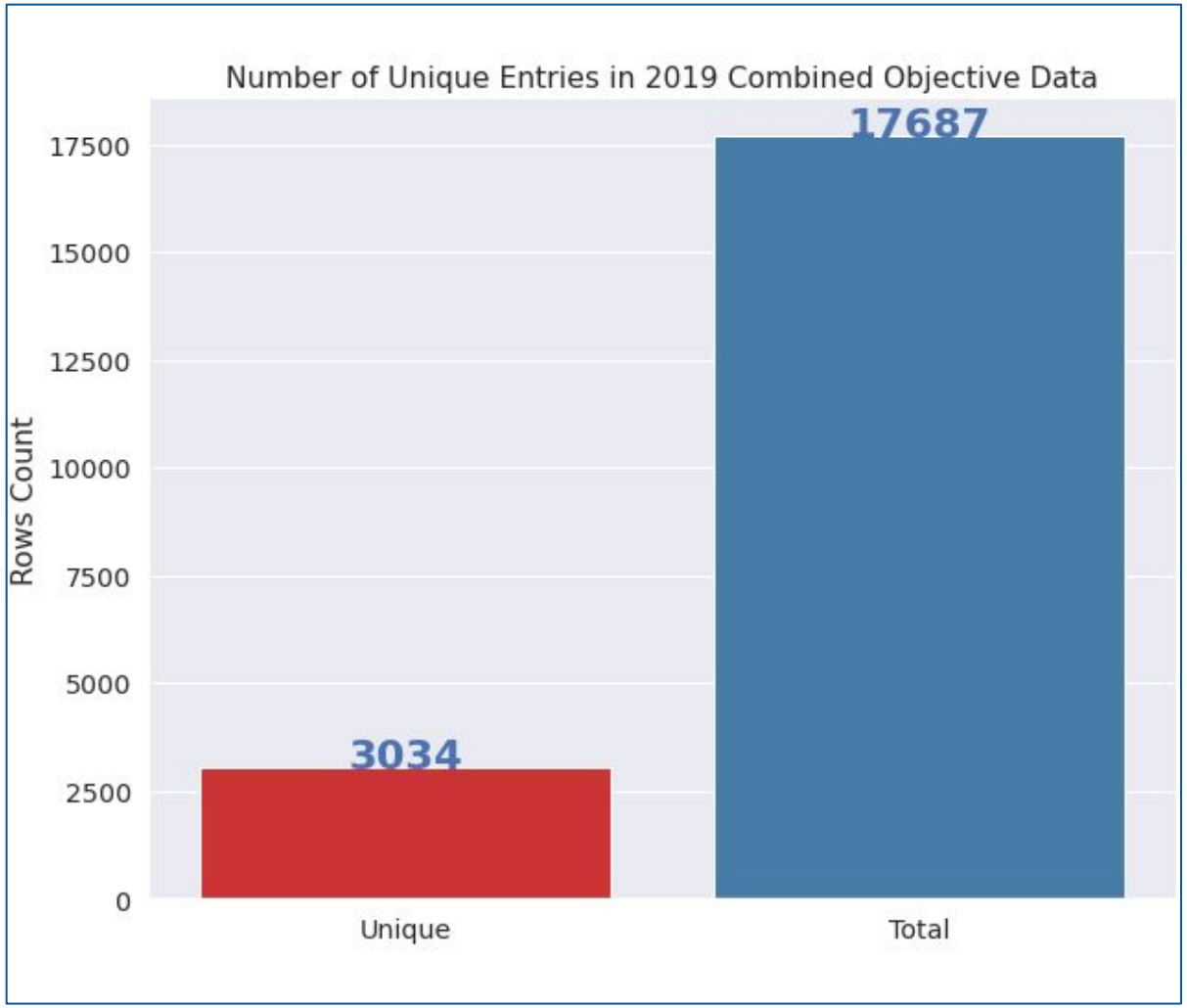
Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns
- Length of objectives varies per column

Combine into one column

Data contains too much duplication



Solution Approach - Data

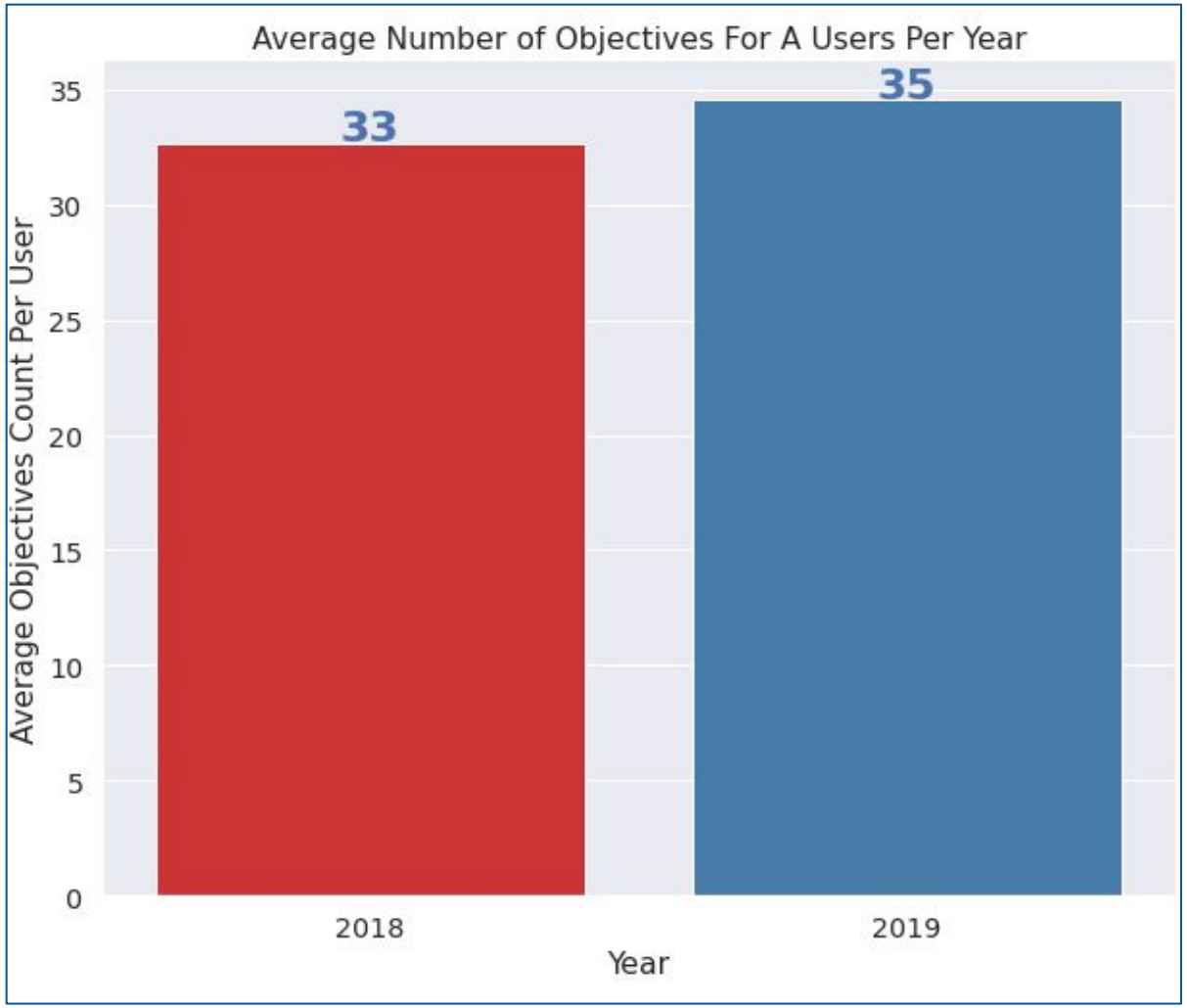
Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns
- Length of objectives varies per column

Combine into one column

Data contains too much duplication

A user has on average 35 objectives!



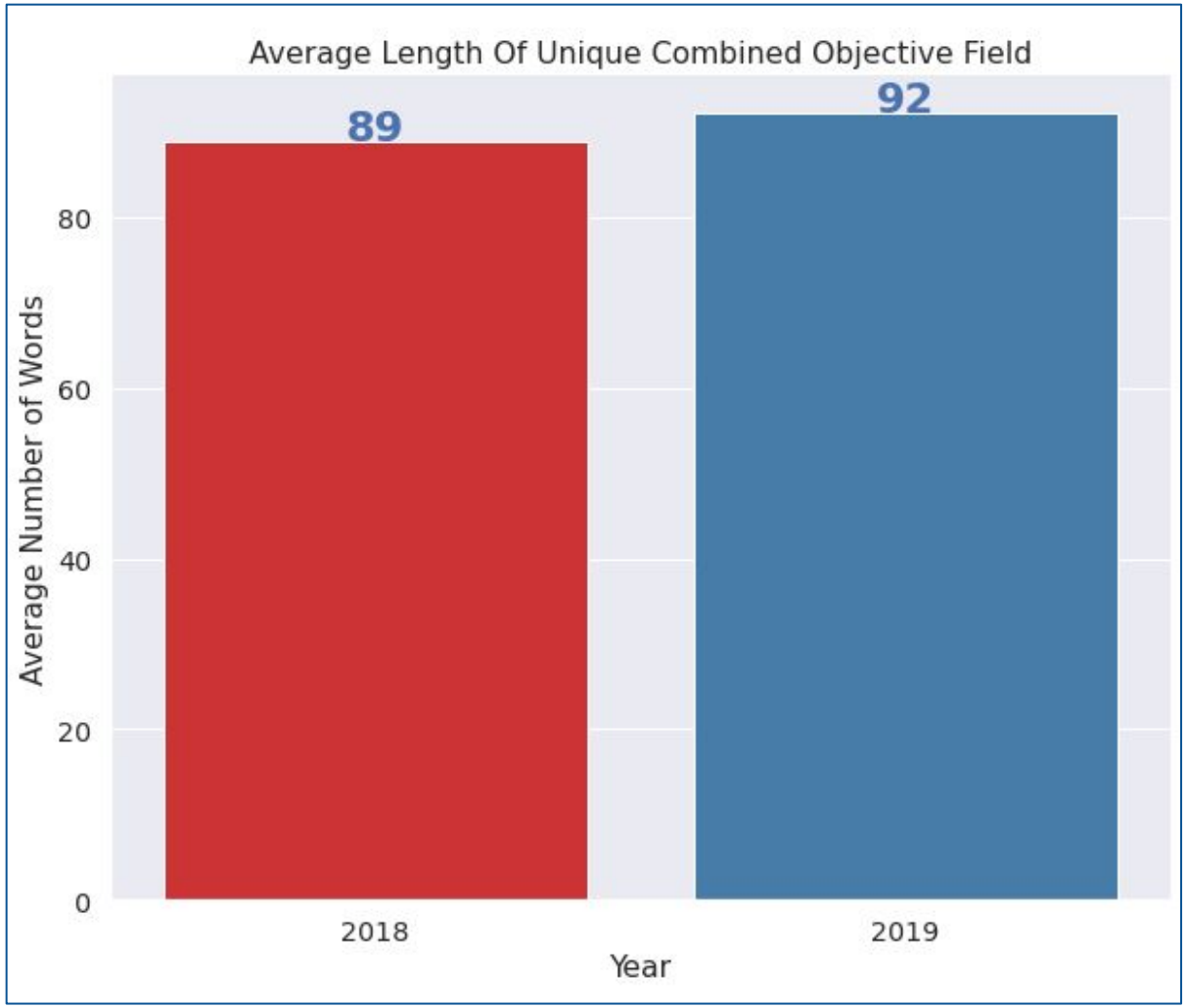
Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns
- Length of objectives varies per column

  Combine into one column
- Data contains too much duplication
- A user has on average 35 objectives!

  Pick the longest unique objective for a user



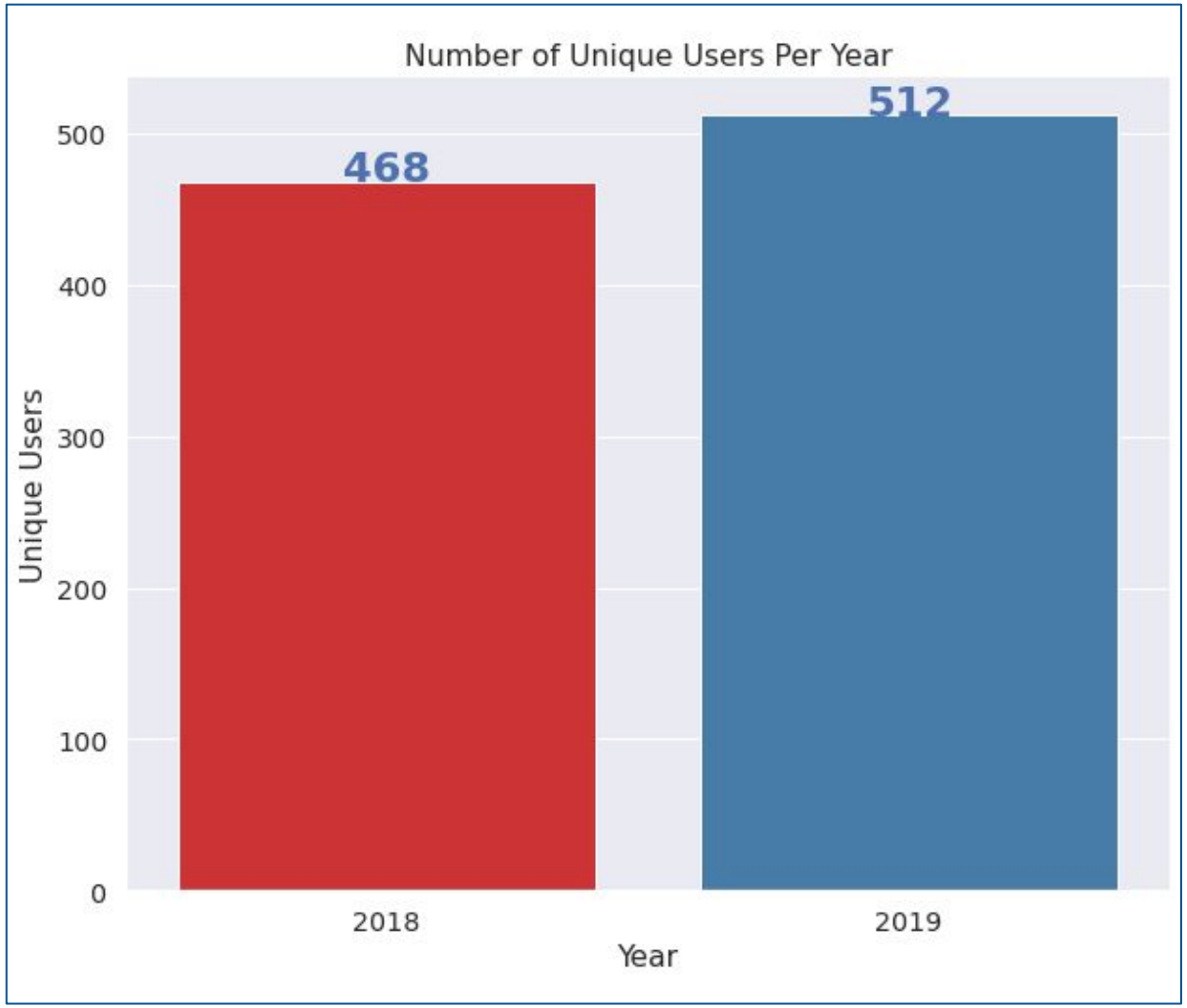
Solution Approach - Data

Processing data from MERCK :

- Data was received as a large csv file with about **30K** records for historical data from 2018 and 2019
- Objectives were spread across multiple columns
- Length of objectives varies per column

  Combine into one column
- Data contains too much duplication
- A user has on average 35 objectives!

  Pick the longest unique objective for a user
- Cleaned dataset contained about **1K** users each with one unique objective of 90 words on average



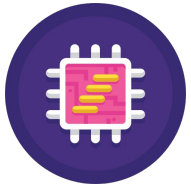
Outline

1) Problem Statement
2) Solution Approach
3) **Deliverables**
4) Evaluation

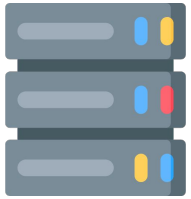
Deliverables

A dockerized web application consisting of:

- Embeddings-Core
  - Contains the logic of embedding generation and data indexing.



- Back-End server
  - API wrapper around the embeddings core.
  - Can be called by the front end.
  - Can be integrated into external systems.



- Front-End client
  - Graphical interface to interact with the application.

Deliverables - Demo

- Login
- User Roles
- Index Data
- Query
- Predictions Explainability

Outline

1) Problem Statement
2) Solution Approach
3) Deliverables
4) **Evaluation**

Outline

1) Problem Statement
2) Solution Approach
3) Deliverables
4) **Evaluation**

- Analytics
- Usability

Evaluation

Two types of evaluations were conducted

- Analytics
  -  Semantic Similarity
  -  Embeddings Generation Time
- Usability
  -  User-centric evaluation of the system conducted as a survey.

Outline

1) Problem Statement
2) Solution Approach
3) Deliverables
4) **Evaluation**

- **Analytics**
- Usability

Evaluation - Analytics

- Semantic Similarity
  - Test the ability of the system to retrieve users with similar objectives.

Evaluation - Analytics

- Semantic Similarity
  - Test the ability of the system to retrieve users with similar objectives.

⚠️ MERCK's data does not have true labels.

Evaluation - Analytics

- Semantic Similarity
  - Test the ability of the system to retrieve users with similar objectives.

⚠ MERCK's data does not have true labels.

✅ Use sentence-pair similarity benchmark: **SICK** dataset.

Evaluation - Analytics

- Semantic Similarity
  - Test the ability of the system to retrieve users with similar objectives.


MERCK's data does not have true labels.


Use sentence-pair similarity benchmark: **SICK** dataset.

- 10K English sentence pairs.
- Annotated Manually with a relatedness score from 1 to 5.
-

ID	Sentence 1	Sentence 2	True Relatedness
752	The presentation is being watched by a classroom of students	The presentation is being attended by a classroom of students	4.8

Evaluation - Analytics

- Semantic Similarity
  - Test the ability of the system to retrieve users with similar objectives.

⚠ MERCK's data does not have true labels.

✅ Use sentence-pair similarity benchmark: **SICK** dataset.

- 10K English sentence pairs.
- Annotated Manually with a relatedness score from 1 to 5.
-

ID	Sentence 1	Sentence 2	True Relatedness
752	The presentation is being watched by a classroom of students	The presentation is being attended by a classroom of students	4.8

ID	Embedding of Sentence 1	Embedding of Sentence 2	Predicted Similarity	True Relatedness
752	[1,2,3,4,5]	[1,2,5,3,5]	4	4.8

Evaluation - Analytics

- Semantic Similarity
  - Test the ability of the system to retrieve users with similar objectives.

  ⚠ MERCK's data does not have true labels.

  ✅ Use sentence-pair similarity benchmark: **SICK** dataset.

- Embedding Generation Time

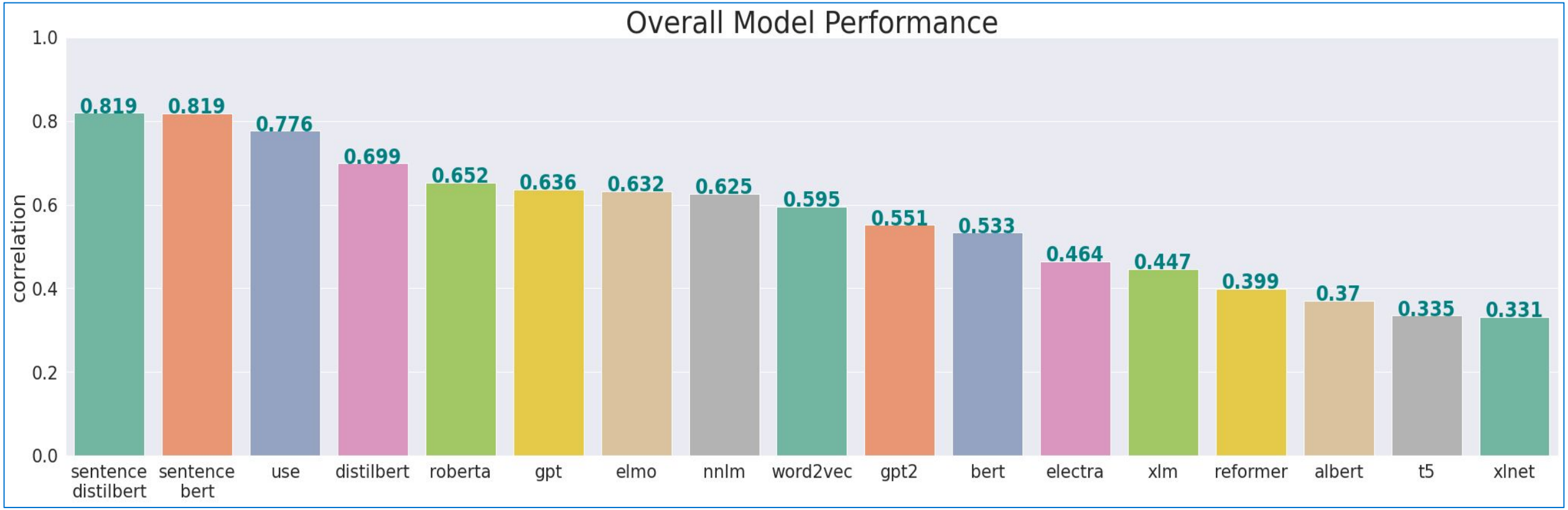
  ✅ Measure running time on benchmark datasets.

Evaluation - Analytics

SICK is a benchmark dataset for evaluating **semantic similarity** of sentences

Embeddings were generated from SICK sentences.

Similarity was estimated between embeddings and measured against true benchmark data

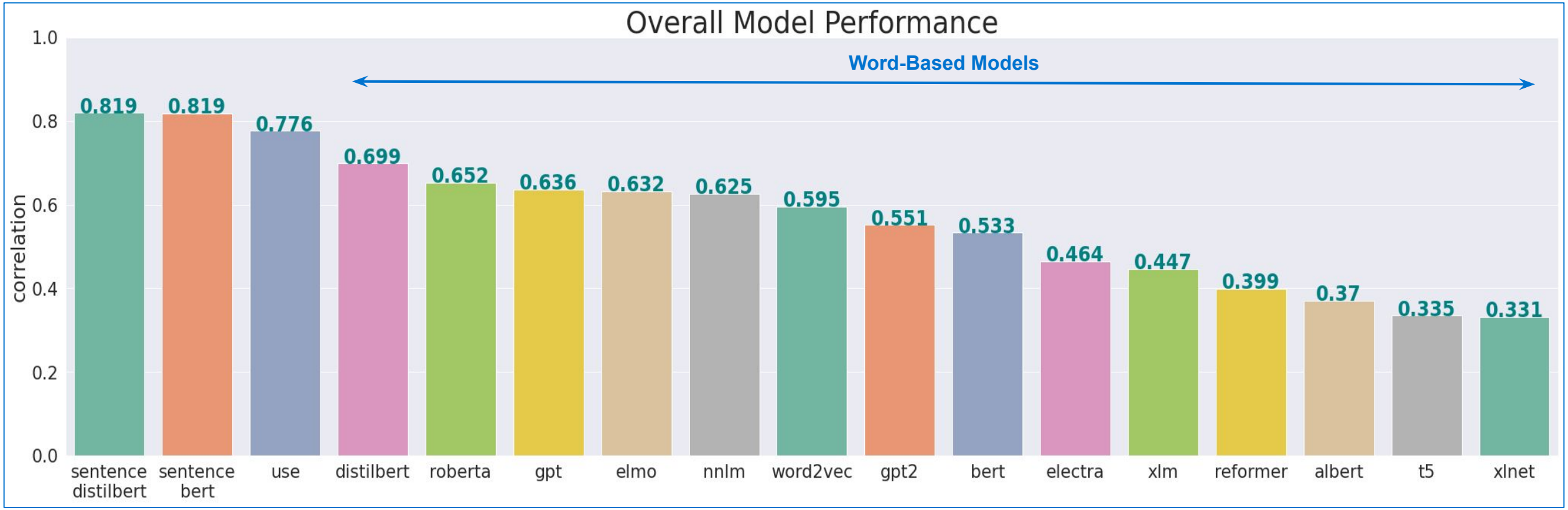


Evaluation - Analytics

SICK is a benchmark dataset for evaluating **semantic similarity** of sentences

Embeddings were generated from SICK sentences.

Similarity was estimated between embeddings and measured against true benchmark data

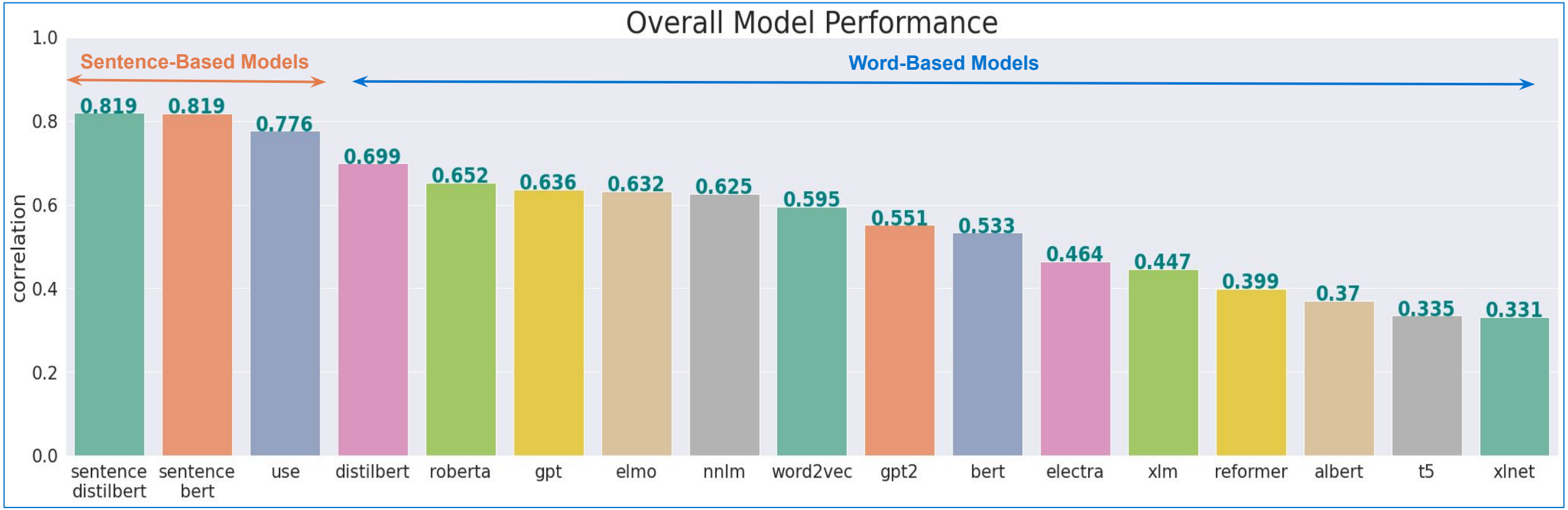


Evaluation - Analytics

SICK is a benchmark dataset for evaluating **semantic similarity** of sentences

Embeddings were generated from SICK sentences.

Similarity was estimated between embeddings and measured against true benchmark data

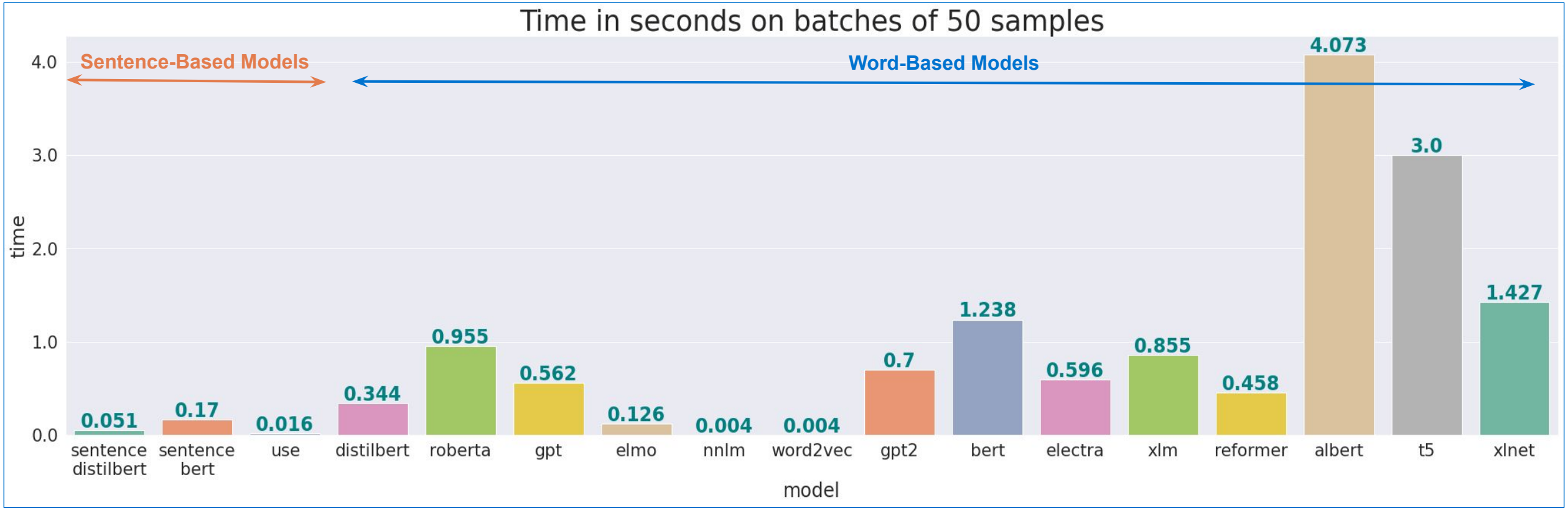


Evaluation - Analytics

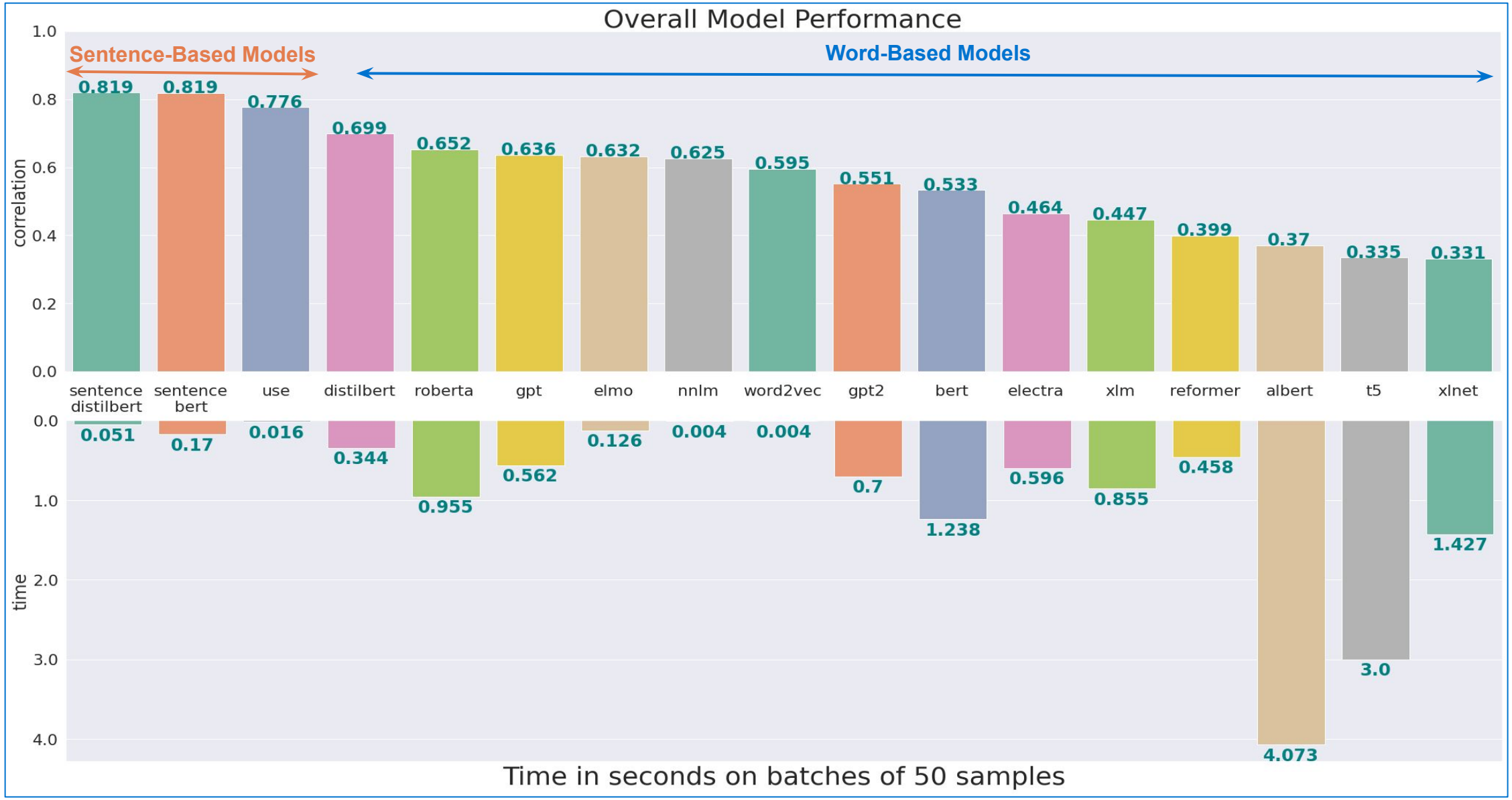
Embedding generation is done in batches of **50 sentence.**

The time taken by each model to generate embeddings was recorded.

The graph below shows the average time in seconds to generate embeddings for a single batch.



Evaluation - Analytics



Outline

1) Problem Statement
2) Solution Approach
3) Deliverables
4) **Evaluation**

- Analytics
- **Usability**

Evaluation - Usability

Users from MERCK were asked to use the system and then fill out a survey.

They were asked to express their agreement/disagreement on a 5 point scale.

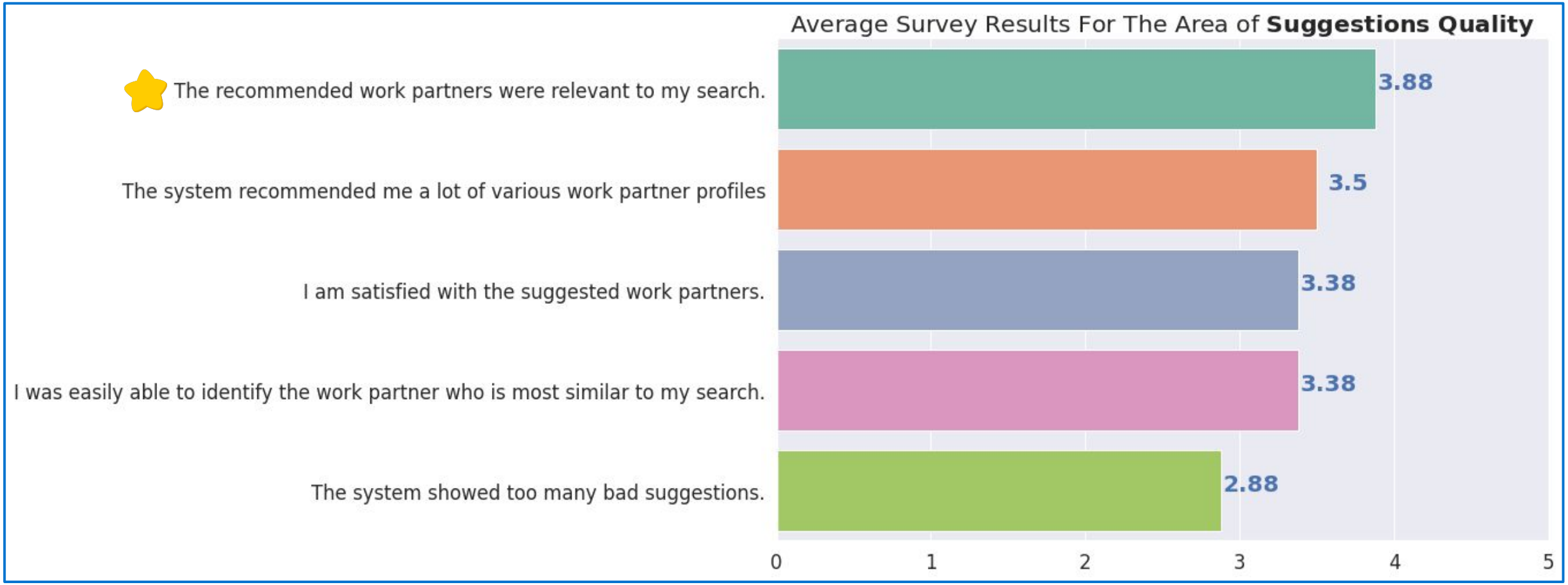
The survey is composed of 20 statements evaluating these areas:

- ⚠️ Quality of system suggestions.
- ⚠️ Effectiveness the system in the process of matching work partners.
- ⚠️ Effort to use the system.
- ⚠️ Layout and explainability.

Eight people in total participated in the survey.

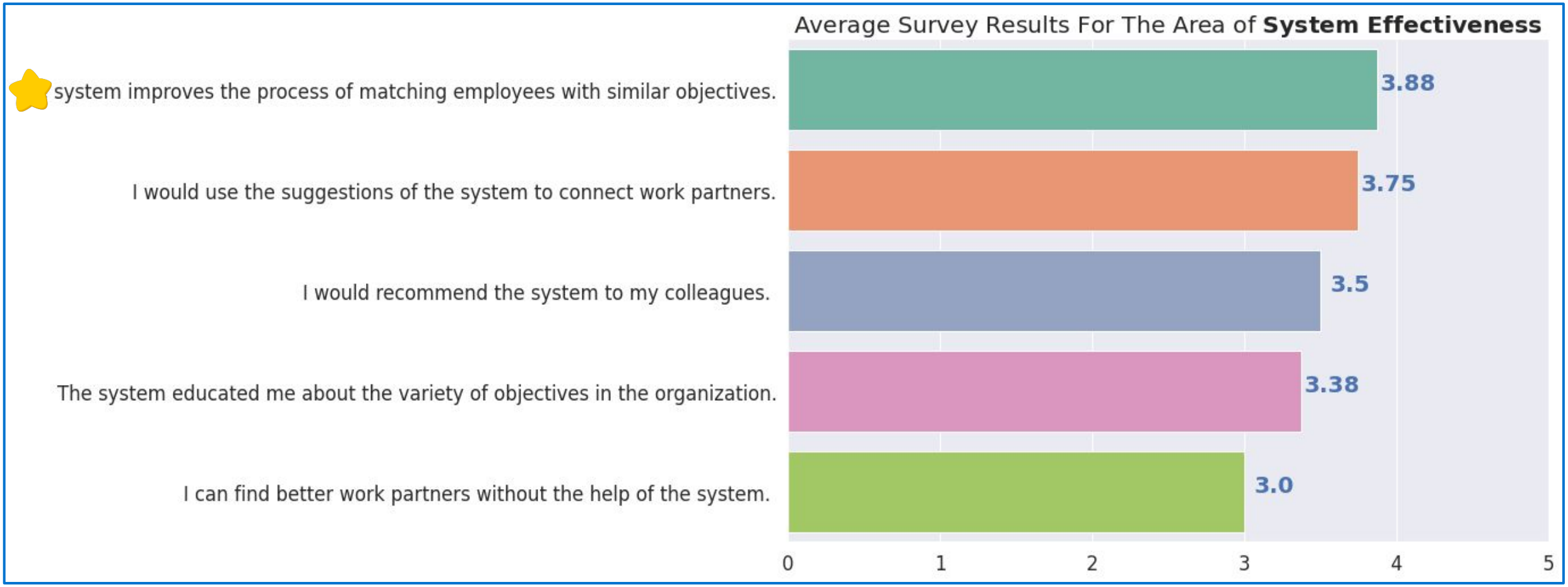
Evaluation - Usability

(1) Quality of system suggestions



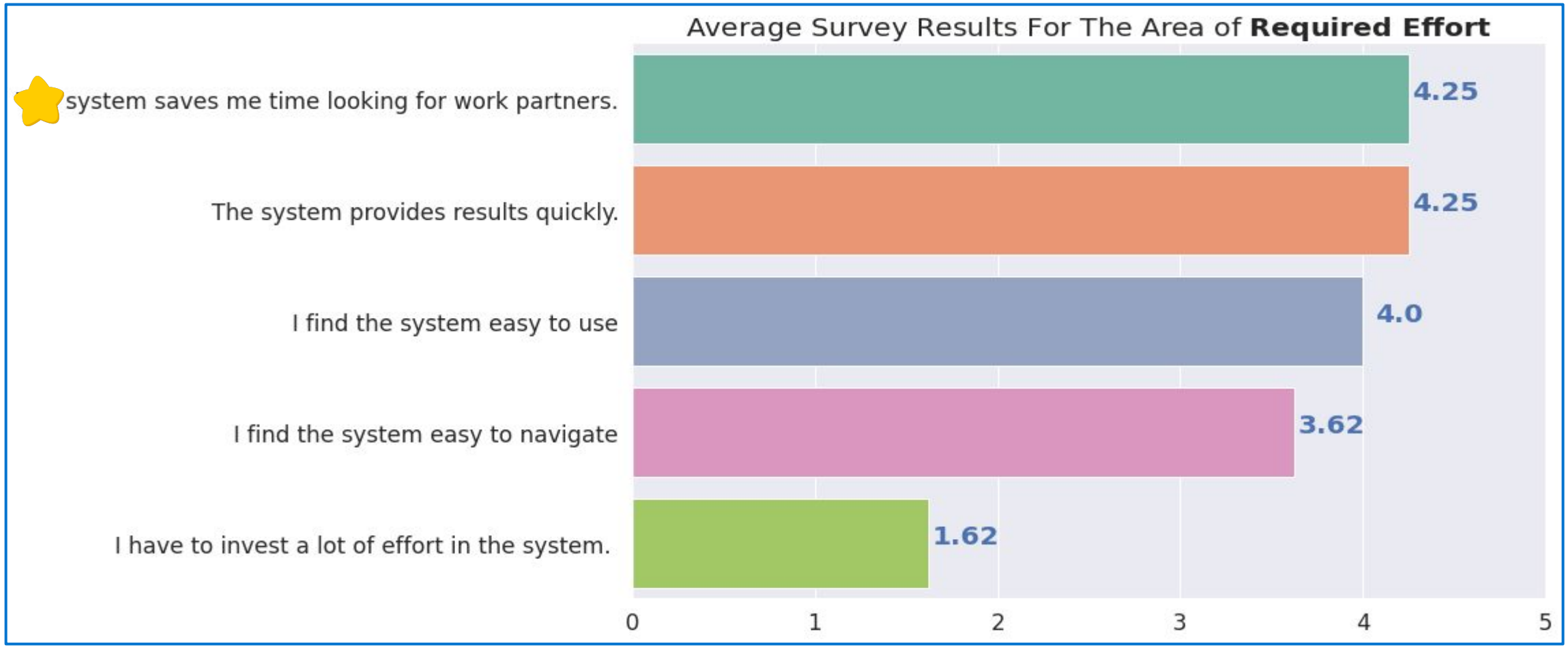
Evaluation - Usability

(2) Effectiveness of the system in the process of matching work partners



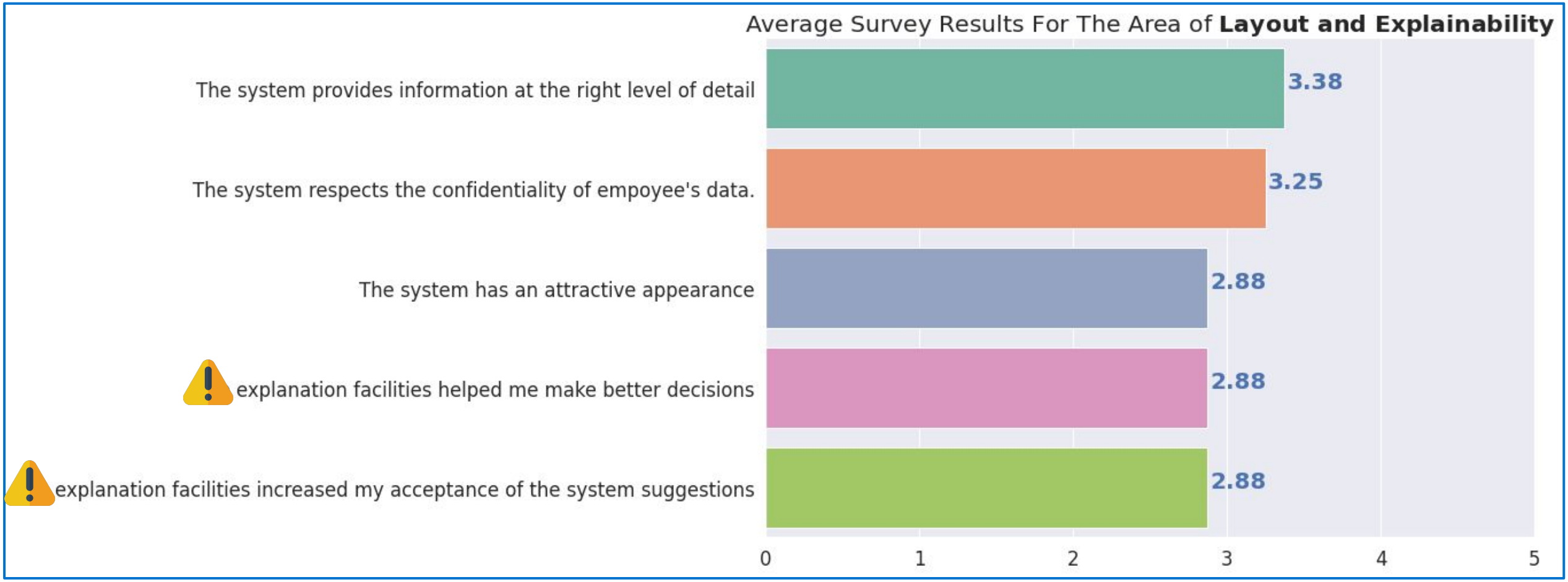
Evaluation - Usability

(3) Effort to use the system



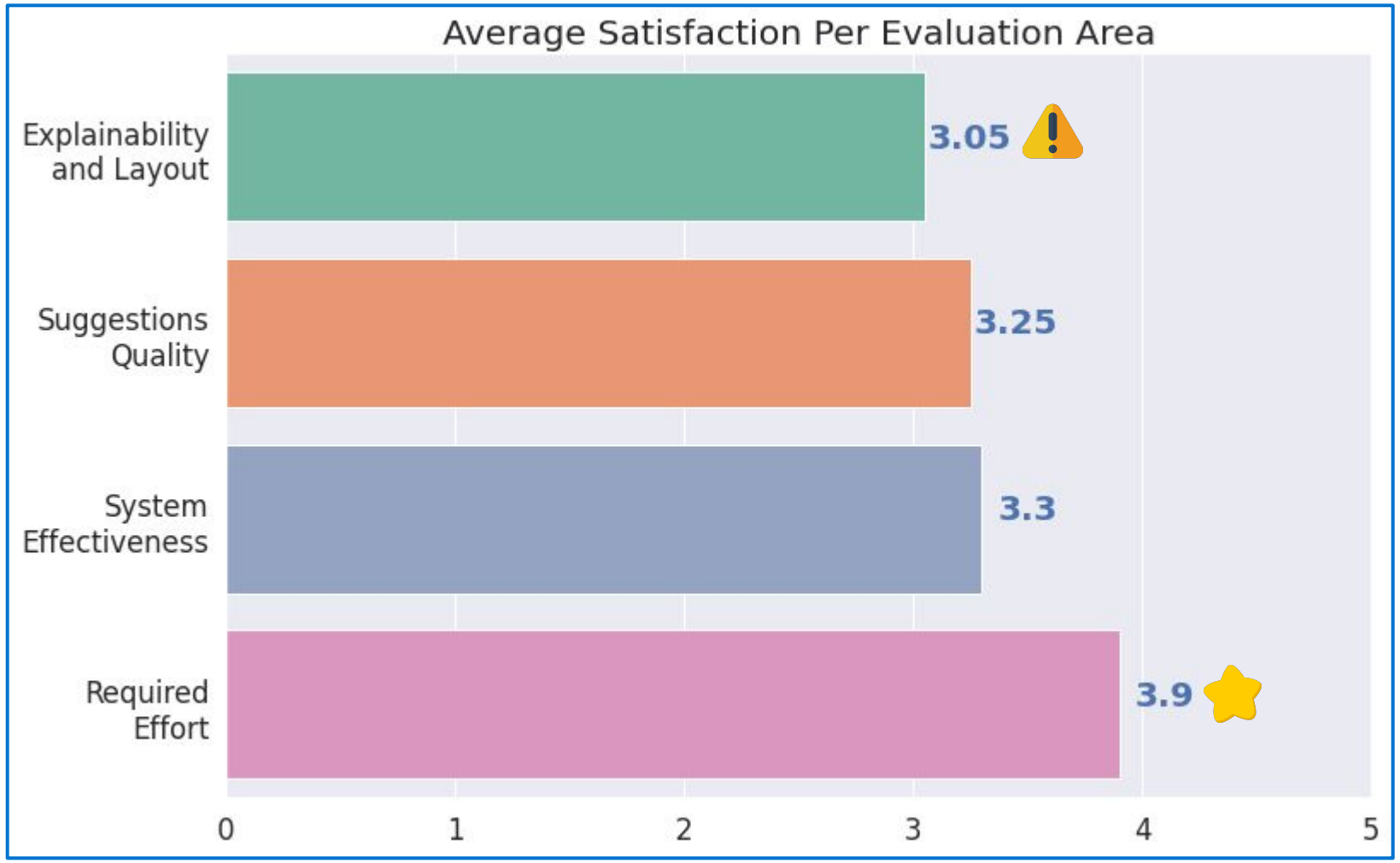
Evaluation - Usability

(4) Layout and explainability



Evaluation - Usability

Average of all areas



Feedback:

- ⚠️ Color coding could be improved.
- ➕ Rank/Filter results by keywords

Questions

Thank you!

Mohab Ghanem

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 7132
Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de

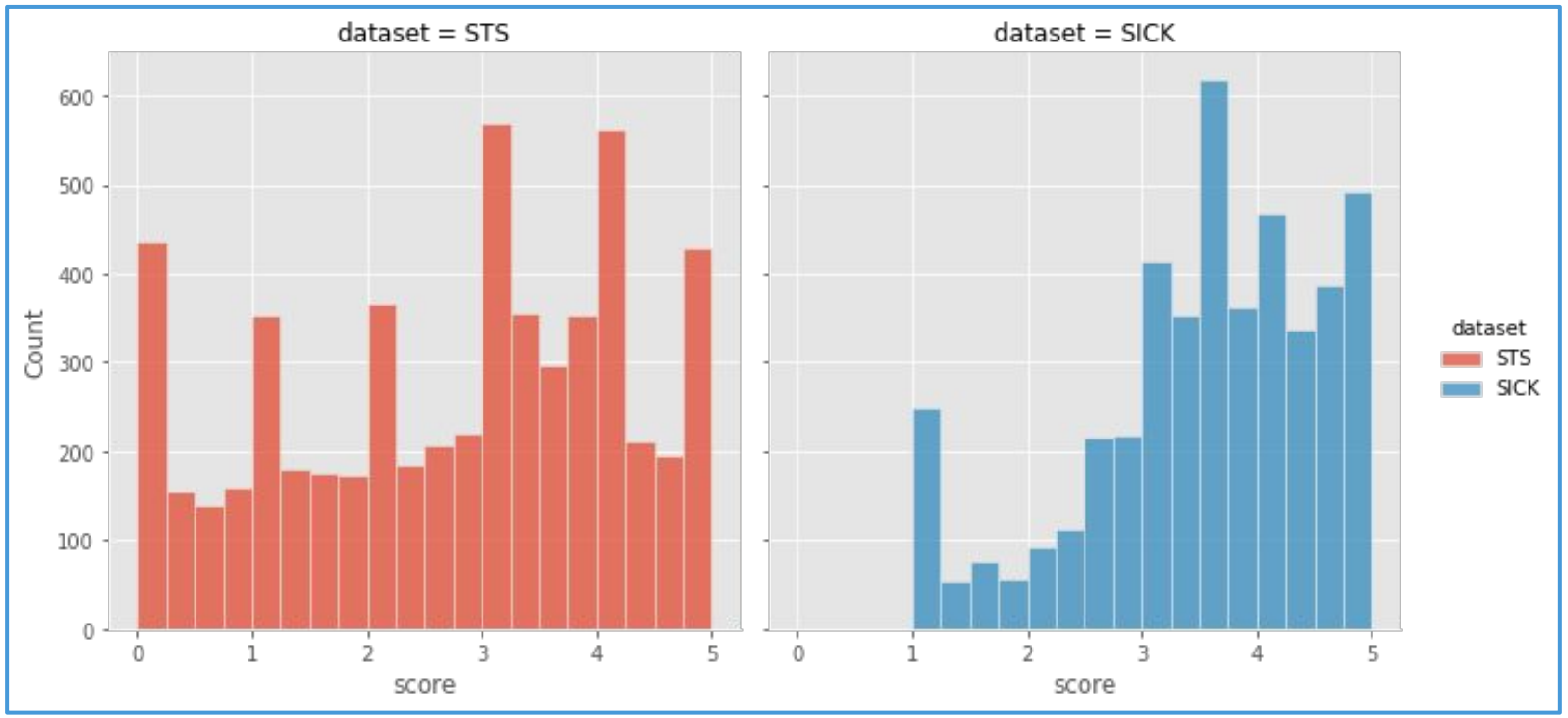
Appendix

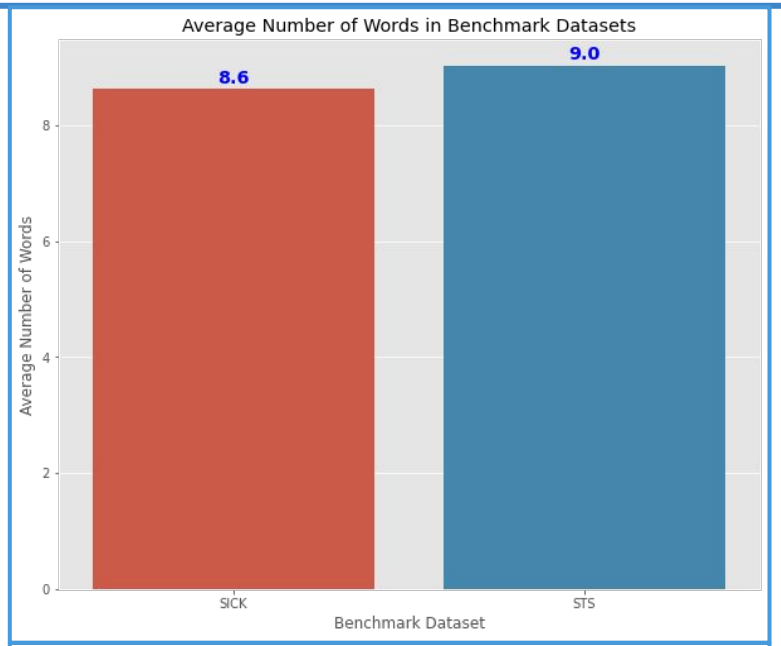
Appendix - Distance Metrics and Models

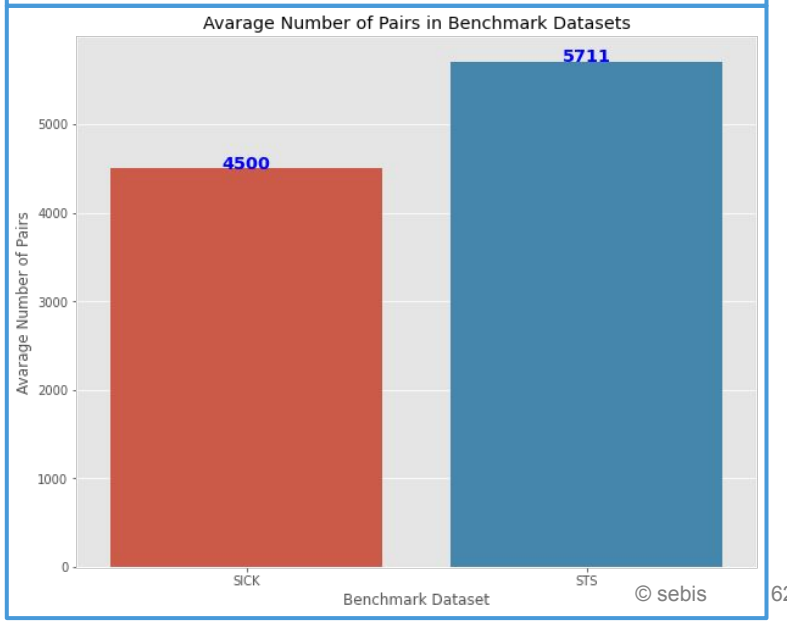
1. Cosine
2. Minkowski
3. Correlation
4. Cityblock
5. Square Euclidean
6. Euclidean
7. Chebyshev
8. Canberra
9. Braycurtis

1. USE
2. Word2Vec
3. NNLM
4. Bert
5. Albert
6. XLM
7. XLNet
8. GPT
9. GPT2
10. Reformer
11. DistillBert
12. Electra
13. T5
14. Sentence Bert
15. Sentence DistilBert

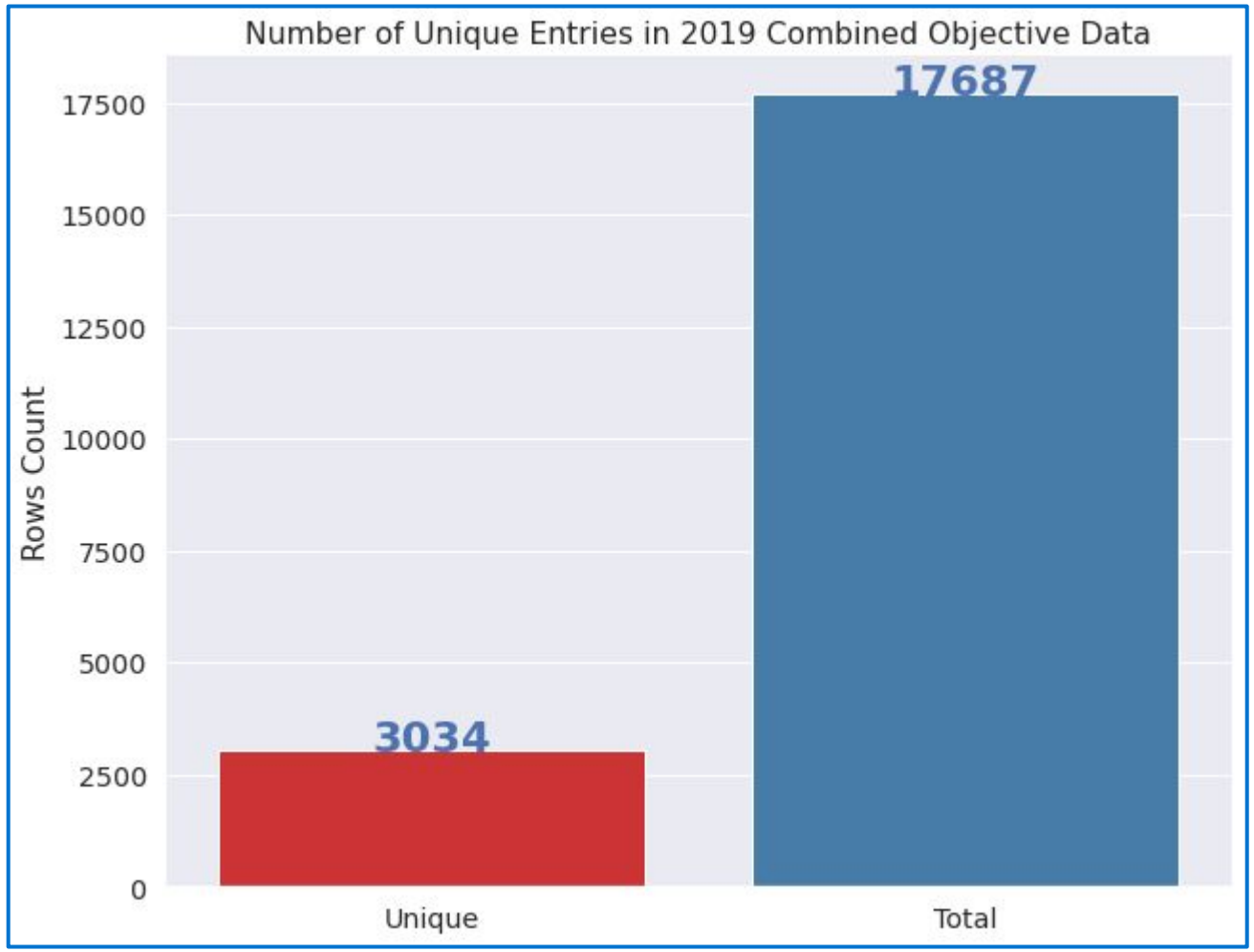
Appendix - Benchmark Datasets

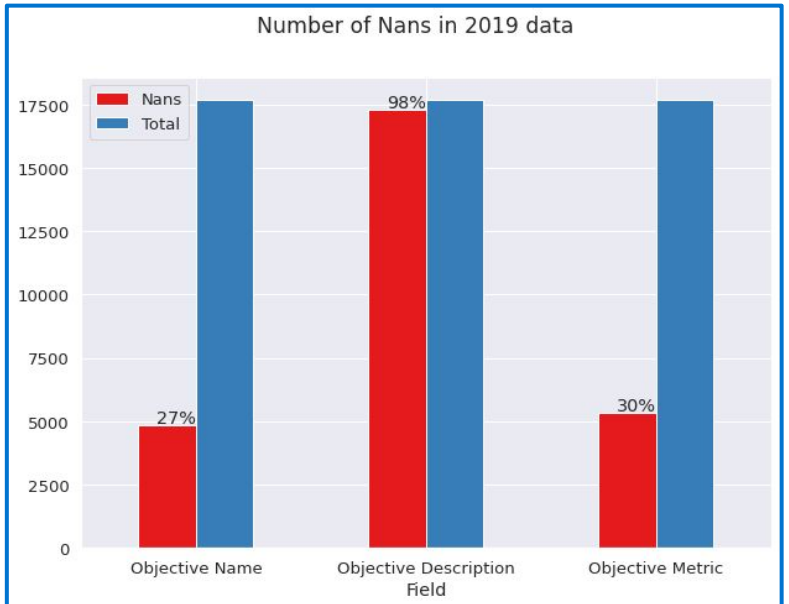


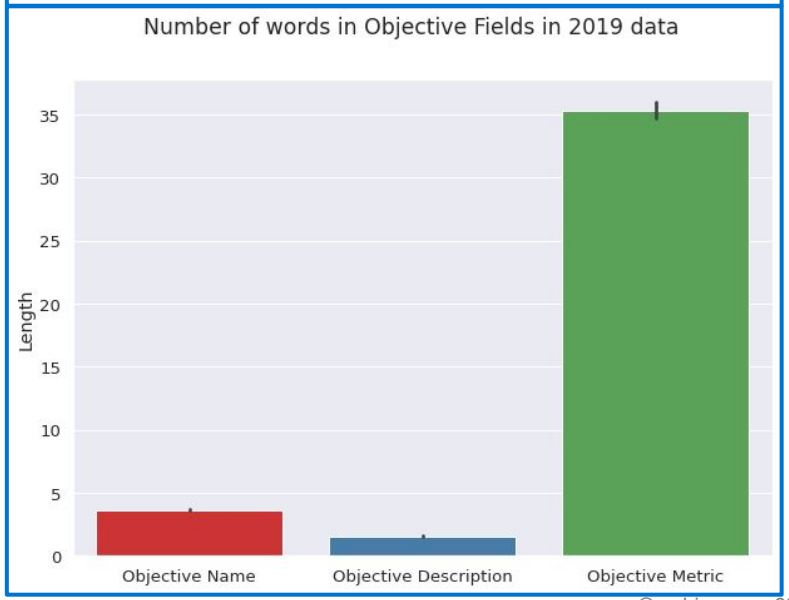




Appendix - 2019 Data



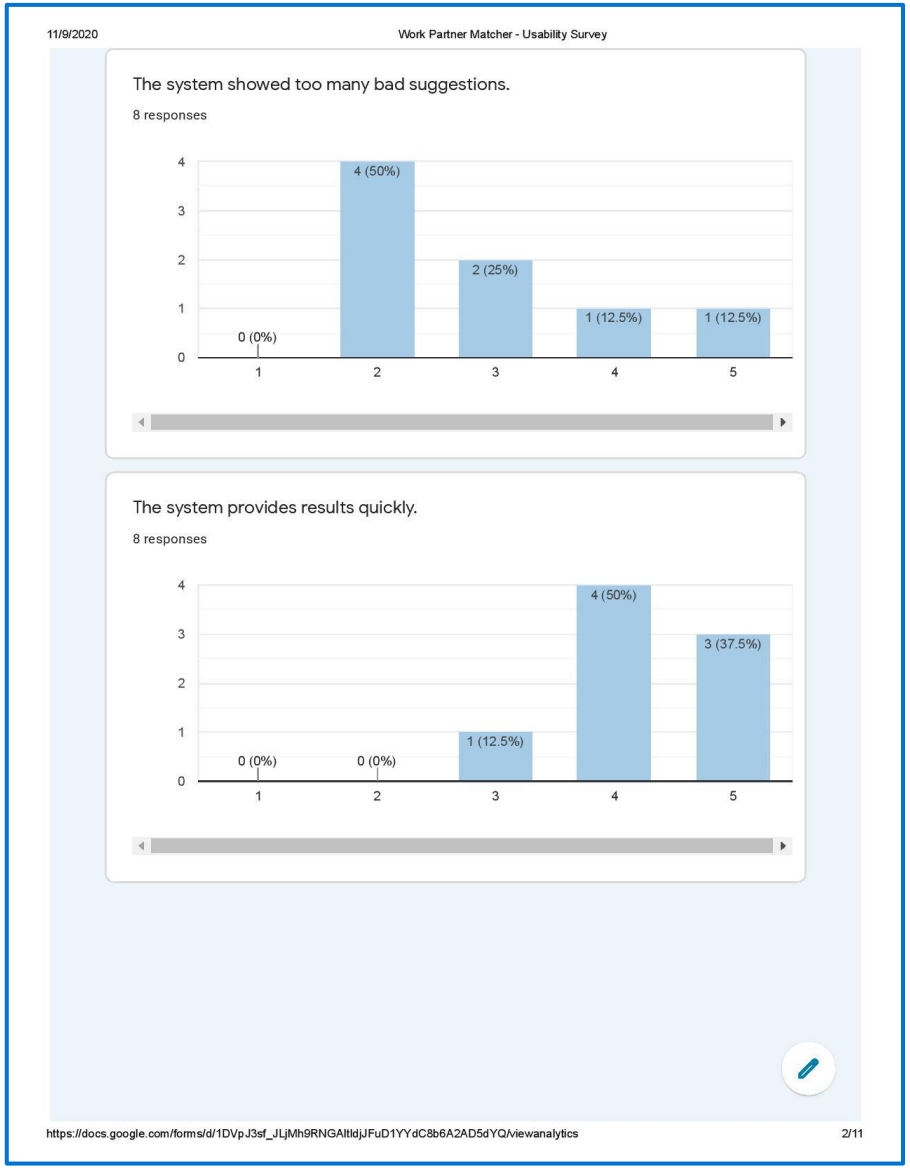




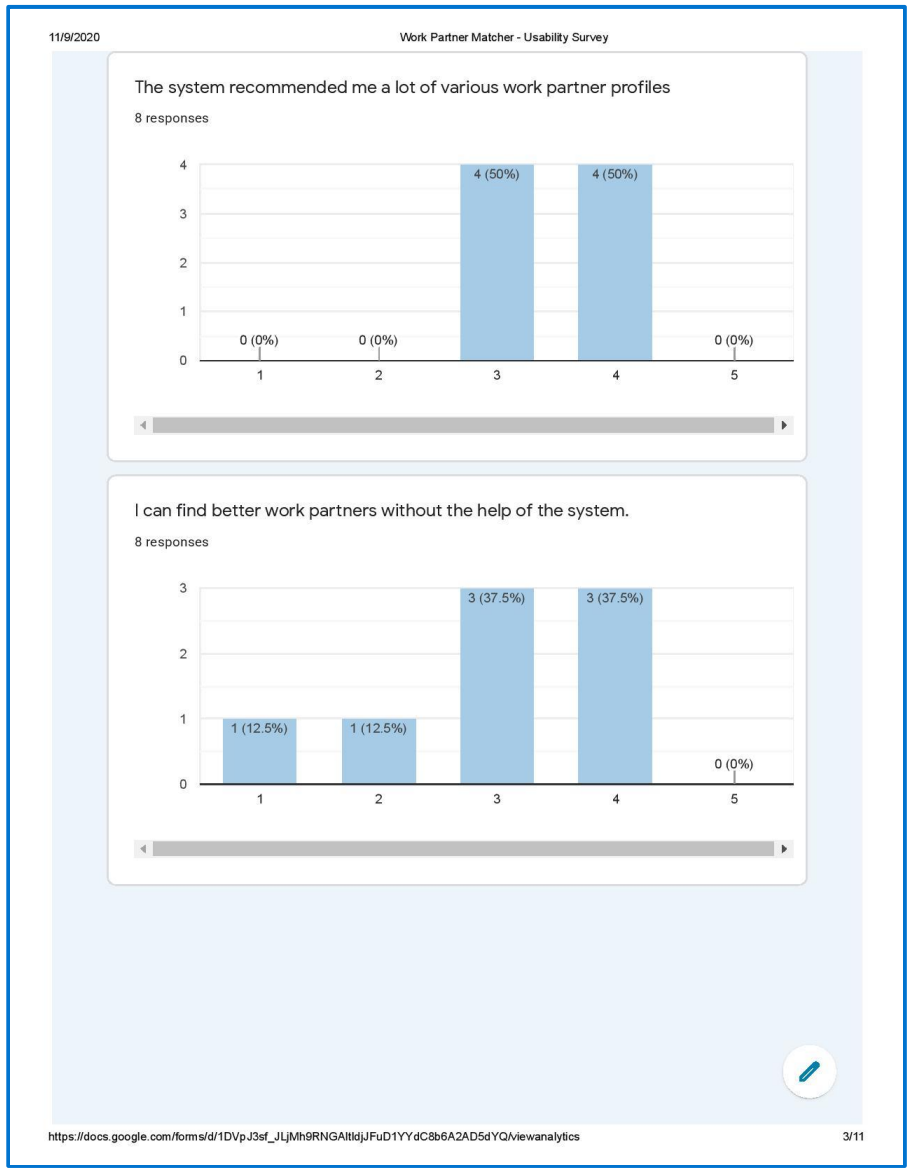
Appendix - Job Description Dataset

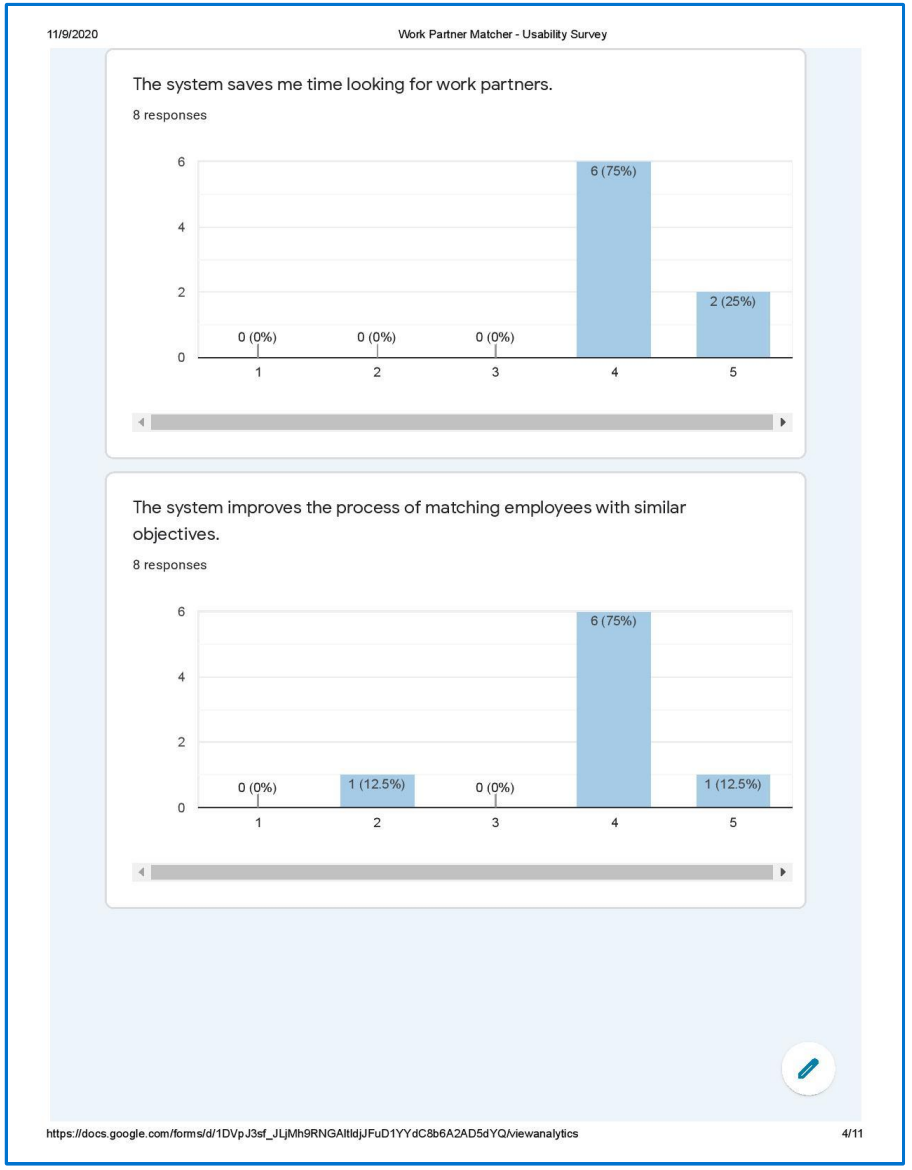
Administrative Assistant : Administrative Assistant - In Bus 26yrs Fashion Handbag wholesaler need assistant for Sales and Inventory Depmts., barnard@americanwest.cc SF52144 As leading supplier of handbags and accessories to the Western Industry, American West is looking for a detail orientated, focused self starter that can multi task in both a sales and operations environment. The Administrative assistant to sales and operations will be responsible for assistance in data entry and customer service for the Sales Department and administrative assistance to the New Product Development and Inventory manager; reporting to the EVP. POSITION SUMMARY: The Sales and Operations Administrative Assistant is responsible for assisting the Sales Department, New Product Development and Inventory Manager with general administration duties, data entry and assisting customers with great customer service.
Sales Representative : Are you ready for something new? Are you tired of your current job and not making any real money? Complete Home Experts is looking for sales people who are tired of making average or below average money. If you want to make the great money you deserve, and change your lifestyle into what you want, then call us. We are filling 10 positions for this quarter. Excellent Customer Service is a must. Attention to detail is a must. Our craftsmanship makes referrals easy!
Customer Service Representative : Superior Staff Resources is currently seeking a Customer Service Representative/Cashier for our client in Albany, NY. . The successful candidate will be able to successfully perform the following duties: • The CSR processes cash, check and credit card payments in a walk-in environment in accordance with all documented guidelines. • The CSR is responsible for tag issuance and distribution at the window in accordance with all documented guidelines. • The CSR processes new account applications in accordance with all documented guidelines. • The CSR is responsible for an end-of-day reconciliation of their daily work and submitting all relevant settlement documentation. •

Appendix - Questionnaire

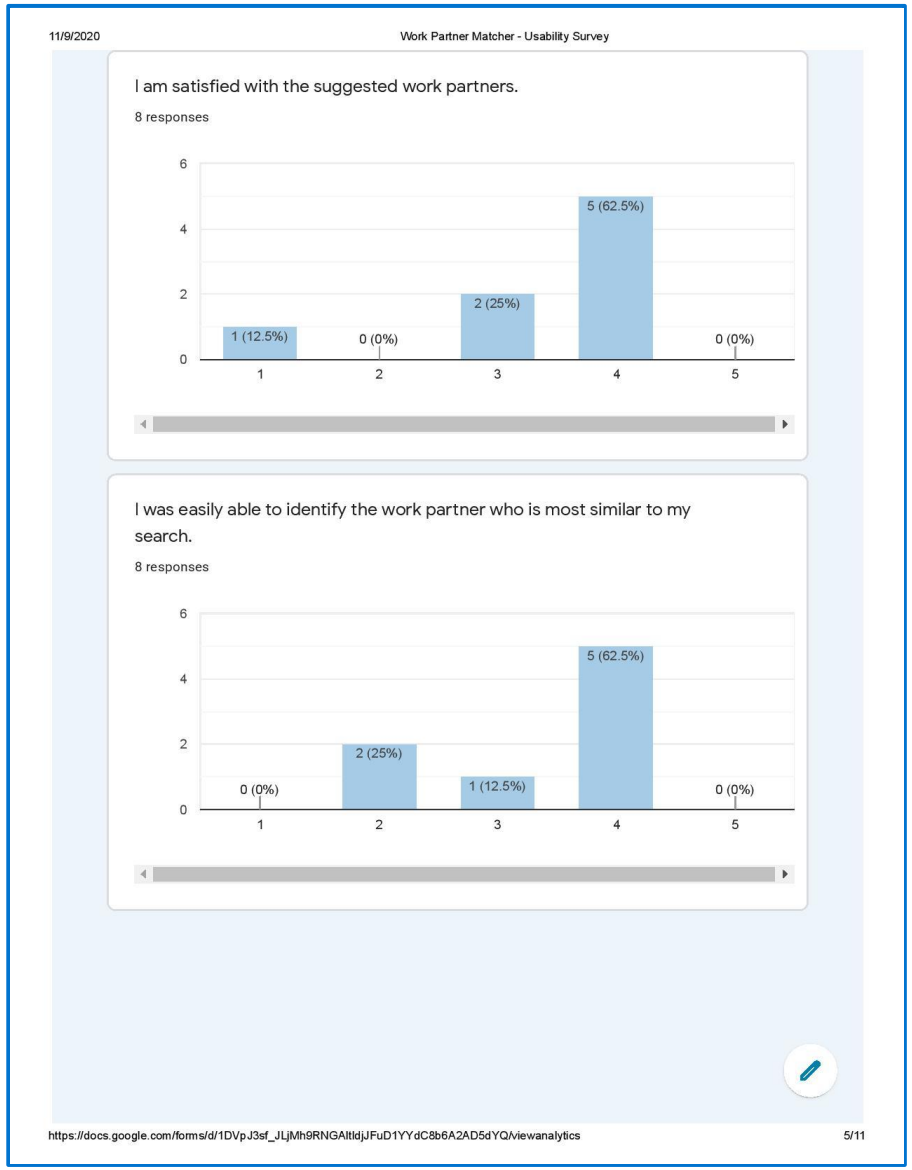


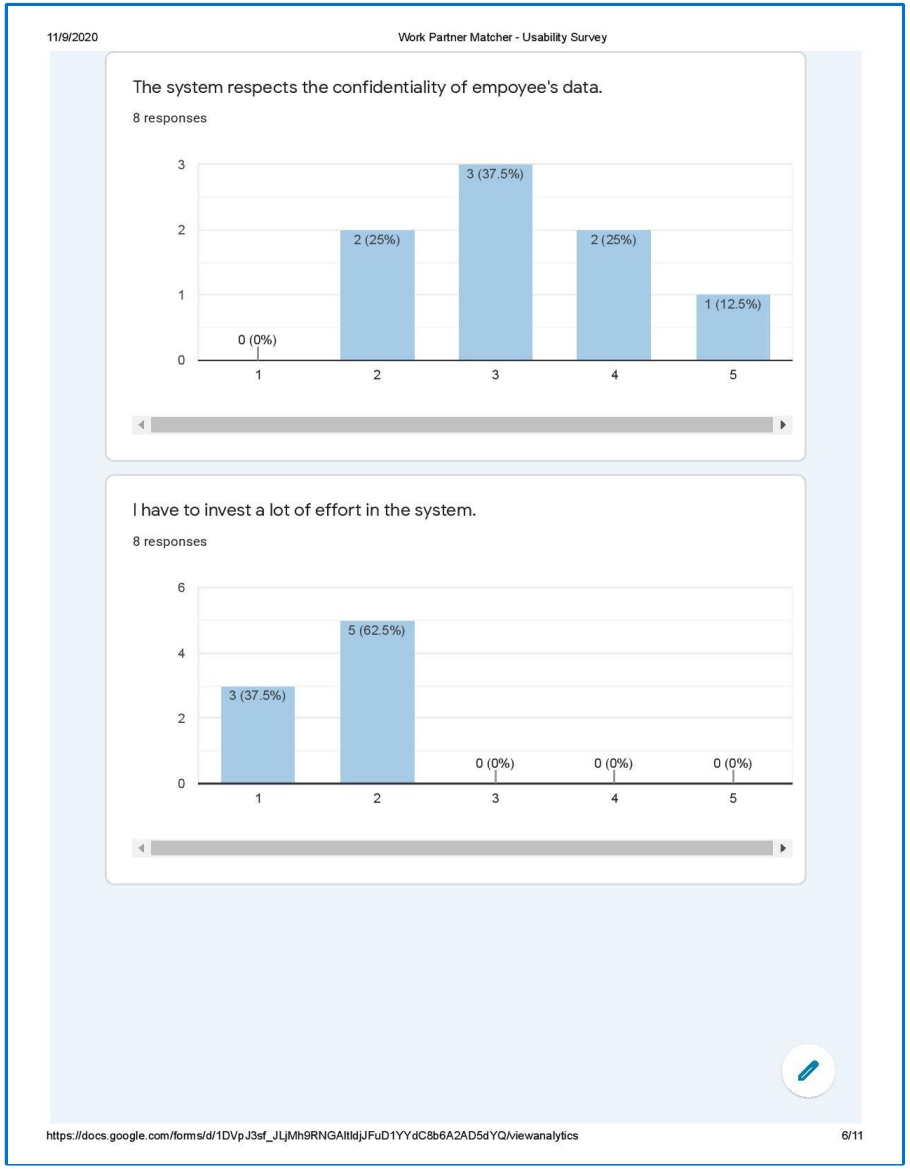
Appendix - Questionnaire



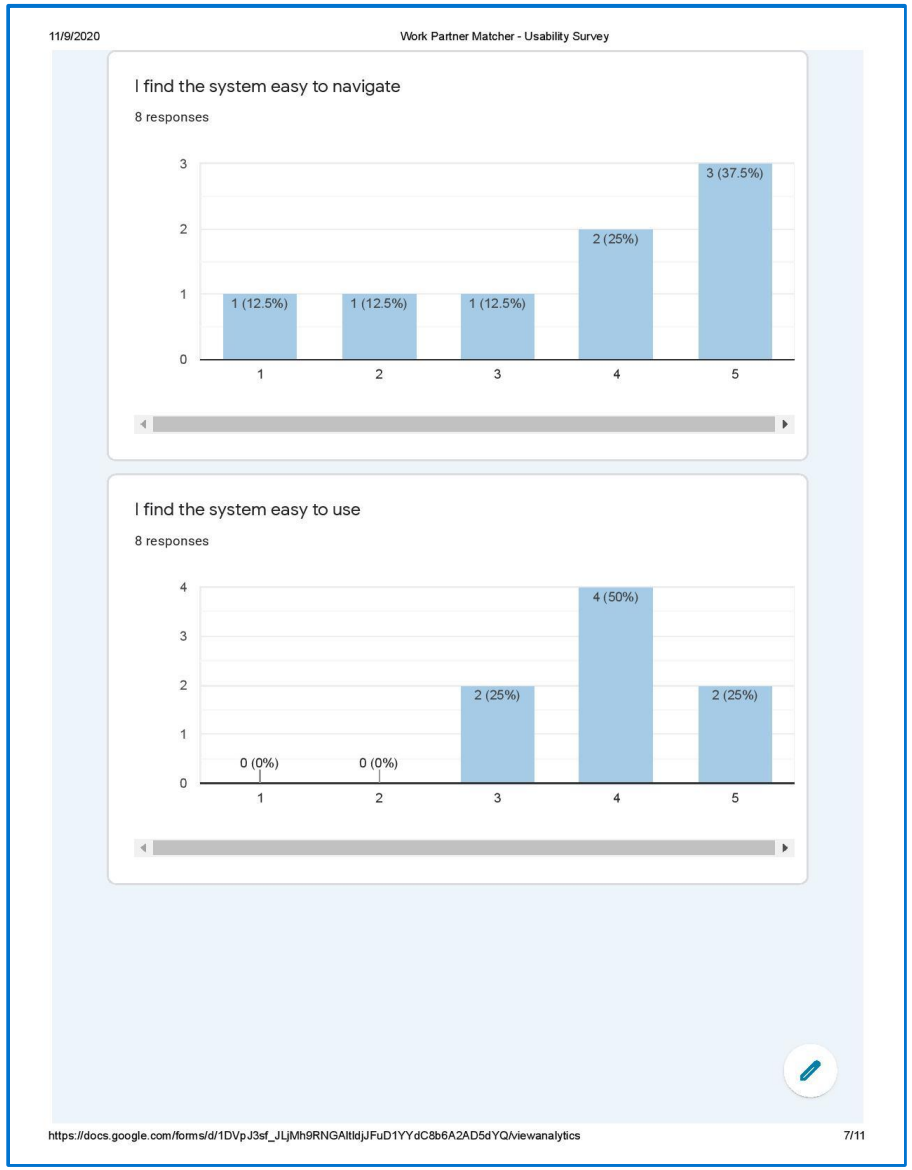


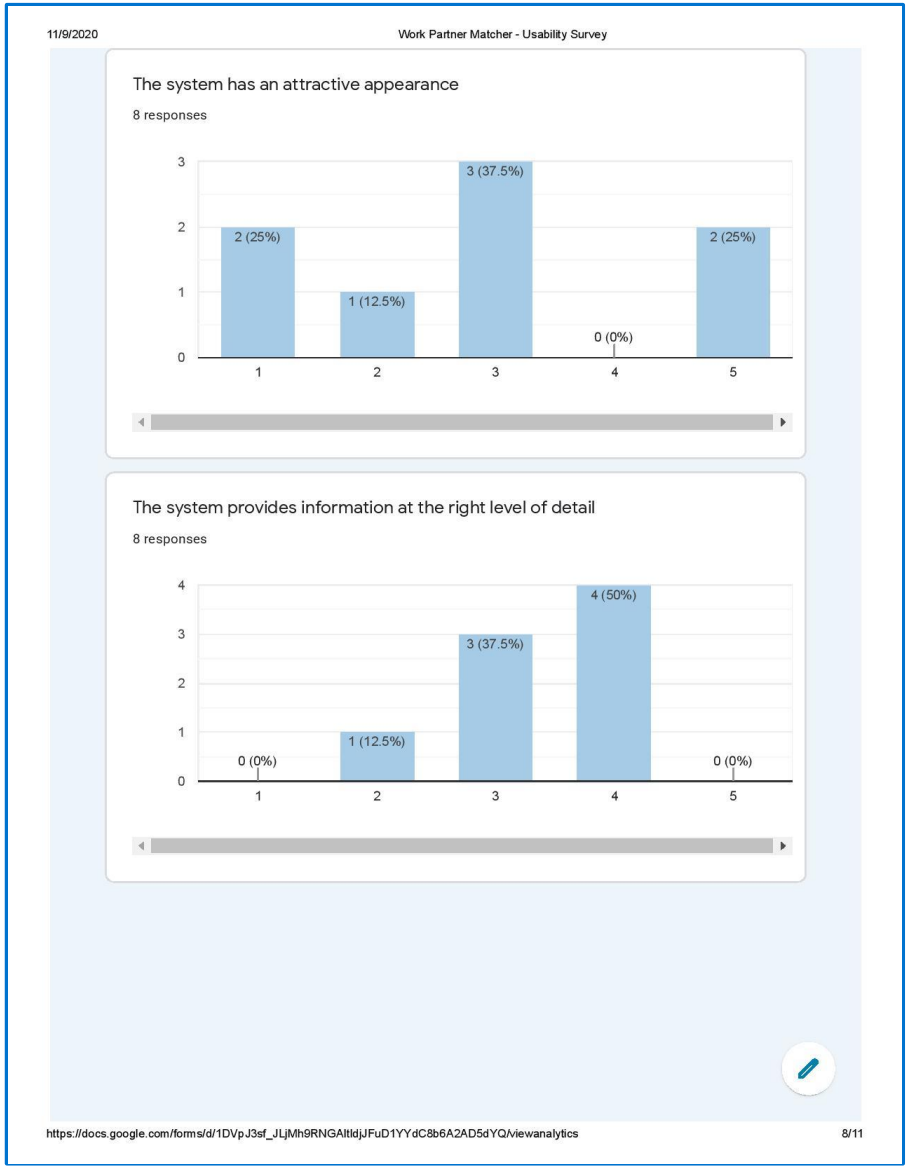
Appendix - Questionnaire



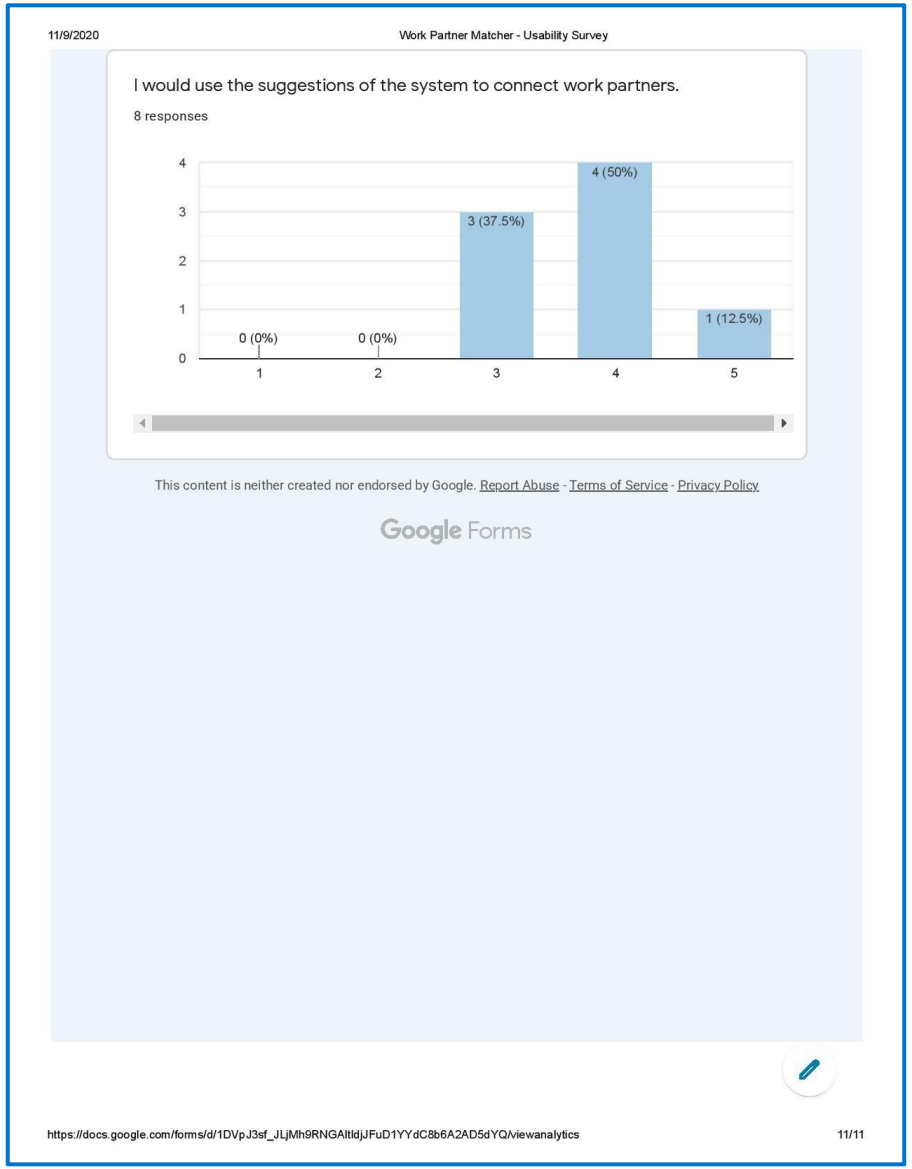


Appendix - Questionnaire





Appendix - Questionnaire



Appendix - Questionnaire