Topic Modeling for Employee Objectives using Word Embeddings

Anum Afzal 01/02/2021

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de

Agenda

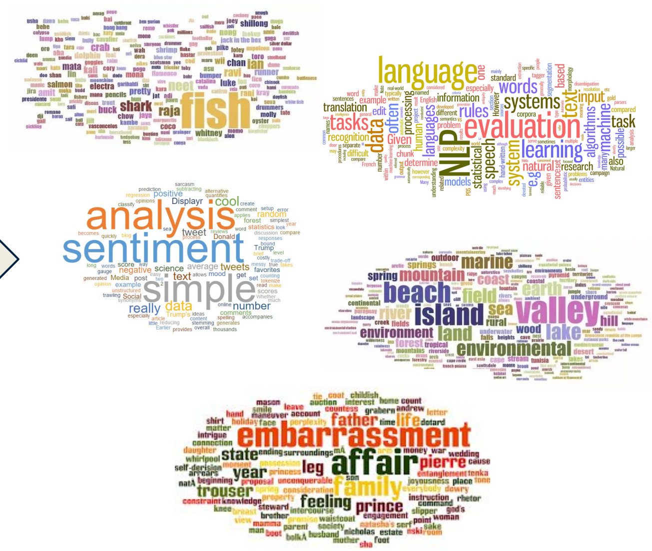
- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Use case

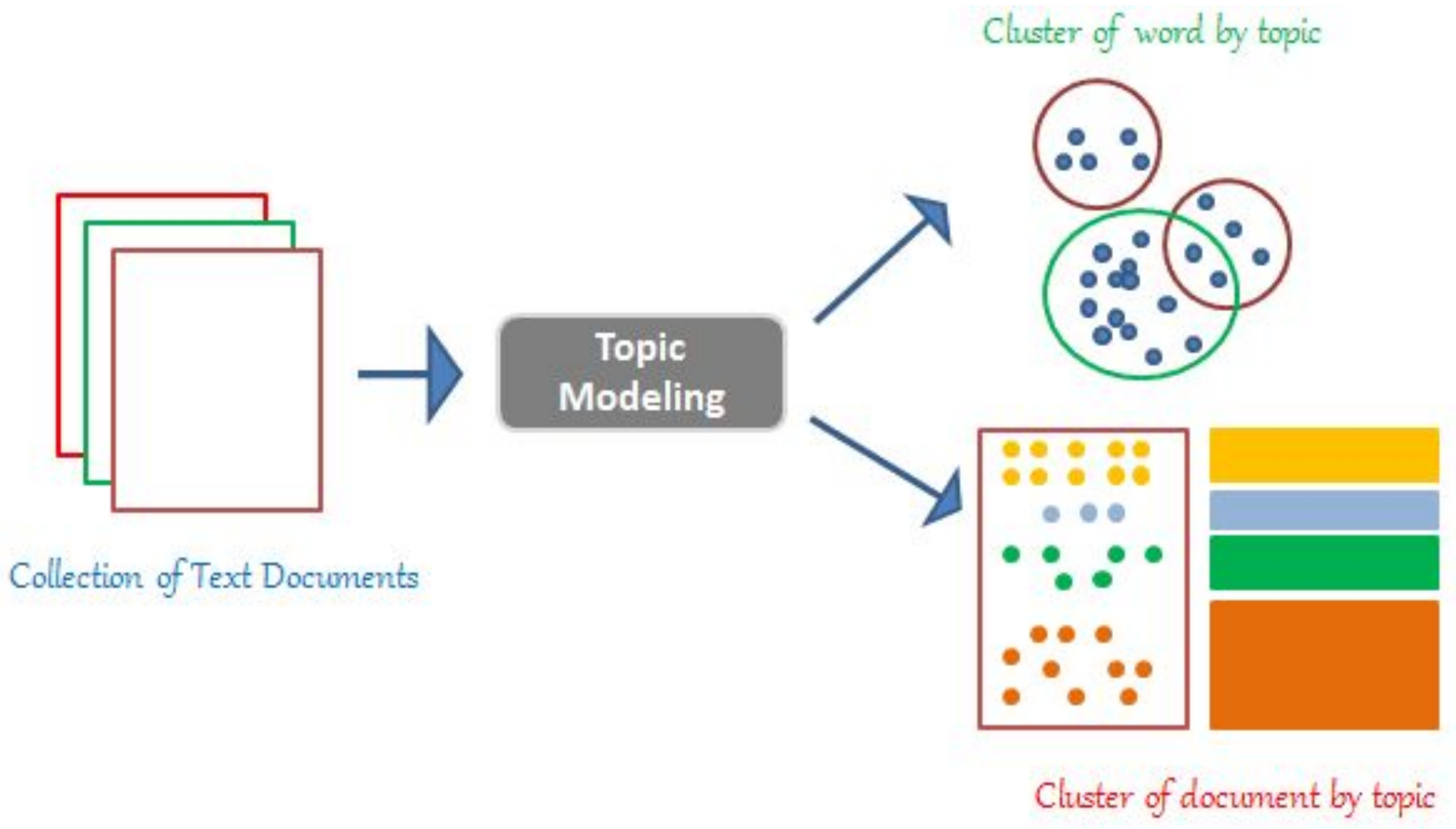
CEO or HOD

Employee Objectives

Magic Box



Topic Modeling for Employee Objective



Some common Topic Modeling approaches include LDA, LSA, PLSA

Agenda

sebis

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Research Questions

sebis

RQ1: Could using embedding vectors lead to better results than Latent Dirichlet Allocation model?

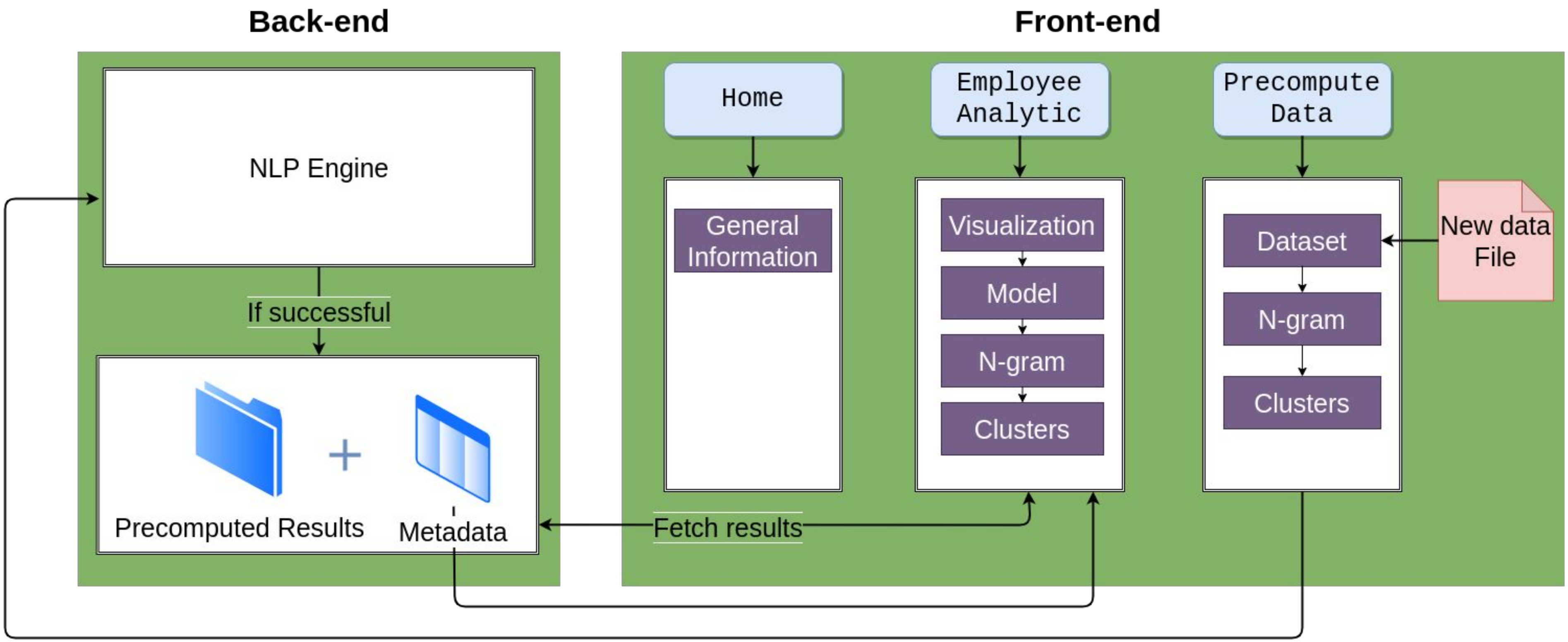
RQ2: If the word embedding models are able to provide better results, then which type of embedding model is better suited?

RQ3: Could using a traditional algorithm such as LDA in tandem with the Embedding models provide better results?

Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

System Architecture - Block Diagram



System Architecture of the Topic Modeling Framework that shows the interaction between the front-end and the back-end. 'Employee Analytic' tab reads meta-data from back-end before displaying options. It also fetches the requested results from back-end. 'Precompute Data' tab reads a new data file and generates results for all selected combination using the NLP engine and stores the result in back-end.

Agenda

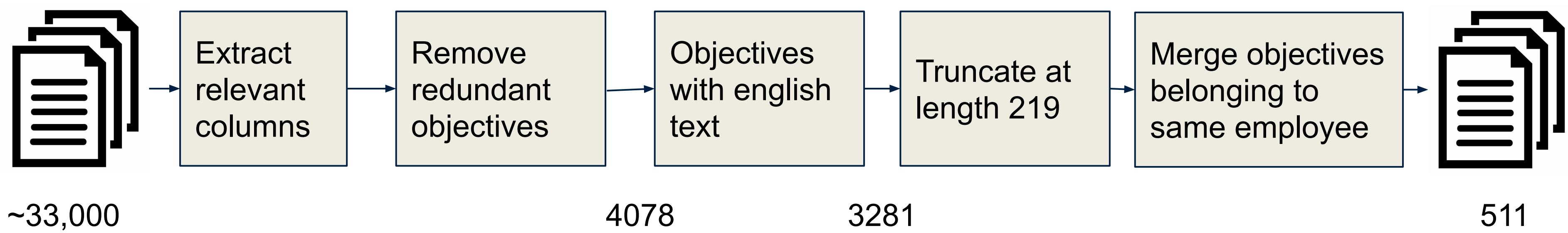
- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Employee Objective Dataset

Column Name	**Column Value**
User ID	*******0000**
Global Key	HR
Functional Area	Human Resources
Location	Temecula
Objective Name	**Operational Excellence**
Objective Description	None
Objective Comment	I appreciate that Kathy has been continuing to work * * * * * * * is a vital part of our support to our * supports * information which is very much appreciated. I also * * * * * * * * * * , TOA to Sick Time and recently * * * * project among *. Thank You, Kathy! * * * * * November 2019.
Objective Metric	**Support * * * * * and other activities focused around * and * for *. Support the * * *, includes * *, * * and other * along the way. Also expected is a regular * * * * * accurate * is recorded in * * * * * * * and is reflected in * * or selected tool. * * * * * * initiatives * * * *. * and * * * * the * of key insights for a further deep dive utilizing *. Continue to * * knowledge of * * * * * * cluster. * * * on a regular basis. Support efforts * from the * * * * * * management and the * for * * Acknowledgement of same.**
Form Template Name	2019 Performance Management

some information is redacted to anonymize the data for privacy reasons as this is a private dataset.

Employee Objective Dataset Analysis Summary

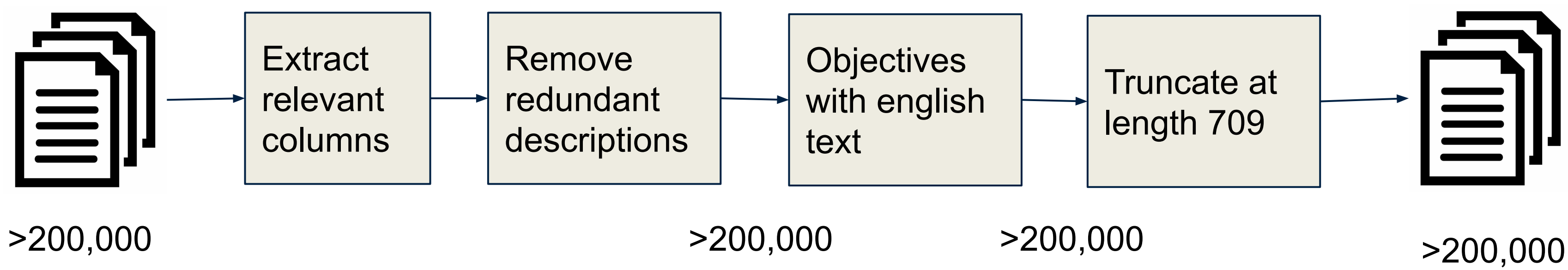


Job Description Dataset

sebis

Column Name	**Column Value**
Id	**12612628**
Title	**Engineering Systems Analyst**
Full Description	**Engineering Systems Analyst Dorking Surrey Salary ****K Our client is located in Dorking, Surrey and are looking for Engineering Systems Analyst our client provides specialist software development Keywords Mathematical Modelling, Risk Analysis, System Modelling, Optimisation, MISER, PIONEEER Engineering Systems Analyst Dorking Surrey Salary ****K**
Location	Dorking
Contract Time	permanent
Contract Type	full_time
Company	Gregory Martin International
Category	Engineering Jobs
Salary	20000 - 30000/annum 20-30K

Job Description Dataset Analysis Summary



Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

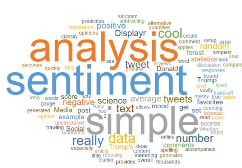
Topic Modeling

Clustering | **Topic word retrieval**

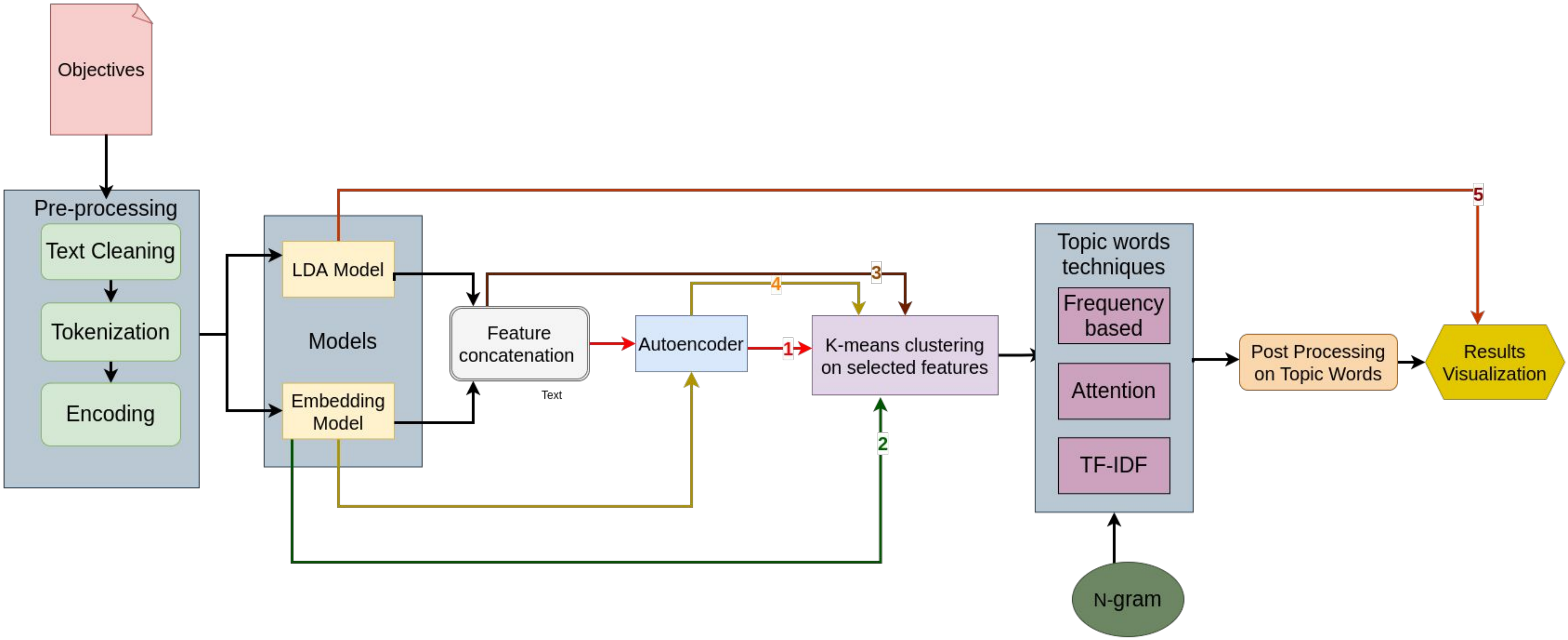
K = 1

K = 2

K = 5



Methodology - Block Diagram



A Block diagram of the methodology used in this study. An objective document is passed through a Pre-processing step first, then one of five type of feature spaces is selected which is used of clustering. Next, one of the three topic word retrieval technique is selected to get the top topic words. Last step is post processing to remove redundant topics from clusters.

Agenda

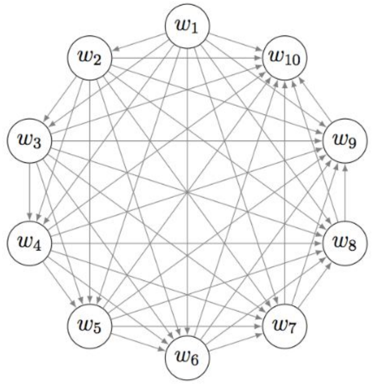
- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Analytical Measures

1) Coherence Score:

- Measures the degree of similarity between topics in a cluster.
- Outputs a value between 0 and 1.

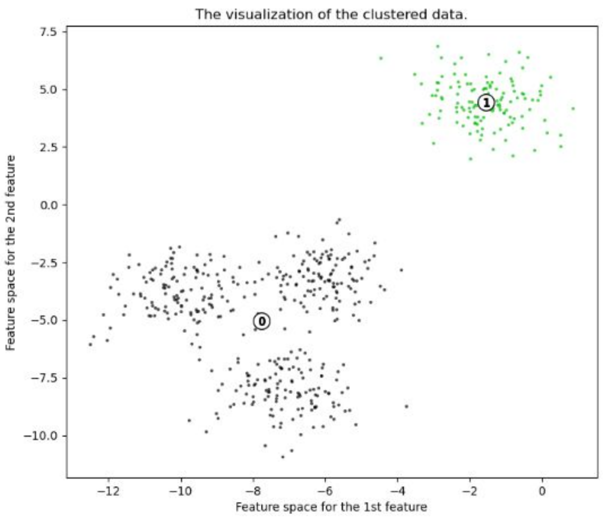
$$\text{score}_{\text{UMass}}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

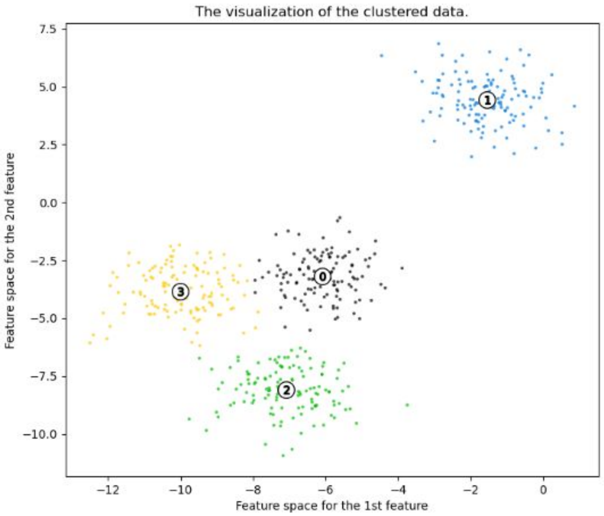


2) Silhouette Score:

- Examines the compactness of the data point features within a cluster and how well the clusters are separated from each other.
- Outputs a value between +1 and -1.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$





Results

Experiments using Embedding Models on Job Description Dataset

Embedding Model	Coherence Scores	Silhouette Score
Sentence BERT	0.5333	0.0638
BERT	0.5857	0.1321
XLNET	0.4938	**0.1324**
Sentence RoBERTa	0.5499	0.0427
ELECTRA	0.5743	0.1169
Sentence DistilBERT	0.6040	0.0435
XLM	**0.6118**	0.0744

Coherence scores and Silhouette Scores when using feature space from Embedding Models for clustering before post-processing step with 10 clusters, frequency-based topic word retrieval approach on Job Description dataset

Results

Experiments using all feature spaces on Job Description Dataset

sebis

Feature Spaces	Coherence Scores	Silhouette Score
LDA model	0.4629	N/A
Embedding Model	0.5333	0.0638
Embedding Model + LDA	0.6001	0.0745
Embedding Model + Autoencoder	**0.6280**	0.1392
Embedding Model + LDA + Autoencoder	0.6083	**0.2456**

Coherence scores and Silhouette Scores when each feature space is used for clustering. Score are captured before post-processing step with 10 clusters, frequency-based topic word retrieval approach on Job Description dataset.

Results - Human Evaluation

System Effectiveness

Layout Adequacy

Results Quality

Explainability and Transparency

Effectiveness of Visualizations

Use Intention

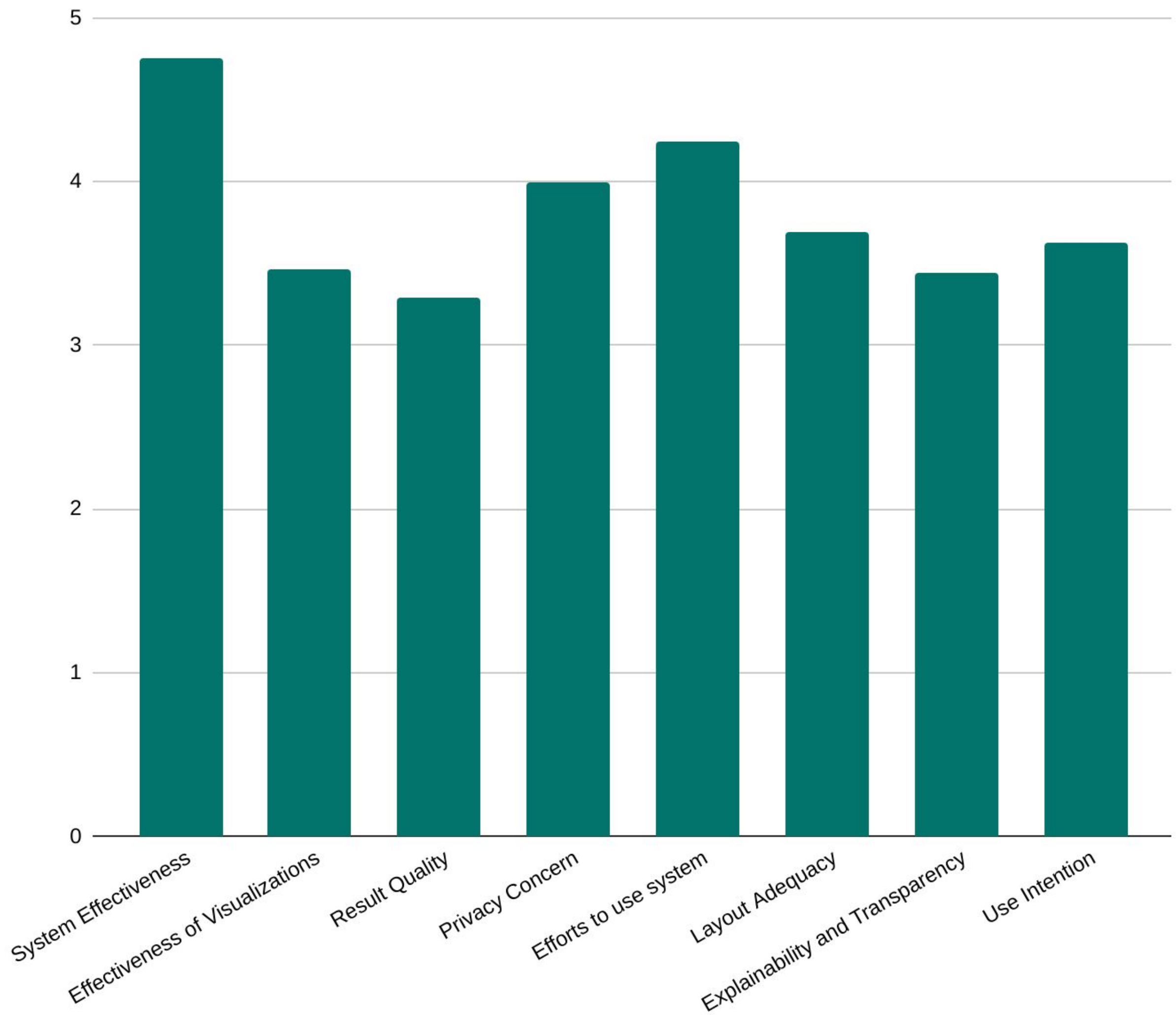
Privacy Concerns

Effort to use the system

The questionnaire consists of 16 statements, divided into 8 evaluation aspects.

Results - Human Evaluation

Summary of Survey



The participant read the statement and expressed their agreement/dis-agreement with the statement on a 1-5 scale: [Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree].

Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Research Questions

RQ1: Could using embedding vectors lead to better results than Latent Dirichlet Allocation model?

Yes, using a Word Embedding model for Topic Model can lead to better results.

RQ2: If the word embedding models are able to provide better results, then which type of embedding model is better suited?

Sentence Transformers Models such as Sentence BERT, Sentence RoBERTa and Sentence DistilBERT provide the best results.

RQ3: Could using a traditional algorithm such as LDA in tandem with the Embedding models provide better results?

No, using Word Embedding in tandem with an LDA model provide almost the same result as using a Word Embedding model alone.

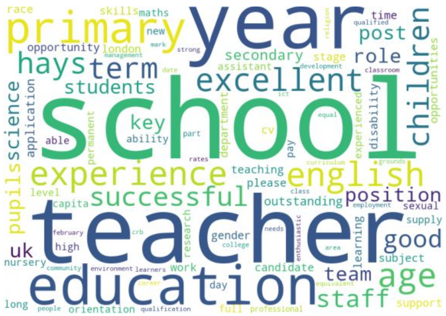
Agenda

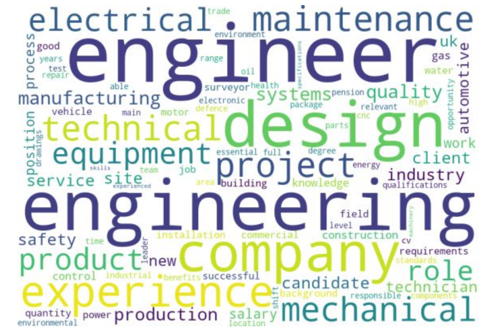
- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Demo

Feature Space from LDA Model

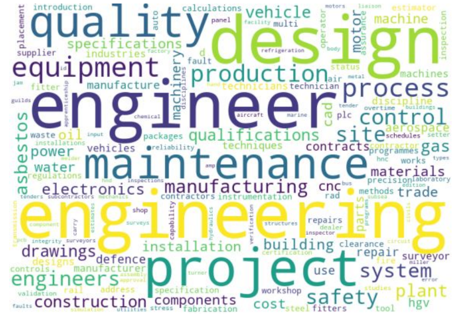
k = 10 on Job Description dataset

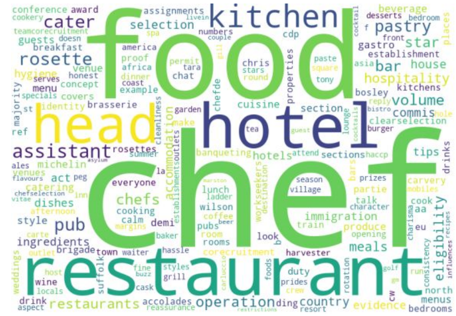




Feature Spaces with Sentence BERT Embedding Model

K = 10, Job Description Dataset





Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion

Discussion - Conclusion

Sentence Transformers such as Sentence BERT and Sentence RoBERTa provide an accurate feature space for clustering.

Clusters obtained using Embedding model is similar to the one obtained from feature concatenation and Autoencoder.

Silhouette score and Coherence score is not a good measure of evaluation for task such as Topic Modeling.

Quality dataset is essential for performing an unsupervised task such as Topic Modeling.

Thank you for your attention!
Questions? Comments?

Anum Afzal
Bsc

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

Boltzmannstraße 3
85748 Garching bei München

anum.afzal@in.tum.de
wwwmatthes.in.tum.de