VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

APPLICATION OF MACHINE LEARNING TECHNIQUES FOR LITHUANIAN ENTERPRISE CLUSTERING

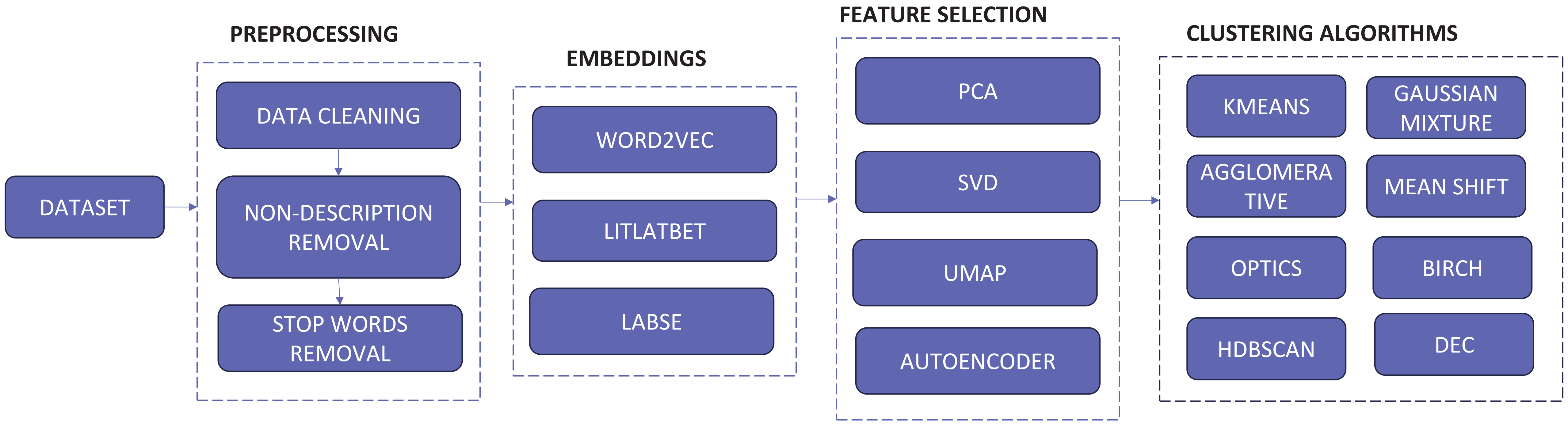
MOTIVATION

- Identification of enterprise activity codes stands as a crucial task enabling establishment or renewal of databases and making informative decision about economic tendencies.
- Gathered insights allow for informative decisions about taxes, needed state-aid and competition analysis.

GOAL

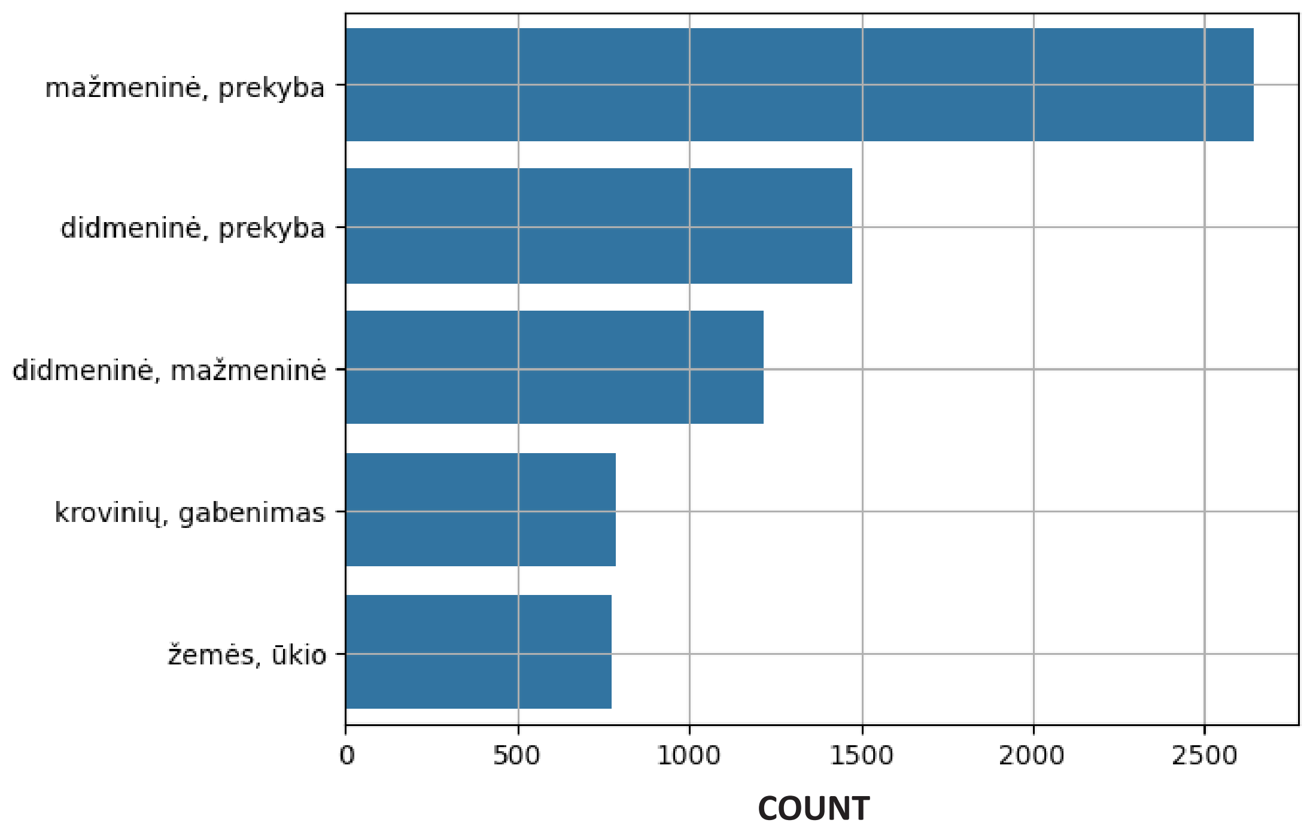
- To apply clustering methods to help identification of the economic activities using descriptions, utilizing predefined **NACE** codes

RESEARCH WORKFLOW

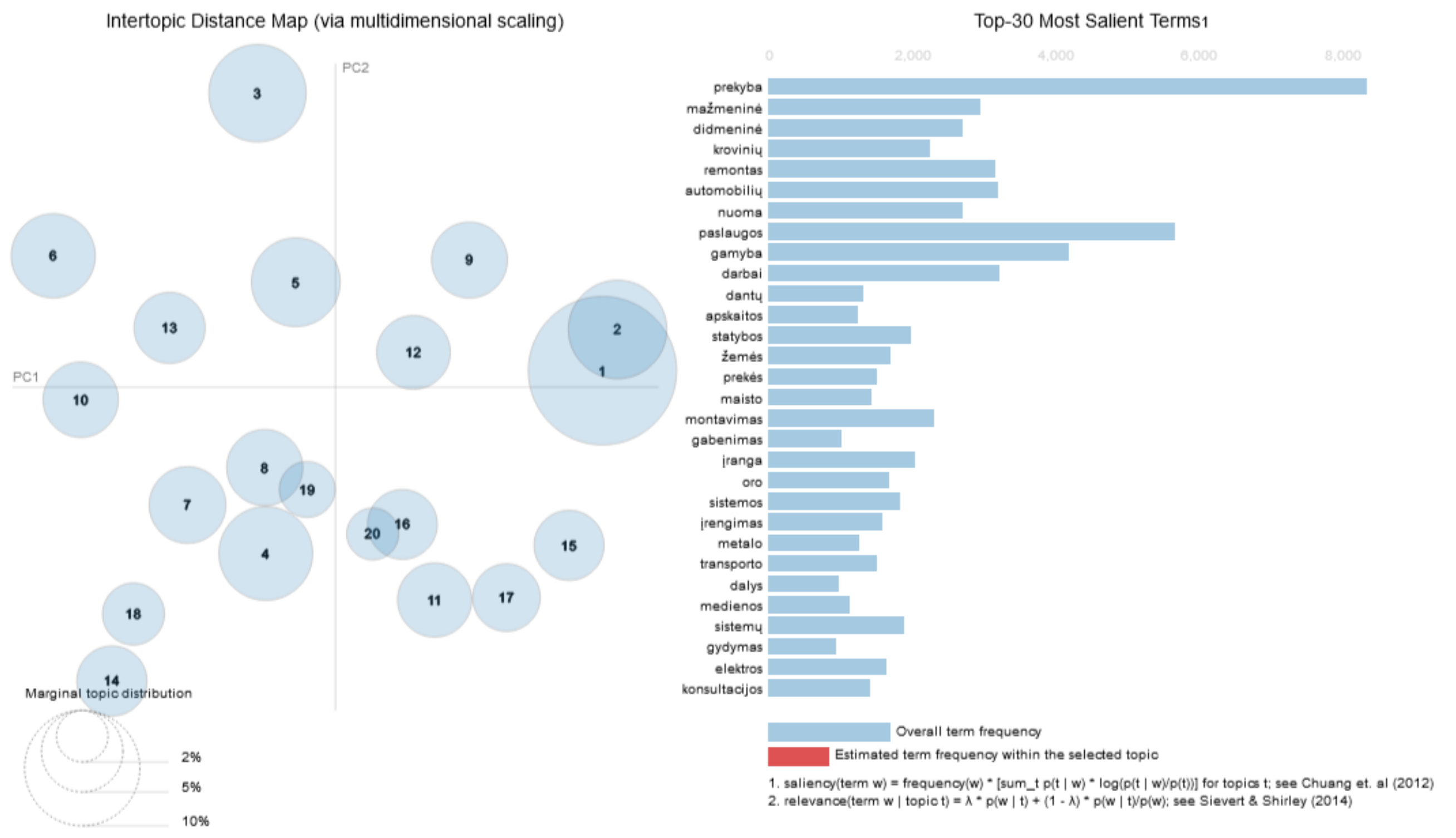


WORDCLOUD

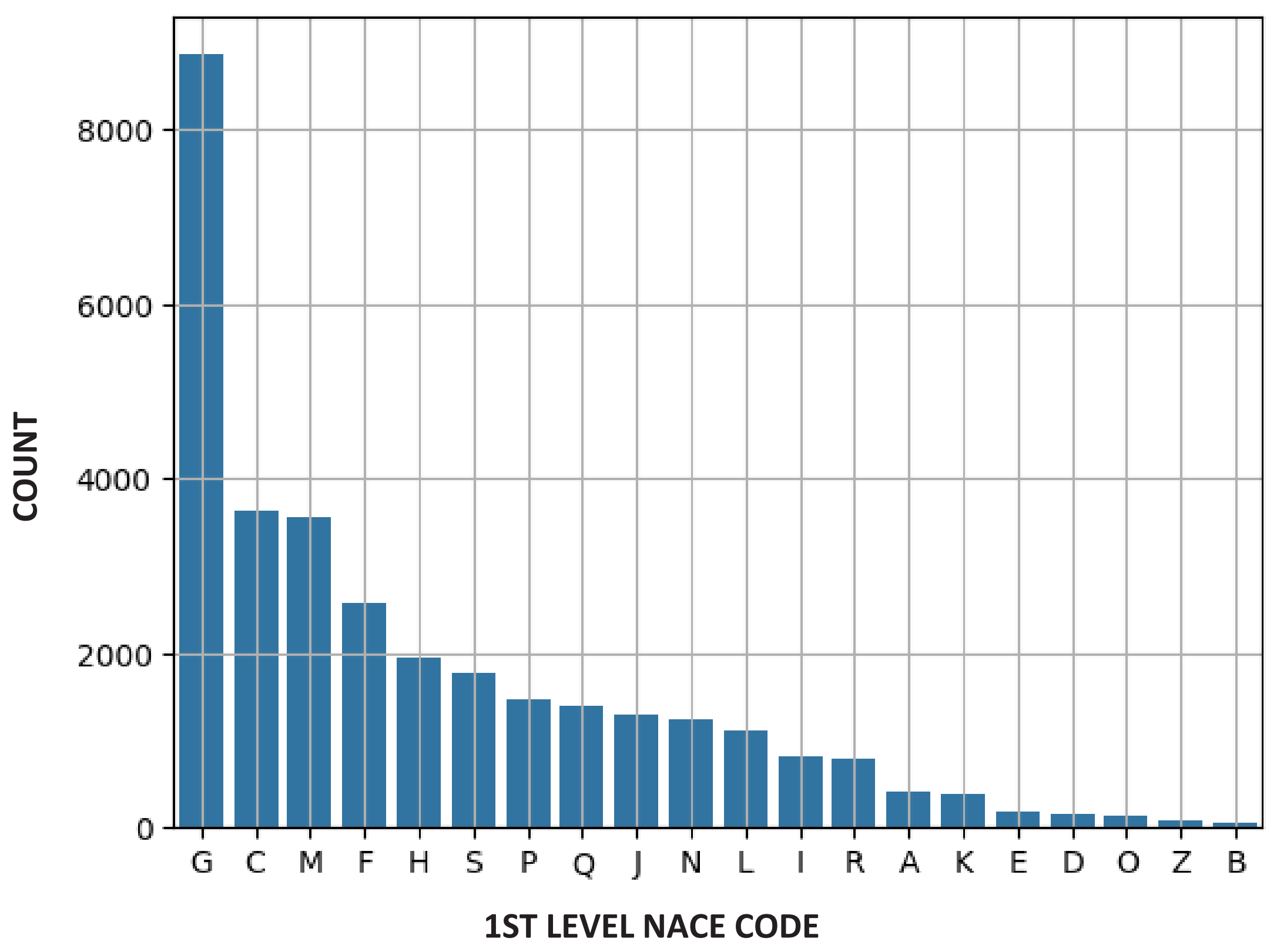
BIGRAMS



TOPIC MODELING



DATASET DISTRIBUTION



AUTHORS:

Eimantas Zaranka
eimantas.zaranka@vdu.lt

Dovilė Kuizinienė
dovile.kuiziniene@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

Co-funded by the European Union

SustAIn Liv Work
CENTRE OF EXCELLENCE OF ARTIFICIAL INTELLIGENCE FOR SUSTAINABLE LIVING AND WORKING

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT

CLUSTERING RESULTS

The best results for each clustering model, where silhouette score > 0

Embeddings	Feature Selection	Clustering Algorithm	# of clusters	Silhouette score	DB Index	CH Index
LaBSE	UMAP	Kmeans	82	0.3764	0.8399	13408.4
LaBSE	UMAP	Agglomerative	82	0.3751	0.8356	12552.3
LaBSE	UMAP	Gaussian Mixture Model	20	0.3351	0.8424	8011.0
Word2Vec	PCA	Mean Shift	14	0.4614	0.8584	3884.9
LaBSE	UMAP	BIRCH	31	0.3495	0.6718	9560.3
Word2Vec	AUTOENCODER	Deep Embedding Clustering	20	0.0679	2.0577	2666.2

DB Index - Davies-Bouldin Index, CH Index - Calinski–Harabasz Index

CONCLUSIONS

- The total of **195** experiments were conducted across all embeddings and feature selection combinations for 4 different levels of **NACE**.
- Experiments showed that for embeddings **LaBSE** and **Word2Vec** and for feature selection **principal component analysis** and **UMAP** are the most promising.
- Most suitable clustering algorithms are **KMeans**, **Agglomerative clustering** and. **Mean Shift clustering.**