# Towards Language Model-based Identification of Market Segments and Fostering of Business Clusters

1th Müller Nicola
s8namuel@stud.uni-saarland.de
2578753

2st Leist Robert
s8roleis@stud.uni-saarland.de
2580448

3nd Eichler Paul
s8pleich@stud.uni-saarland.de
2578569

4rd Recktenwald Tobias
s8tsreck@stud.uni-saarland.de
2577468

5th Nazari Hameed
abna00001@stud.uni-saarland.de
7004543

***Abstract*—With this guide, we give an example of what a write-up for the mini-project report could look like. We provide a LaTeXtemplate that you can use in your own write-up, if you want to. We explain some important points regarding specific types of sections that need to be included in the report.**

## I. Introduction

Deciding on location of a business is a vital strategic decision. Finding the right location can have a number of benefits: First, it is crucial to most businesses to be located close to its customers and suppliers, for more effective marketing or reduced transportation costs, secondly, being located in an active regional cluster can significantly benefit a business development [1]–[4].

Clusters are local concentrations of businesses, manufacturers, suppliers, and institutions from the same field. Identifying regional clusters is done by combining geographical data of companies' locations with information on their corresponding market segments. This geographic proximity of related businesses enables efficient cooperation and knowledge transfer, which has been shown to significantly increase the productivity of all businesses within the cluster [1]–[4]. Thus, establishing or further developing business clusters is of high interest to companies and local governments. Therefore, significant effort is spent identifying clusters and market segments in regional economies.

However, the identification of market segments, and moreover that of business clusters, requires significant amount of manual labor. Typically, one first has to collect information about companies with a location in the targeted region. Afterwards, for each company, their publicly available information must be analyzed to better understand their products, strategic position and market positioning.

This process is very labor-intensive, as businesses do not provide any specific data on the latter points, but it has to be inferred from their product portfolio and websites. Additionally, even simpler data like the number of employees is often unavailable, or only available inside hard to parse company websites, especially for small regional companies.

At the same time, recent advances in natural language processing (NLP), e.g. in the form of the popular ChatGPT [5], have led to surprisingly good results in numerous applications. Especially, they have proven be very efficient at extracting data and summaries from human written texts.

The main contribution of this work is an automatic clustering pipeline to identify market segments and business clusters effectively from publicly available information. Utilizing those clusters, our system is then able to provide a strategic location, which we will demonstrate by using the German federal state Saarland as an example. The novelty of our approach is that we rely on language models to process the companies' information.

Our clustering pipeline can be divided into $4$ distinct steps: $1$) gathering data from the websites of companies in Saarland, $2$) computing embeddings of the data, $3$) reducing the embeddings' dimensionality, and $4$) applying clustering algorithms.

$1$) To construct data points for companies in Saarland, we rely on publicly available descriptions from their websites, allowing us to circumvent the challenges of gathering economic data. Since these descriptions are intended to attract customers and investors, they contain all relevant information for characterizing the companies.

$2$) Given the company descriptions, we utilize state-of-the-art language models to compute embeddings of the text data, corresponding to low-dimensional vectors encompassing the data's essential information.

$3$) We further reduce the dimensionality of the embeddings using non-linear dimensionality reduction techniques, which enables us to preserve the non-linear relationships between the embedded text data while avoiding the curse of dimensionality.

$4$) Lastly, we apply well-established clustering algorithms in the low-dimensional feature space to assign each company to a cluster corresponding to a potential market segment.

Our location recommendation system builds upon the clustering pipeline by computing an embedding of a novel company's description, assigning it to a cluster, and then identifying the most similar companies within the cluster. The system then recommends that the novel company locate its site near an established company acting in the same market segment, which fosters business cluster development in Saarland.

PE: We might want to shorten this part a little as we will explore it in all details later

We provide a graphical user-interface for our methods, which visualizes the clustering of companies in Saarland and their locations in an interactive manner, allowing end-users to analyze the various market segments. Further, the interface supports making location recommendations for new company sites using text inputs.

In summary, our main contributions are:

- We circumvent the challenges of traditional economic data collection by relying on publicly available company descriptions.
- We leverage state-of-the-art language models to apply clustering algorithms to the high-dimensional text data.
- We show how our methods can be extended to handle additional features.
- We demonstrate our approach's feasibility using Saarland's economy and provide a user-friendly graphical interface for our methods, allowing for analyzing market segments and location recommendations for establishing business clusters.

NM: Adjust this when related work is finally done

The remaining parts of this paper are structured as follows: In section II, we first examine the economy in Saarland and the theory behind business clusters. Then we look upon existing work on using clustering for economic analysis, and lastly, we investigate work on language models for extracting information from text data. Next, in section III, we present our methods in more detail, discussing each step of our clustering pipeline and location recommendation system, and in IV, we present how we apply our methods to the economy of Saarland. Lastly, we examine our results in section V and conclude this work in section VI.

## II. Related Work

- General stuff about Saarland's economy and its market segments?
- How market segments are normally identified?
- Theory behind business clusters and maybe related work?
- Related Work on clustering in economy?
- Mention that we collected traditional economic data?

NM: Related work for business stuff is still missing!

### A. Identification and Impacts of Business Clusters

Clusters and their advantage have been studied in many works. For example [3], examined the well known technology cluster in the Silicon Valley, famous for being the home of many innovative startups even in recent years.

[4] provides detailed discussion of the cluster phenomenon. It discusses preexisting literature on clusters and reviews the two main postivies effects commonly observed in clusters, positive feedback, and productivity & growth. Additionally, it provides examples for well-known clusters.

Our region of choice for testing our approach has also been subject of scientific work. It should be noted that most of them examine it for its location on the border of France and Luxembourg or because of its mining and steel producing past. These factors are however of no interest to this work. For example, [6] has focused on the agglomeration of automotive industry in Saarland and the neighboring regions, but focuses on the challenges that the country borders pose to this industry.

The paper [7] specifically compares the economic strategies for development of the regions economy and with a focus on how clusters are developed. Most importantly, they identify some regional clusters. We will revisit these when evaluating our clustering algorithm and compare on the clusters identified by our approach. It should be noted however that the paper was published in 2009, therefore the analysis conducted by this paper was based on the economy at least 14 years before this paper.

PE: This is on the sparse side but going deeper into the techniques would require more in depth research on what are accepted approaches. Would be good if the business people help here

### B. Natural Language Processing

In natural language processing (NLP), translating text data to machine-processable representations is a fundamental challenge for numerous tasks, such as machine translation [8], sentiment analysis [9], and information retrieval [10], due to the complexity and variability of natural language. A popular approach for representing text data is the utilization of so-called *embeddings*: vectorized representations of text with the property that vectors of text data with similar semantic and syntactic properties will have a smaller distance to vectors of text data with different properties. Thus, embeddings enable quantifying semantic and syntactic similarity in terms of distances in a low-dimensional, dense vector space, making them ideal for applying clustering to our sparse dataset of high-dimensional company descriptions.

In general, embeddings are learned by training a neural network to perform a task requiring understanding semantic and syntactic information, such as predicting words given their surrounding contexts or predicting subsequent words in sentences. In this process, the network learns a mapping from its text input to low-dimensional representations, containing the information needed to generate correct outputs. Hence, the outputs of this mapping encode all relevant semantic and syntactic information in the text inputs, meaning that, after training, the network's corresponding layers can be reused to compute embeddings for other downstream tasks.

Mikolov et al [11]. popularized learning word embeddings from raw text data in 2013 by introducing their Word2Vec model. GloVe, developed by Pennington et al. [12], improved upon Word2Vec by effectively encoding both local context information and global statistical information into the embeddings, thereby offering more detailed semantic capture. More recently, transformer-based models like BERT, developed by Devlin et al. [13], and GPT models by OpenAI [14] have leveraged context-dependent word embeddings by encoding the words themselves and the context in which they are used. So-called sentence transformers extended these approaches

by computing embeddings for entire sentences to capture more contextual information than individual word embeddings. Reimers and Gurevych [15] introduced the Sentence-BERT model consisting of two Siamese BERT-networks, i.e., two BERT-networks with tied weights, which compute sentence embeddings by embedding each word individually and then applying a pooling layer calculating the maximum or mean over all word embeddings. Starting with pre-trained BERT-networks, Sentence-BERT is fine-tuned using text classification or cosine-similarity tasks, yielding a model capable of generating high-quality sentence embeddings using only a fraction of the original BERT model's computational effort.

PE: Maybe we should bring out the connection to our approach here a little bit more

## III. Procedure

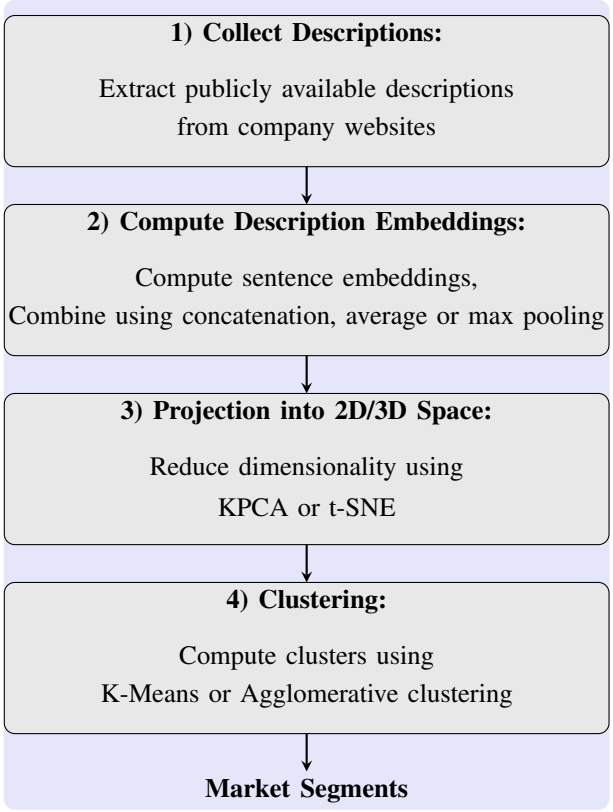


Fig. 1. Overview of our clustering pipeline.

PE: s/Market Segments/Local Clusters; Market segment seems to be defined as segmenting your customer base wiki market segmentation

Our goal is to collect data points on companies in Saarland and cluster them to identify market segments and recommend locations for companies who plan to establish new sites, such that it supports the development of business clusters. We will present our clustering pipeline and then examine how we can use the results to provide location recommendations.

### A. Identifying related companies through Clustering

Figure 1 shows a graphical representation of our general clustering pipeline, consisting of 4 steps: 1) Collection of descriptions from company websites, 2) converting the descriptions to embeddings, 3) reducing the dimensionality of the embeddings, and 4) applying a clustering algorithm. In the following, we will examine each of these steps in detail.

1) *Collect Descriptions:* The traditional approach for building such a dataset would be to gather several attributes, like revenue, number of employers, and products, for many companies in Saarland and then store each in a fixed-size vector.

PE: TODO: backup with accepted approach

However, it is not straightforward to encode non-numerical attributes, like products and customer base, such that clustering algorithms can parse them without introducing an unwanted ordering of the attributes' values. Further, some attributes may be unavailable for some companies, meaning we would need to discard them because we require fixed-size vectors, or we would need to insert dummy values, which could lead to unwanted effects.

Lastly, representing a company using a small number of publicly available attributes might be insufficient for capturing all relevant characteristics. For instance, consider the smartphone manufacturers Apple and Samsung: Both companies build smartphones, have over 100.000 employees, have revenues of over 200 billion dollars, and have production and sales sites worldwide. Given these attributes, one could argue that Apple and Samsung are very similar companies. Yet, it is commonly known that Apple, with its minimalist design, high-end smartphones targeted toward customers looking to differentiate themselves from other people, and Samsung, with its more utilitarian smartphones targeted toward a broad customer base, are companies with fundamentally different characters.

PE: This is a very shallow argument to me, would propose to remove

To address these challenges to dataset construction, we propose representing companies by the description on their websites instead of numerical attributes. The benefit of using this data is that companies usually design their website to appeal to a very targeted audience. Additionally, they usually include very detailed information on their products, their size and capabilities, as they want to present themselves publicly in a way that attracts customers and investors. Using the text directly circumvents the problem of creating a targeted aggregator for specific information. This makes our approach simpler to implement and a lot more scalable & automatable for larger target regions.

2) *Compute Description Embeddings:* To achieve the fixed-size inputs required by clustering algorithms, we map each company description to an embedding that describes the data in a low-dimensional space. We rely on a sentence transformer neural network architecture to compute embeddings for each sentence in a company's description. We propose three ap-

proaches to attain a single embedding for each description: we repeatedly concatenate individual sentence embeddings until a fixed size is achieved, take the average over each feature dimension, or take the maximum over each feature dimension. The first approach, which we will call concatenation, preserves each sentence embedding's information but may introduce bias and high dimensionality, whereas the second and third approaches, which we will call average and max pooling, achieve a lower dimensionality but may discard relevant information by only keeping the average of the features or the largest features.

3) *Projection into 2D / 3D Space:* Although the embeddings are of much lower dimensionality than the original text data, the number of features might still be too large relative to the size of our dataset. This might invoke the curse of dimensionality, meaning that the few data points are so distant from each other in the high-dimensional space that it is impossible to derive meaningful clusters. Further, clustering data in high-dimensional space prevents the visualization of clustered data points. To avoid the course of dimensionality and to enable user-friendly visualization of our results, we will project the embeddings to 2D or 3D space using two dimensionality reduction techniques.

*Kernel Principal Components Analysis* (KPCA) is a dimensionality reduction technique that computes a linear transformation that projects the data in the kernel space onto the dimensions with the highest variance, such that a non-linear transformation in the original feature space is attained. Hence, by projecting the data on the dimensions with the highest variance, KPCA preserves the most relevant information. The *t-distributed stochastic neighbor embedding* (t-SNE) technique computes a probability distribution representing the pairwise similarity between data points in the feature space and then projects the data points to a low-dimensional space according to a second probability distribution that minimizes the Kullback-Leibler divergence to the first distribution. This enables t-SNE to preserve local structures in the data accurately. The crucial difference between KPCA and t-SNE is that KPCA computes a deterministic projection that can be applied to novel data points, whereas t-SNE does not. However, the preservation of local structures by t-SNE might be more suitable for identifying clusters.

4) *Clustering:* We will now present two clustering approaches that depend on the choice of dimensionality reduction technique. If KPCA is used, we cluster the data using the $K$-Means algorithm, which, starting from $K$ randomly initialized cluster centers, assigns each data point to the nearest cluster, updates the values of the cluster centers as the mean of each data point in the cluster, and then repeats until the cluster assignments do not change anymore. We combine $K$-Means with KPCA since it returns a set of cluster centers that can be used to classify novel data points. If t-SNE is used, we cluster the data using agglomerative clustering, which initially assigns each data point to a separate cluster and then iteratively merges close clusters until all data points are in the same cluster. The intermediate clustering that achieves the largest overall separation between data points can then be chosen as the final clustering. The advantage of agglomerative clustering over $K$-Means is that it does not require a pre-specified number of clusters and can compute hierarchically structured clusters. Once the clustering algorithm converges, we receive cluster labels for each company in our dataset, corresponding to the predicted market segments.

Although our general pipeline represents companies using their descriptions, we emphasize that our approach can also handle utilizing other features. For instance, we can include the companies' products in the encoding by computing the corresponding sentence embeddings and merging them with the description embedding using concatenation, average or max pooling. In section V, we will investigate the results of combining company descriptions with other features.

Figure 2 depicts the clustering analysis tab of our user-interface. On the left side, users can define how the description embeddings are computed, the technique and target dimensionality of the dimensionality reduction, and the number of clusters. Additionally, users can specify a subset of additional features, including industry, products, customer base, market positioning, and revenue, which will be combined with the descriptions. After clicking "Compute clustering", the clustered data points are shown on the right side as a 2D or 3D scatter plot. Further, users can see the location and clustering assignment of all companies in our dataset on a map. When hovering the mouse over a company's data point, additional information about the company is shown, and when double clicking the labels on the legend, only the companies of the corresponding cluster are shown.

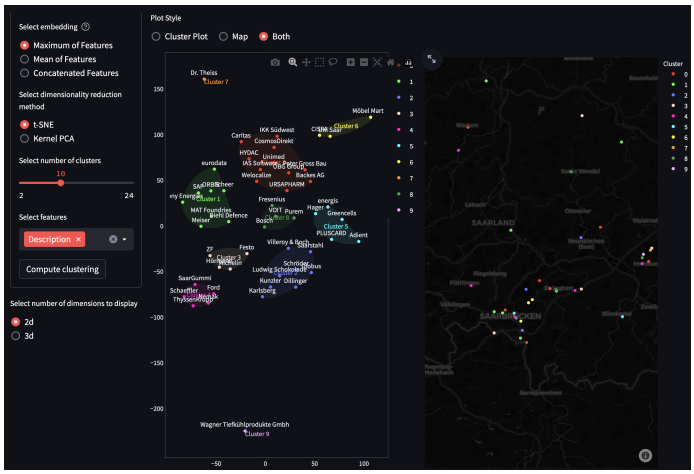


Fig. 2. Clustering analysis page of our user-interface.

## B. Fostering Business Clusters through Location Recommendations

We utilize parts of our clustering pipeline to give location recommendations for companies seeking to establish new sites in Saarland. Given the description of the novel company, we compute an embedding using a sentence transformer and

apply KPCA or t-SNE to get a low-dimensional feature vector. We then assign the feature vector to a cluster according to the $K$-Means or Agglomerative clustering algorithm and then compute which company within the cluster is most similar to the novel company. Note that when using KPCA and $K$-Means, we can assign the novel data points to a pre-computed clustering, and when using t-SNE and Agglomerative clustering, we compute a new clustering. Thus, the second approach accounts for the possibility that introducing a novel company might change existing market segments, whereas the first approach does not. Given the cluster label of the novel company, the location recommendation is to establish the new company site nearby the most similar company within the cluster, meaning we recommend placing the novel site near the most similar company acting in the same market segment. Thus, our location recommendations support the geographical concentration of companies acting in the same market segment, which helps to foster business clusters in Saarland.

Our user-interface offers a dedicated tab for location recommendations. In figure 3 we see the text boxes where users can provide their company's name and description. Additionally, users can specify their company's industry, products, customer base, market position, and revenue which can be used to support the location recommendations. After clicking "Submit", users are shown the output depicted in figure 4. The table at the top shows the names and features of the companies in the cluster, to which the user's company was assigned. The scatter plot at the bottom left shows the clustered data points, where the point of the user's company is highlighted. Lastly, the map at the bottom right displays the location of the novel company's cluster members and the distances between them. The user can also overlay a heatmap which show the similarity between his company and all other companies in out dataset.

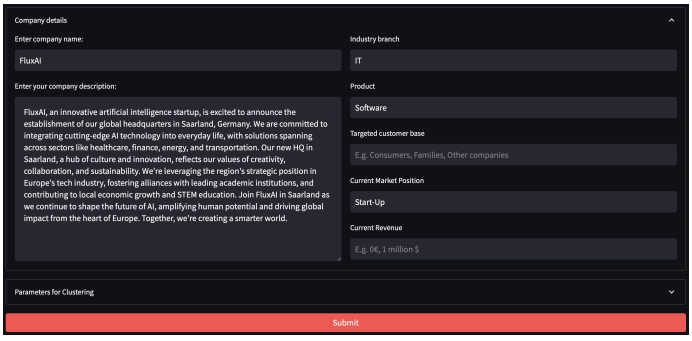


Fig. 3. Input to the location recommendation feature of our user-interface.

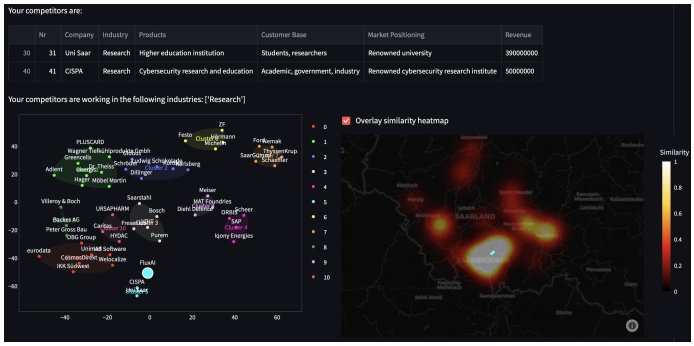


Fig. 4. Output of the location recommendation feature of our user-interface.

## IV. Experiments

To build our dataset of company descriptions, we manually extracted the descriptions of 50 of the largest companies in Saarland. When a description was only available in German, we translated it to English using the online translator DeepL. A list of all companies in our dataset can be found in the appendix D.

For the purpose of the experimental evaluation we conducted the data collection manually. As our target region was very narrow this was easily feasible. However, in the future we might create a simple HTML scraper to scale our approach to larger regions. To test our location recommendation feature, we created company descriptions using GPT-4 [5].

Additionally, we evaluate the impact of providing additional features to the clustering pipeline,for which also gathered the features industry, products, customer base, market positioning, and revenue for each company.

We utilize the popular pre-trained sentence transformer "all-MiniLM-L6-v2" from the HuggingFace hub to compute sentence embeddings for each description. An embedding for the full description is then attained using our concatenation, average, or max pooling approach. When using additional features, we compute embeddings for each feature and combine them with the description embedding in the same way the individual sentence embeddings were combined.

PE: Links and references plz

Afterward, the embeddings are projected onto 2D or 3D space using KPCA or t-SNE. We chose suitable hyperparameters for KPCA by conducting a grid search over the values $kernel = \{radial\ basis\ function, sigmoid, linear, polynomial\}$ and $gamma = \{0.03, 0.032, 0.034, \ldots, 0.05\}$, where we found the best values to be $kernel = polynomial$ and $gamma = 0.05$. For t-SNE, we manually fine-tuned the perplexity hyperparameter until we achieved a suitable separation of data points.

To choose the number of clusters for $K$-Means, we started with $K = \sqrt{N}$, where $N$ is the size of the dataset, and then increased $K$ until the clustering aligned with our expectations. For Agglomerative clustering, we chose the final number of clusters through visual inspection of the dendrogram returned by the algorithm.

## V. Results

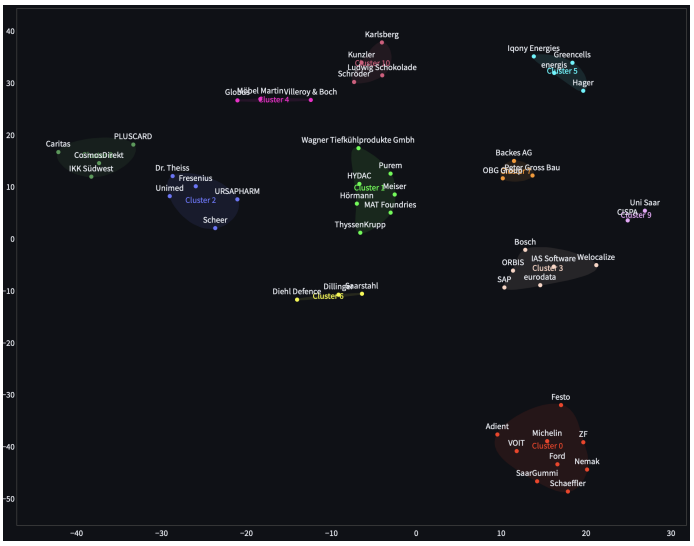


Fig. 5. Clustering of company descriptions using t-SNE and Agglomerative clustering. The company descriptions and industries were used as features.

In the following, we will investigate the results of the best-performing configuration of our clustering and location recommendation system. These results were obtained using the following configuration: the description embeddings were computed using max pooling, the embeddings were projected into 2D space using t-SNE, and the clustering was achieved using Agglomerative clustering. We also used the companies' industry as an additional feature. We found that this configuration yields the most sensible predictions of market segments and location recommendations. Figure 5 displays the resulting clustered data points.

### A. Clustering

In the following, we analyze each of the found clusters and interpret them with respect to potential market segments:

- *Cluster 0 (red)*: This cluster consists of automotive manufacturers and suppliers, such as Ford and Michelin. Hence, we follow that this cluster represents the automotive industry market segment.
- *Cluster 1 (light green)*: The steel industry companies MAT Foundries, ThyssenKrupp, and Meiser are part of this cluster. Further, cluster members are companies like Hörmann and HYDAC, which also offer products manufactured from steel. Accordingly, we assign this cluster to the steel products market segment. The food manufacturer Wagner can be considered misclassified, as it should be contained in Cluster 10.
- *Cluster 2 (dark blue)*: We attribute this cluster to the healthcare market segment since it contains the medicine manufacturers Dr. Theiss and URSAPARM, Fresenius, which produces dialysis machines, and Unimed, which provides accounting services for hospitals. Note that the IT consulting firm Scheer can be considered misclassified and should be added to Cluster 3.
- *Cluster 3 (beige)*: This cluster contains the IT companies SAP, ORBIS, eurodata, IAS Software, and Welcoalize, representing the IT products and IT consulting market segment. Although the cluster member Bosch is not an IT company, it also offers software products, so we do not consider it misclassified.
- *Cluster 4 (purple)*: This cluster consists of the companies Globus, Möbel Martin, and Villeroy&Boch, which offer very different products but are all large retailers. Hence, we assign this cluster to the retail market segment.
- *Cluster 5 (light blue)*: The energy provider energis and the energy service providers Iqony Energies and Greencells are contained in this cluster. Hager is also a cluster member specializing in electrical installation and energy distribution. Accordingly, we infer that this cluster represents the energy market segment.
- *Cluster 6 (yellow)*: We assign this cluster to the steel production market segment since it contains the steel manufacturers Dillinger and Saarstahl. The defense company Diehl Defence can be considered misclassified, as it should be in a single cluster because it is the only defense company in our dataset.
- *Cluster 7 (orange)*: This cluster consists of the construction companies Backes AG, Peter Gross Bau, and OBG Group, so it represents the construction work market segment.
- *Cluster 8 (dark green)*: Because the insurance companies CosmosDirekt and IKK Südwest are part of this cluster, it could indicate the insurance market segment. However, this cluster also contains the healthcare service provider Caritas, which should be in Cluster 2, and the credit card provider PLUSCARD.
- *Cluster 9 (lavender)*: The only two members of this cluster are Uni Saar and CISPA, making it representative of the academia market segment.
- *Cluster* 10 *(pink)*: This cluster consists of food producers such as Kunzler and Ludwig Schokolade, and thus it represents the food market segment.

We conclude that our clustering methods are effective since we saw that each of the 11 clusters mostly contained similar companies, allowing us to infer various market segments.

### B. Comparison to clusters identified by other work

As mentioned [7] also identified clusters for the region of Saarland. Specifically, they identified the following clusters by industry ( [7], Table 4, p. 1226)

- Coal and mining industries
- Metal industry
- Automobile industry
- Energy sector
- Information and communication technology
- Biotechnology and nanotechnology

Comparing these results with the output of our pipeline we can see that

- Coal and mining industries: With the decline of coal mining in recent years, there are no more coal mines in Saarland. As the last one closed in 2012, this cluster does not exist anymore.
- Metal industry is covered by cluster 1 (light green) and cluster 6 (yellow).
- Automobile industry is covered by cluster 0 (red).
- Energy sector is covered by cluster 5 (light blue).
- Information and communication technology is covered by cluster 3 (beige).
- Biotechnology and nanotechnology is covered by cluster 2 (dark blue).

One can see that the results of our pipeline do agree with the general sectors identified by [7]. However, our approach did lead to more fine-grained separation. Additionally, we for example include construction companies, leading to additional companies. However, these sectors are usually not contributing significantly to the region's economy, which is most probably why they are not mentioned in [7]. In summary, the output of our pipeline matches very well with traditionally identified clusters.

### C. Location Recommendation

To test our location recommendation system, we generated descriptions for four fictional German companies with vastly different characteristics:

- *FluxAI*: A young artificial intelligence start-up with ambitious goals of revolutionizing healthcare, finance, energy, and transportation.
- *ABC Auto*: A well-established automotive manufacturer offering a wide range of cars focusing on sustainability.
- *Bratwurst Bliss*: A hot dog manufacturer that values traditional sausage making and locally sourced ingredients.
- *PanzerTech*: A weapons manufacturer with a wide spectrum of products, from armored vehicles to cybersecurity solutions.

For FluxAI, our system recommends placing the new site close to CISPA, which is sensible since CISPA is a world leader in information security research, which also includes artificial intelligence, and thus knowledge transfer between both companies could be highly beneficial. It also makes sense that our system did not recommend locating FluxAI near IT companies with a more commercial focus, like SAP, since they primarily work on existing technology instead of revolutionary innovations. For ABC Auto, the location recommendation is to establish a new site close to VOIT, which specializes in manufacturing car components with a focus on hybrid and electric cars. Thus, placing ABC Auto near VOIT would allow ABC Auto to source parts for its hybrid and electric cars from a company nearby, leading to a more robust supply chain. For Bratwurst Bliss, the system recommends settling near the meat producer Kunzler, which could lead to knowledge transfer regarding sausage making. Lastly, our system recommends placing PanzerTech's new site close to Diehl Defence, corresponding to the only weapons manufacturer in our dataset. Hence, this shows that our system does not need a large number of samples from each market segment to give reasonable location recommendations.

We conclude that our location recommendation system can reliably identify companies in our dataset with similar characteristics to companies seeking to establish sites in Saarland and thus can be used to foster the development of business clusters.

### D. Limitations

We observed in subsection V-A that a small number of companies were assigned to the wrong clusters. These companies mostly belonged to sectors that are underrepresented in our dataset, such as defense and financial services, so we would likely achieve more concise cluster assignments if we added more companies from these sectors to our dataset.

Another reason for invalid cluster assignments could be that our utilized sentence transformer was not fine-tuned for company descriptions, so that it could have produced inaccurate embeddings for some descriptions. Further, projecting the embeddings into 2D or 3D space could also cause inaccurate cluster assignments due to a large loss of information. Projecting the embeddings into higher dimensions might avoid this information loss, yet it would complicate or prevent a user-friendly visualization of the clustering results.

## VI. Conclusion

The goal of this project was to collect data on companies in Saarland and apply clustering algorithms such that the results can be used to identify regional clusters and give location recommendations for novel companies such that business clusters are fostered.

To circumvent the challenges of feature selection and encoding non-numerical features, we represented companies by their publicly available descriptions. We leveraged state-of-the-art language models to transform this highly informative text data into machine-readable embeddings. Lastly, we utilized two approaches to project the embeddings into low-dimensional space where clustering can be applied. Further, our methods allowed straightforwardly including other standard company attributes, such as industry, customer base, market positioning, and revenue. To make location recommendations, we passed the descriptions and features of novel companies to our clustering pipeline and then recommended establishing new sites next to the most similar companies in our dataset.

A qualitative evaluation of our methods has shown that they can reliably cluster similar companies, which allowed us to infer major industry segments in Saarland. Further, our location recommendation system demonstrated its capability of identifying similar companies, given the description of a new company, independently of the total number of similar companies in the dataset.

To allow end-users to leverage our methods, we implemented a user-friendly interface with many functionalities, such as visualizing the clustered data points, modifying the

clustering parameters, showing the location of companies, and giving detailed and interpretable location recommendations.

We see many opportunities for future work. Collecting data from more companies in Saarland could improve the clustering results, although the number of (large) companies in Saarland is limited. Thus, applying our methods to larger economies would be of great interest. This would also allow more comparisons to traditionally identified clusters. Further, fine-tuning our language models using text data from the business sector might lead to more expressive embeddings. To achieve this, efficiently, an automated aggregation system could be developed. As we utilize publicly available websites, this would be easily possible without utilizing 3rd party services. To this end, we have demonstrated the feasibility and potential of applying language models to economic analysis.

## REFERENCES


[1] M. Porter, "Clusters and the new economics of competition," *Harvard business review*, vol. 76, no. 6, p. 77—90, 1998. [Online]. Available: http://europepmc.org/abstract/MED/10187248

[2] S. Kamath, J. Agrawal, and K. Chase, "Explaining geographic cluster success—the gems model," *American Journal of Economics and Sociology*, vol. 71, no. 1, pp. 184–214, 2012. [Online]. Available: http://www.jstor.org/stable/23245182

[3] M. S. Gertler, P. Oinas, M. Storper, and P. Scranton, "Discussion of "regional advantage: Culture and competition in silicon valley and route 128" by annalee saxenian," *Economic Geography*, vol. 71, no. 2, pp. 199–207, 1995. [Online]. Available: http://www.jstor.org/stable/144358

[4] A. Kuah, "Cluster theory and practice: Advantages for the small business locating in a vibrant cluster," *Journal of Research in Marketing and Entrepreneurship*, vol. 4, pp. 206–228, 10 2002.

[5] OpenAI. (2023) Introducing chatgpt.

[6] C. K. Hahn, "The transboundary automotive region of saar-lor-lux: Political fantasy or economic reality?" *Geoforum*, vol. 48, pp. 102–113, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016718513000948

[7] M. Trippl and A. Otto, "How to turn the fate of old industrial areas: A comparison of cluster-based renewal processes in styria and the saarland," *Environment and Planning A*, vol. 41, pp. 1217–1233, 05 2009.

[8] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" *arXiv preprint arXiv:1804.06323*, 2018.

[9] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Systems with Applications*, vol. 117, pp. 139–147, 2019.

[10] X. Ye, H. Shen, X. Ma, R. Bunescu, and C. Liu, "From word embeddings to document similarities for improved information retrieval in software engineering," in *Proceedings of the 38th international conference on software engineering*, 2016, pp. 404–415.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.


## APPENDIX

NM: Upload all LaTeX files to overleaf when its done since they want a link! Update readme in repo!

- **Link to Code**: Provide the link to your code repository. You can use any service to host your code (Github, gdrive etc.). The code should be running, accessible and all of your work must be reproducible. In case your code is not accessible or not in a working state, relevant number of points will be deducted.
- **Link to Overleaf**: Add a link to the Overleaf project which contains all of the LaTex related files for this report.
- **Additional Figures**: Add any and all additional figures produced during you work with relevant explanations in the appendix.

### A. Additional Clustering Results

### B. Link to Code

https://github.com/NicolaMueller42/data_science_project

### C. Link to Overleaf

### D. Companies in our Dataset

We list the 50 Saarland companies in our dataset:

- ZF
- Saarstahl
- Dillinger
- Bosch
- Festo
- Ford
- Schaeffler
- Fresenius
- Wagner
- Villery&Boch
- Michelin
- ThyssenKrupp
- Hager
- Purem
- SaarGummi
- VOIT
- Meiser
- SAP
- Nemak
- Peter Gross Bau
- OBG Group
- URSAPHARM
- Hörmann
- Kunzler
- Diehl Defence
- HYDAC
- Backes AG
- Karlsberg
- Globus
- Caritas
- Uni Saar
- Welocalize
- Möbel Martin
- IKK Südwest
- Dr. Theiss
- Unimed
- CosmosDirekt
- ORBIS
- eurodata
- energis
- CISPA
- Greencells
- Scheer
- Iqony Energies
- PLUSCARD
- IAS Software
- Ludwig Schokolade
- MAT Foundries
- Adient
- Schröder