Towards Language Model-based Identification of Business Clusters

Submission Type: Research in Progress

Abstract

One of the most important decisions every business has to decide on a physical location. Identifying the right location is crucial for effective marketing strategies reaching the intended customer base, it can reduce transportation costs, and locating in a vibrant regional business clusters can have significant positive effects on business development. However, strategically acquiring & analyzing data on regional economies, the corresponding companies along with their target market segments is very complex and labor-intensive. We propose to circumvent these challenges by leveraging state-of-the-art language models to transform publicly available company websites, into machine-readable data points and applying clustering to make location recommendations. We demonstrate accuracy of identified clusters by applying our method to data from the German municipality Saarland.

Introduction

Deciding on the location of a business is a vital strategic decision: First, most businesses must be located close to their customers and suppliers for more effective marketing or reduced transportation costs. Secondly, being located in an active regional cluster can significantly benefit business growth [Porter(1998),Kamath et al.(2012)Kamath, Agrawal, and Chase,Gertler et al.(1995)Gertler, Oinas, Storper, and Scranton,Kuah(2002)].

Business clusters are local concentrations of companies, manufacturers, suppliers, and institutions from the same field. Identifying clusters combines geographical data of companies' locations with information on their corresponding market segments. This proximity of related companies enables efficient cooperation and knowledge transfer, which has been shown to significantly increase the productivity of all companies within the cluster [Porter(1998), Kamath et al.(2012)Kamath, Agrawal, and Chase,Gertler et al.(1995)Gertler, Oinas, Storper, and Scranton,Kuah(2002)]. Thus,

establishing or further developing business clusters is of high interest to companies and local governments. Therefore, significant effort is spent identifying clusters in regional economies.

However, them requires a significant amount of manual labor. Typically, one must first collect information about companies in the targeted region. Afterward, for each company, their publicly available information must be analyzed to understand their products, strategic position, market positioning, and further characteristics. This process is labor-intensive, as businesses do not provide specific data on the latter points, but it must be inferred from their product portfolio and websites. Additionally, even basic data like the number of employees is often unavailable or difficult to collect automatically, especially for small regional companies. At the same time, recent advances in natural language processing (NLP), e.g., the popular ChatGPT [OpenAI(2023)], have yielded surprisingly good results in numerous applications. Especially they have proven to 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 from publicly available information. Additionally, our system can utilize the found clusters to provide strategic location recommendations, which we will demonstrate 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, allowing us to circumvent many of the challenges attributed to gathering economic data. Our clustering pipeline can be divided into 4 distinct steps: 1) gathering descriptions from the websites of companies in the region of interest, 2) computing embeddings of the text data, 3) reducing the embeddings' dimensionality, and 4) applying clustering algorithms. 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. We provide a graphical user-interface for our methods, which visualizes the clustering of companies and their locations in an interactive manner, allowing end-users to analyze the various predicted market segments. Further, the interface

supports making location recommendations for new company sites using text inputs.

In summary, our main contributions are:

- Circumventing the challenges of economic data collection by relying on public text data.

- Leveraging state-of-the-art language models to apply clustering algorithms to text data.

- Showing how our methods can be extended to handle additional features.

- Demonstrating our approach's feasibility using Saarland's economy and providing a user-friendly graphical interface.

Related Work

This section will examine previous work on business clusters and recent advances in natural language processing.

Business clusters and their advantages have been studied in many works. For example, [Gertler et al.(1995)Gertler, Oinas, Storper, and Scranton] examined the well-known technology cluster in Silicon Valley, famous for being the home of many innovative start-ups. [Kuah(2002)] provides a detailed discussion of the cluster phenomenon. It discusses preexisting literature on clusters and reviews the main positive effects commonly observed in clusters, positive feedback and productivity & growth. Additionally, it provides examples of well-known clusters. The region of choice for testing our approach, Saarland, has also been the subject of scientific work. It should be noted that Saarland is mostly investigated 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, [Hahn(2013)] has focused on the agglomeration of the automotive industry in Saarland and the neighboring regions but focuses on the challenges that the country borders pose to this industry. The paper [Trippl and Otto(2009)] specifically compares the economic strategies for developing the region's economy with a focus on how clusters are developed. Most importantly, they identify some regional business clusters. We will revisit these when evaluating our clustering algorithm and compare the clusters and market segments identified by our approach. It should be noted, however, that the paper was published in 2009, meaning their analysis is outdated to some

extent.

In natural language processing (NLP), translating text data to machine-processable representations is a fundamental challenge for numerous tasks, such as sentiment analysis [Rezaeinia et al.(2019)Rezaeinia, Rahmani, Ghodsi, and Veisi], and information retrieval [Ye et al.(2016)Ye, Shen, Ma, Bunescu, and Liu], 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, making them ideal for applying clustering to our sparse dataset of high-dimensional company descriptions. Mikolov et al [Mikolov et al.(2013)Mikolov, Chen, Corrado, and Dean]. popularized learning word embeddings from raw text data in 2013 by introducing their Word2Vec model. More recently, transformer-based models like BERT, developed by Devlin et al. [Devlin et al.(2018)Devlin, Chang, Lee, and Toutanova], and GPT models by OpenAI [Radford et al.(2018)Radford, Narasimhan, Salimans, Sutskever et al.,OpenAI(2023)] 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 [Reimers and Gurevych(2019)] introduced the Sentence-BERT model, capable of generating high-quality sentence embeddings using only a fraction of the original BERT model's computational effort.
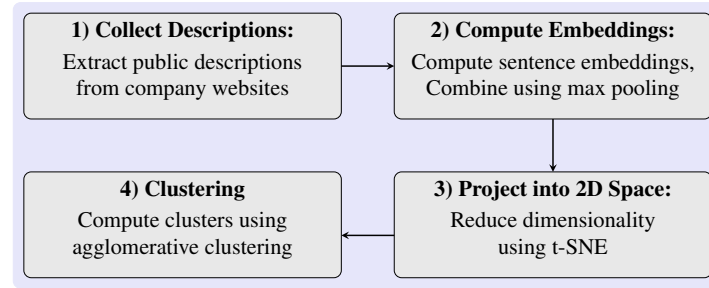
Procedure

Figure 1: Overview of our clustering pipeline.

Our goal is to collect data points on companies in the region of interest 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.

Figure 1 shows a graphical representation of our general clustering pipeline, consisting of 4 steps: 1) Collection of company descriptions, 2) computing embeddings, 3) reducing the embeddings' dimensionality, 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 an economic dataset would be to gather several attributes, like revenue, number of employers, and products, for many companies in the region of interest and then store each in a fixed-size vector. However, it is generally unclear which are the best attributes for representing companies. Further, encoding non-numerical attributes, like products and customer base, so that clustering algorithms can parse them is not straightforward. Lastly, some attributes may be unavailable for some companies, meaning we would need to discard them or we would need to insert dummy values, which could lead to unwanted effects. 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, since companies design their website to attract customers and investors, they include the most relevant information. Hence, we circumvent the problem of creating a targeted aggre-

gator for specific information, making our approach simpler to implement and significantly more scalable & automatable for larger target regions.

2) *Compute Embeddings:* To achieve the fixed-size inputs, we map each company description to an embedding that describes the data in a low-dimensional space. We rely on a sentence transformer model to compute embeddings for each sentence in a company's description. We then attain a single embedding for each description by taking the maximum over each feature dimension, which we denote as max pooling.

3) *Projection into 2D Space:* The embeddings' number of features might be too large relative to the size of the dataset, potentially invoking the curse of dimensionality [Köppen(2000)]. Further, clustering data in high-dimensional space prevents visualizing the results. Hence, we project the embeddings to 2D space. To do so, we apply the *t-distributed stochastic neighbor embedding* (t-SNE) technique, which 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 accurately preserve local structures in the data, making it suitable for identifying clusters.

4) *Clustering:* We cluster the data using agglomerative clustering, which initially assigns each data point to a separate cluster, then iteratively merges close clusters until all are in the same cluster. The intermediate clustering that achieves the largest separation between data points is chosen as the final clustering. Once the algorithm converges, we receive cluster labels for each company in our dataset.

Although our general pipeline represents companies using their descriptions, we emphasize that it can also utilize additional 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 max pooling.

Figure 2a 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. Further, users can see the location and clustering assignment of all companies in the given dataset on a map. When hovering the mouse over a company's data point, additional information about the company is shown.
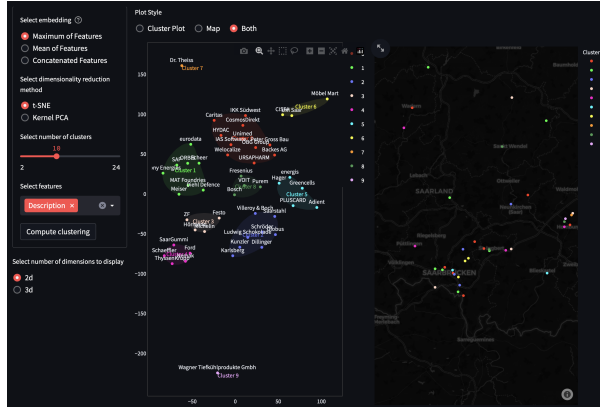
(a) Clustering analysis page of our user-interface.
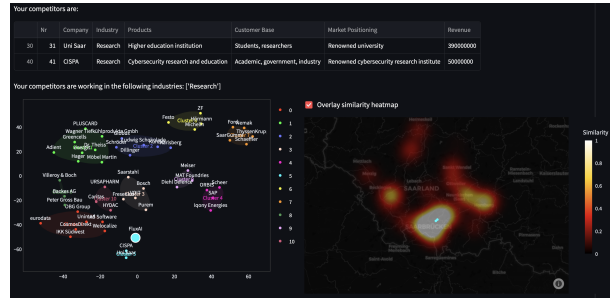
(b) Output of the location recommendation feature of our user-interface.

Fostering Business Clusters through Location Recommendations

To give location recommendations for companies seeking to establish new sites in the region of interest, we utilize parts of our clustering pipeline. Given the novel company's description, we compute an embedding and apply dimensionality reduction. We then assign the data point to a cluster and compute which company within the cluster is most similar to the novel company. Given the novel company's cluster label, the location recommendation is to establish a new site nearby the most similar company within the cluster. Thus, our location recommendations support the geographical concentration of companies acting in the same cluster, encouraging their growth.

Our user-interface offers a dedicated tab for location recommendations. Users can provide their company's name and description in text boxes. Additionally, they can specify their company's industry, products, customer base, market position, and revenue. After clicking "Submit", users are shown the output depicted in figure 2b. 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 that shows the similarity between their company and all other companies in the dataset.

Results

To build a dataset of company descriptions, we extracted the descriptions of 50 of the largest companies in Saarland, which are listed in the appendix 1.2. For the experimental evaluation, the data was collected manually. However, in the future, we aim to create an automated scraper to tackle larger regions. To test our location recommendation feature, we created company descriptions using GPT-4 [OpenAI(2023)]. Additionally, we evaluate the impact of providing additional features to the clustering pipeline, for which we also gathered the features industry, products, customer base, market positioning, and revenue of each company. We utilize the popular pre-trained sentence transformer model "all-MiniLM-L6-v2" [sen(2023)] to compute sentence embeddings for each description.
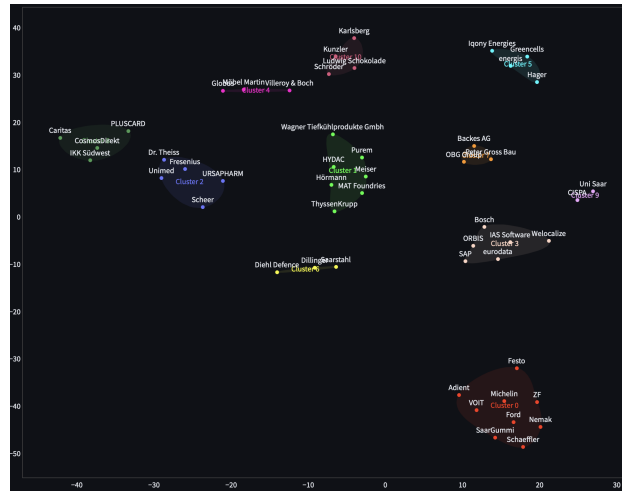

Figure 3: Clustering of company descriptions using t-SNE and Agglomerative clustering. The company descriptions and industries were used as features.

We will now investigate the results of our clustering and location recommendation system. Besides the company descriptions, we used the companies' industry as an additional feature since this

yielded the most concise clustering. Figure 3 displays the resulting clustered data points. To evaluate the clustering of our pipeline, we will compare the found clusters to those identified in [Trippl and Otto(2009)] (Table 4, p. 1226): *Coal and mining industries*: With the decline of coal mining in recent years, there are no more coal mines in Saarland, therefore no companies are present in our dataset. *Metal industry*: Covered by cluster 1 (light green) and cluster 6 (yellow), which contain steel producers like SaarStahl and Dillinger and companies offering steel products like Hörmann, a producer of various types of doors. *Automobile industry*: Saarland's large automobile market segment is covered by cluster 0 (red), including companies like Ford, ZF, and Michelin. *Energy*: Cluster 5 (light blue) comprises the energy provider energis and the energy service providers Iqony Energies and Greencells, representing the energy market segment. *Information and communication technology*: The IT products and IT consulting market segment, to which companies like SAP, ORBIS, and eurodata belong, is covered by cluster 3 (beige). *Biotechnology and nanotechnology*: Covered by cluster 2 (dark blue), with members such as Fresenius, which produces dialysis machines, and USRAPHAM, which manufactures medicine. One can see that the results of our pipeline agree with the general sectors identified by [Trippl and Otto(2009)]. However, our approach did lead to more fine-grained separation. Additionally, we identified sectors not covered in their work, such as cluster 7 (orange), which represents the construction work market segment, and cluster 10 (pink), which represents the food production market segment, both not as influential in the regions economy.

To test our location recommendation system, we generated descriptions for three 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. *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 leader in information security research, including artificial intelligence, and thus knowledge transfer between both companies could be highly

beneficial. 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. 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 many 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.

Conclusion

In this work, we demonstrated how language models and clustering algorithms could be combined to identify market segments and give location recommendations for novel companies such that business clusters are fostered.

We represented companies by their public descriptions, which allowed us to leverage state-of-the-art language models to transform the highly informative text data into embeddings. Given the embeddings, we applied a clustering algorithm to infer related companies, which enables us to make location recommendation. A qualitative evaluation of our methods has shown that they can reliably cluster similar companies, which allowed us to infer clusters in Saarland. Further, our location recommendation system demonstrated its capability of identifying similar companies, given the description of a new company. To allow end-users to leverage our methods, we implemented a user-friendly interface with many functionalities, such as visualizing the identified clusters and location recommendations.

A promising opportunity for future work would be fine-tuning our language models using text data from particular business sectors. To achieve this efficiently, an automated aggregation system could be developed, along with optimizations to the clustering pipeline.

References

[sen(2023)] 2023. "HuggingFace all-MiniLM-L6-v2," .

   URL `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[Devlin et al.(2018)Devlin, Chang, Lee, and Toutanova] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:181004805* .

[Gertler et al.(1995)Gertler, Oinas, Storper, and Scranton] Gertler, M. S., Oinas, P., Storper, M., and Scranton, P. 1995. "Discussion of "Regional Advantage: Culture and Competition in Silicon Valley and Route 128" By AnnaLee Saxenian," *Economic Geography* (71:2), pp. 199–207.

   URL `http://www.jstor.org/stable/144358`

[Hahn(2013)] Hahn, C. K. 2013. "The transboundary automotive region of Saar-Lor-Lux: Political fantasy or economic reality?" *Geoforum* (48), pp. 102–113.

   URL `https://www.sciencedirect.com/science/article/pii/S0016718513000948`

[Kamath et al.(2012)Kamath, Agrawal, and Chase] Kamath, S., Agrawal, J., and Chase, K. 2012. "Explaining Geographic Cluster Success—The GEMS Model," *American Journal of Economics and Sociology* (71:1), pp. 184–214.

   URL `http://www.jstor.org/stable/23245182`

[Köppen(2000)] Köppen, M. 2000. "The curse of dimensionality," in *5th online world conference on soft computing in industrial applications (WSC5)*, vol. 1, , vol. 1.

[Kuah(2002)] Kuah, A. 2002. "Cluster Theory and Practice: Advantages for the Small Business Locating in a Vibrant Cluster," *Journal of Research in Marketing and Entrepreneurship* (4), pp. 206–228.

[Mikolov et al.(2013)Mikolov, Chen, Corrado, and Dean] Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:13013781* .

[OpenAI(2023)] OpenAI 2023. "Introducing ChatGPT," .

[Porter(1998)] Porter, M. 1998. "Clusters and the new economics of competition," *Harvard business review* (76:6), p. 77—90.
URL `http://europepmc.org/abstract/MED/10187248`

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

[Reimers and Gurevych(2019)] Reimers, N., and Gurevych, I. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:190810084* .

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

[Trippl and Otto(2009)] Trippl, M., and Otto, A. 2009. "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* (41), pp. 1217–1233.

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

1   Appendix

1.1   Link to Code

`https://github.com/nicola-mueller/LM-based_Company_Clustering`

1.2   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