

Detecting Business Clusters using Language Models

Research-in-Progress Paper

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 since businesses must be located close to their customers and suppliers for more effective marketing or reduced transportation costs. Further, being located in an active regional cluster can significantly benefit business growth (Porter, 1998), (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. Thus, establishing or further developing business clusters is of high interest to companies and local governments. However, identifying business clusters 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) 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 using public text data. 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. Further, the interface supports making location recommendations for new company sites using text inputs. In summary, our main contributions are: 1) Circumventing the challenges of

economic data collection by relying on public text data, 2) Leveraging language models to apply clustering algorithms to text data, 3) Demonstrating our approach's feasibility using Saarland's economy and providing a user-friendly graphical interface.

Related Work

Business clusters and their advantages have been studied in many works. (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, (Trippel & Otto, 2009) 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.

In natural language processing (NLP), translating text data to machine-processable representations is a fundamental challenge for numerous tasks, due to the complexity and variability of natural language. A popular approach for representing text data are 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. (Mikolov, Chen, Corrado, & Dean, 2013) popularized learning word embeddings from raw text data by introducing their Word2Vec model. More recently, sentence transformers extended these

approaches by computing embeddings for entire sentences to capture more contextual information than individual word embeddings. (Reimers & Gurevych, 2019) introduced the Sentence-BERT model, capable of generating high-quality sentence embeddings using only little computational effort.

Procedure

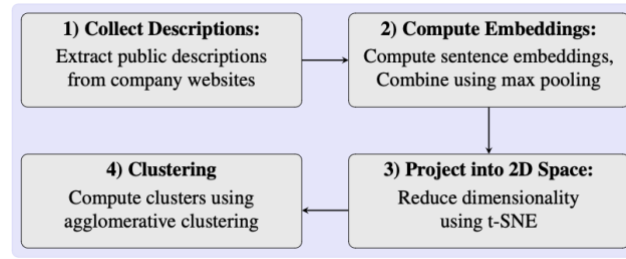


Figure 1. Overview of clustering pipeline

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. Figure 1 shows a graphical representation of the 4 steps of our general clustering pipeline, which we will examine in the following:

1) Collect Descriptions: The traditional approach for building an economic dataset would be to gather attributes, like revenue, number of employers, and products, for many companies in the region of interest and then store each in a vector. However, it is generally unclear which are the best attributes for representing companies. Further, encoding non-numerical attributes 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 insert dummy values, which could lead to unwanted effects. To address these challenges, we propose representing companies by the description on their websites instead of numerical attributes. Given that

companies design their website to attract customers and investors, these descriptions include the most relevant information. Hence, we circumvent the problem of feature selection, making our approach simple and scalable.

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 dataset's size, potentially invoking the curse of dimensionality. Further, clustering data in high-dimensional space prevents visualizing the results. Hence, we need to project the embeddings into 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 KL 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.

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 2 depicts the clustering analysis tab of our user-interface. On the left side, users can modify the hyperparameters of the pipeline. 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.

To give location recommendations for novel companies, we utilize parts of our clustering pipeline. Given the 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 related companies, fostering business cluster development. 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 2. 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.

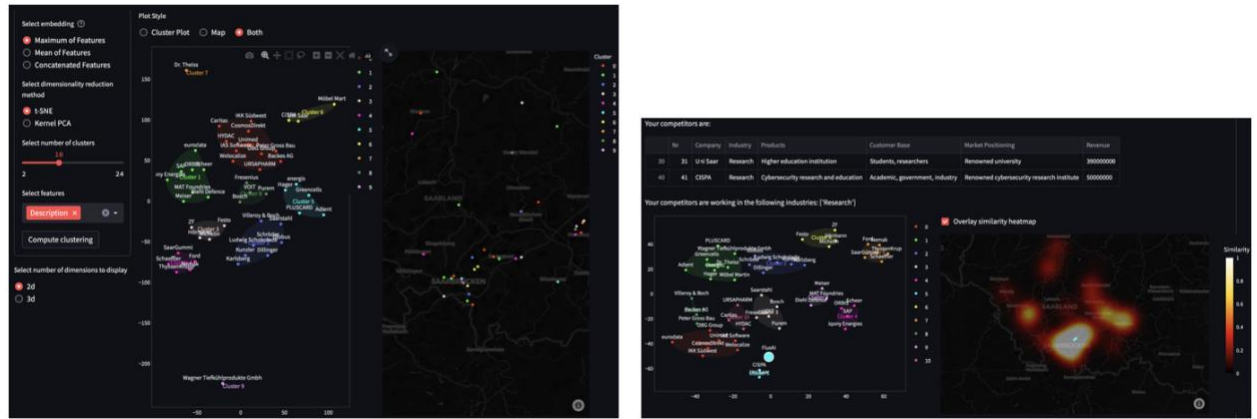


Figure 2. Clustering analysis & location recommendation in our user-interface

Results

To build a dataset of company descriptions, we manually extracted the descriptions of 50 of the largest companies in Saarland. 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 ChatGPT. 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" from the HuggingFace hub to compute sentence embeddings for each description.

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. We will compare the found clusters to those identified in (Tripl & Otto, 2009):

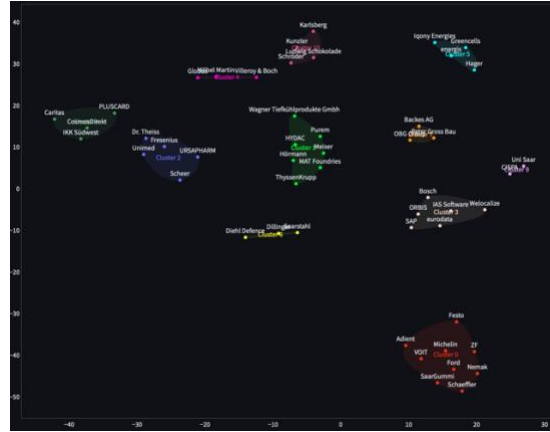


Figure 3. Clustering of companies in Saarland.

Coal and mining industries: There are no more coal mines in Saarland, therefore no such 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.

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 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 medical companies such as Fresenius, and USRAPHAM.

One can see that the results of our pipeline agree with the general sectors identified by previous work. 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.

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, 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. 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 business clusters 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 language models to transform the highly informative text data into embeddings. Given the embeddings, we applied a clustering algorithm to infer related companies, which enabled us to make location recommendations. A qualitative evaluation of our methods has shown that they can reliably cluster similar companies 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 recommending locations.

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.

Bibliography

- Porter, M. (1998). Clusters and the new economics of competition. *Harvard business review*, 77-90.
- Kamath, S., Agrawal, J., & Chase, K. (2012). Explaining Geographic Cluster Success—The GEMS Model. *American Journal of Economics and Sociology*, 184-214.
- Gertler, M. S., Oinas, P., Storper, M., & Scranton, P. (1995). Discussion of "Regional Advantage: Culture and Competition in Silicon Valley and Route 128" By AnnaLee Saxenian. *Economic Geograph*, 199-207.
- Kuah, A. (2002). Cluster Theory and Practice: Advantages for the Small Business Locating in a Vibrant Cluster. *Journal of Research in Marketing and Entrepreneurship*, 206-228.
- Hahn, C. K. (2013). The transboundary automotive region of Saar-Lor-Lux: Political fantasy or economic reality? *Geoforum*, 102-113.
- Trippel, M., & 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*, 1217-1233.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv*.