




Introducing PAUSE: A deep natural language model for the private markets

[Sonja Horn](#), [Lele Cao](#) | EQT Ventures x Motherbrain
November, 2021

This is the extension of an [EQT Ventures medium](#) post on PAUSE with the same title, written by Sonja Horn and Lele Cao.

Introducing PAUSE

To train a high performing deep learning model for competitor mapping, the data scientists in EQT's Motherbrain team had to finetune the pretrained sentence embedding models using more relevant supervision signals. These signals are basically textual descriptions of company pairs with binary annotations (labels) indicating their relatedness. 1 means the statements are related, 0 means they are not.

Company A	Company B	Similar?
<i>Owner and operator of data centers in UK ... distribute data in data centers and the global digital economy.</i>	<i>Independent co-location / data center provider in Slovenia.</i>	 YES 1
<i>Provider of human-computer interaction technology designed to ... documentation is common to both CX and ES.</i>	<i>Developer of a mobile stock-trading application designed to make ... enabling traders to discover and invest in markets without a hassle.</i>	 NO 0
<i>Provide a reliable and fast veterinary diagnostic service ... We intend to be your partner in the daily medical diagnosis!</i>	<i>Provider of laboratory services. We care about ... Biochemical and haematological examinations are available 24 hours a day.</i>	 YES 1

166,832 pairs of company descriptions ...

Table 1: Illustration of the first three samples (shortened and anonymized) of our dataset used to train company embeddings for identifying similar companies.

In EQT's database, there are currently more than eight million companies with textual descriptions and only 12,326 of these have annotations indicating their relatedness to other companies. These form 166,832 company pairs with noisy binary labels indicating whether they are similar (55,139 pairs) or dissimilar (111,693 pairs).

Table 1 demonstrates a few annotated sentence pairs from our finetuning dataset. The unbalance between annotated and unannotated sentence pairs exemplifies what is true for most private capital datasets: the supervision signal is strictly limited. Yet, deep learning only performs well if exposed to a large number of annotations.

The team asked themselves: is there a way to apply deep learning in a way that allows Motherbrain to leverage an entire company dataset, not just annotated descriptors? After some careful literature survey, brainstorming, and experimentation, we proposed a novel

method -- PAUSE - Positive and Annealed Unlabeled Sentence Embedding - which generates numerical representations from company descriptions, enabling a measurement of closeness between any two companies. It uses an entire dataset, not just annotated descriptors, meaning the model needs far fewer supervision signals to undertake the competitor mapping task successfully. In practise, the proposed method requires only about 5 - 10% of the annotated descriptors compared to other approaches. Despite the lack of supervision signal, PAUSE archives (and sometimes surpasses) state-of-the-art results on many benchmarking tasks in natural language processing.

The highlights of PAUSE include:

- **Low application barrier:** reduces the required labeling effort to about 10%.
- **Easy to train:** the model can be trained in an end-to-end fashion.
- **Wide application range:** can be applied to many underlying neural network architectures.
- **Stable and effective:** good model performance is largely guaranteed with low volatility.

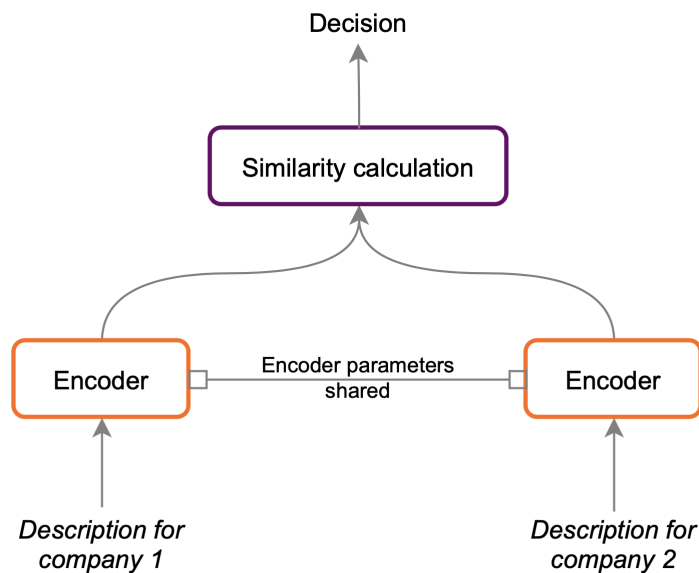


Figure 1: The dual-encoder model comprising two identical encoders.

A simplified diagram of PAUSE can be seen in Figure 1. First off, company descriptions are passed through separate encoders with shared parameters. The encoder outputs are then used in a similarity calculation which determines if the companies should be considered competitors or not. We call the entire architecture a *dual-encoder* that outputs a similarity score given a pair of textual company descriptions. As exemplified in Figure 2, the trained dual-encoder is deployed on our platform to automatically generate similar companies for any company in scope.

The encoder in Figure 1 is capable of generating a 64 dimensional vector (i.e. embedding) for any input company, which may be useful in many analytics use cases in private capital. Therefore, we also deploy the trained encoder alone on the backend of our platform as a streaming prediction job which listens to the change of company descriptions (collected

through many data sources). Whenever there is a change of description for any company, the job immediately produces a new company embedding by triggering an encoder prediction request. The company embeddings are then stored in our data warehouse.

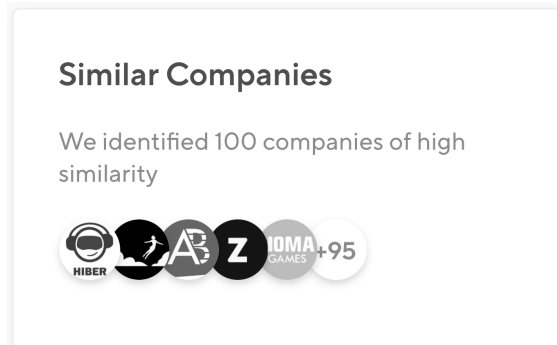


Figure 2. A screenshot of similar company recommendations on the Motherbrain platform

The encoder in Figure 1 is capable of generating a 64 dimensional vector (i.e. embedding) for any input company, which may be useful in many analytics use cases in private capital. Therefore, we also deploy the trained encoder alone on the backend of our platform as a streaming prediction job which listens to the change of company descriptions (collected through many data sources). Whenever there is a change of description for any company, the job immediately produces a new company embedding by triggering an encoder prediction request. The company embeddings are then stored in our data warehouse.

To demonstrate how the trained encoder can support various types of relational analytics tasks, we can not use the embeddings in their current form (64 dimensions). Instead, we reduce the dimensionality of the embeddings from 64 to merely 2 using the dimensionality reduction method PCA for a set of companies in our data warehouse. The dimensionality reduction allows us to plot the companies in a 2 dimensional plane, where the resulting closeness of companies can be considered a bad-yet-acceptable approximation of their location in the original 64 dimensional embedding space, since applying PCA will cause information loss.

We visualize the 2-dimensional embeddings in a scatter plot, as shown in Figure 3, where a subset of companies are visualized for simplicity. Despite the enormous dimensionality reduction, and without overlaying any clustering algorithm, we can qualitatively identify several clusters (represented by ellipses in different colors in Figure 3) representing food delivery, food products, sustainability, and fintech. It is also obvious that the adjacent clusters tend to share some commonalities annotated by the overlapped areas between two clusters, such as sustainable fintech and sustainable food. You might also have noticed a company with a question mark in the center that does not seem to belong to any cluster. Its purposefully engineered description might give you an idea of its neutral location:

Dummy Company: *“Provider of loan, cleaning supplies, groceries, snacks, drinks & electronics, in a sustainable and environmentally friendly way.”*

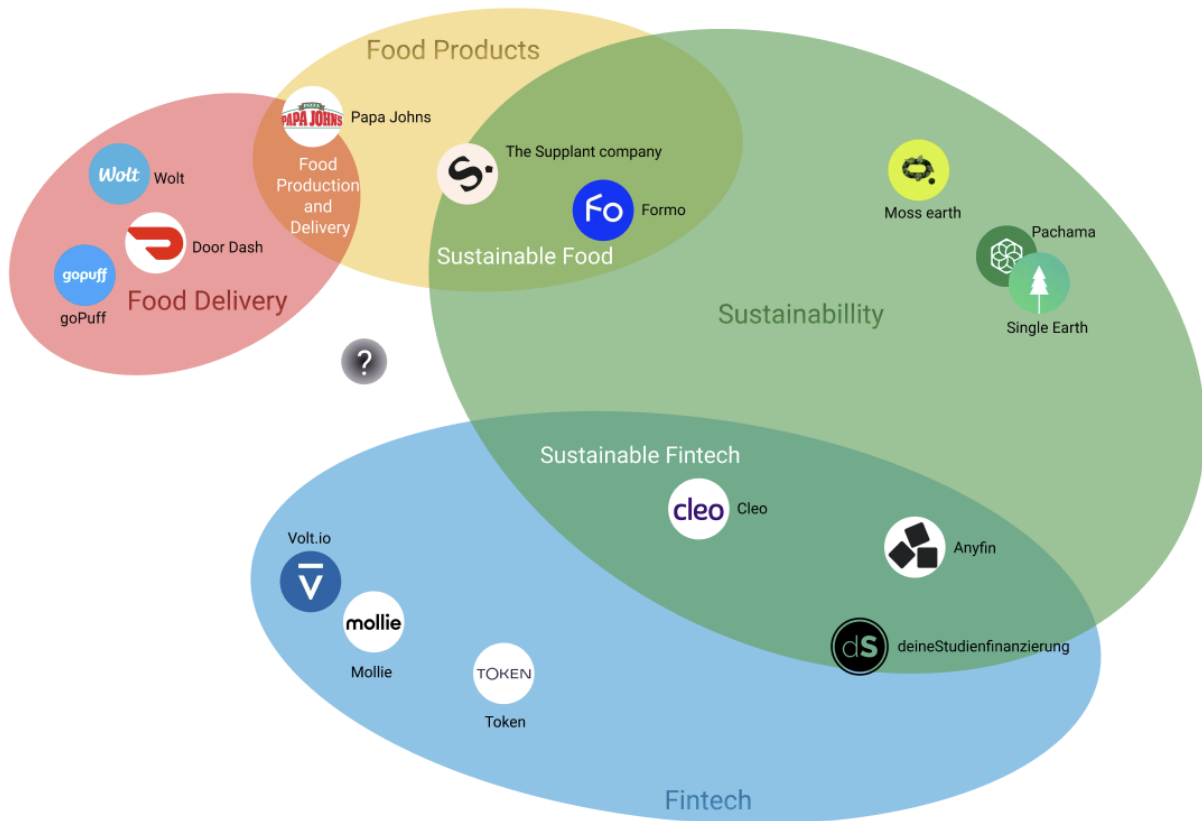


Figure 3. A demonstration of company location in relation to their embeddings produced by the trained encoder (approximated with PCA in a 2D plane).

We are pleased to announce that this work has been accepted as a long paper by the [EMNLP 2021](#) main conference, which is acknowledged as one of the best NLP (Natural Language Processing) academic conferences in the world. As of 2021, according to [Microsoft Academic](#), EMNLP is the 14th most cited conference in computer science, with a citation count of 332,738, between ICML (#13) and ICLR (#15). For details of the PAUSE algorithm, we recommend reading [our paper](#) and trying [the source code](#).

Besides competitor mapping, obtaining good embeddings of company descriptions is beneficial to many other use cases in the domain of private capital, such as sector tagging and trend detection. Moreover, the core of the PAUSE algorithm lies in applying a new optimization function for training deep learning models with a much lower requirement on manual annotation than what has previously been seen in this domain. As a result, it could also be used to address tasks like deal sourcing and success prediction.

Private capital is undergoing a massive transformation, shifting from networking heavy to neural network heavy. We can expect to see more efforts on adopting cutting-edge models trained with more data from different sources, in different modality, and with lower requirement of annotation. PAUSE and EQT's Motherbrain team contributes to this development, and will continue to move the needle when it comes to applying advanced technology in the private capital sector.