Company Recommendation in Private Equity: A Similarity-Based Approach for the Healthcare Sector

Anna Abrahamyan

Archimed*

Supervisor: Ankit Tewari

June 24, 2025

Abstract

In the fast-evolving and dynamic field of healthcare investments, private equity firms face increasing challenges in identifying high-potential opportunities. This thesis presents a scalable, data-driven recommendation system designed to streamline the sourcing process by identifying companies similar to what current company the investor is interested in. Our system suggests companies by measuring company similarity based on textual descriptions and some additional features. Using advanced Natural Language Processing (NLP) techniques, the system computes company similarity using sentence embeddings derived from BERT and OpenAI embedding models then combined with features. Two modeling strategies are compared and evaluated: (1) representing all company features as a unified textual input and (2) treating each feature type distinctly for hybrid similarity computation. Similarity is computed using cosine similarity across all approaches, and ranking algorithms are applied to refine recommendations. The system integrates domain-specific nuances, supporting both data-driven analysis and the qualitative judgment of investment professionals. The resulting framework offers a robust tool for enhancing decision-making efficiency in private equity for healthcare, with potential for broader application in identifying investment opportunities.

Contents

*ARCHIMED is a global investment firm focused on driving the sustainable development of healthcare industries./

1 Introduction

Private equity (PE) firms play a significant role in the global financial ecosystem by investing in privately held companies, with the goal of generating substantial returns over a medium to long term horizon, typically spanning four to ten years [3]. In simpler terms, private equity refers to investments made in companies that are not listed on public stock exchanges, even though investors may occasionally be involved in the privatisation of publicly listed company.[7] These investments are often managed by specialized intermediaries and involve the acquisition of significant ownership stakes in target companies. This hands-on investment approach is coupled with strategic guidance, operational oversight, and long-term support, all of which are intended to drive growth and increase company value over time.[3]

Traditionally, the decision to invest in a company relies on manual review of financial and operational indicators, often supplemented by the intuitive judgment or "gut feeling" of experienced investors.[7] However, as company volume and data complexity increase this manual, intuition-driven approach is becoming increasingly difficult to scale and time consuming. It becomes increasingly difficult for human analysts to manually evaluate every potential target in a consistent and scalable way. In response to these growing challenges, private equity firms are beginning to adopt artificial intelligence (AI) and machine learning (ML) technologies to streamline various stages of the investment lifecycle, including deal sourcing, portfolio monitoring, and comparative market analysis. These tools can help analysts parse large volumes of data, detect hidden patterns to suggest similar companies and support more data-driven decisions. However, building such systems for private equity presents unique challenges. Unlike consumer platforms that rely on large volumes of user-item interaction data (e.g., movie ratings or purchase history), PE firms operate in data-sparse environments, with limited publicly available financials, proprietary deal flow, and minimal historical interaction data. Financial data for many potential targets, especially in the healthcare sector, is either private, incomplete, or inconsistently reported. Additionally, deal flow information is often proprietary, and PE firms rarely have access to comprehensive records of previous investment evaluations.

Building an advanced recommendation system in private equity (PE) presents several challenges. One of the most significant is the sparsity and scarcity of the historical deal data and financial data from small companies has been considered as one of the major factors hampering advancements of data-driven methods in the PE market. Next, the degree of complexity involved in applying computational methods in PE is exacerbated by the qualitative and intuitive nature of investment decision making. PE managers often report their decisions being based more on their 'gut feeling' than hard-coded rules and data. The holding period for PE investment is relatively long, spanning a few to many years. Therefore, the estimation for any models using the exit outcome (for example, return on investment specific to the type of exit) may be less relevant as the market environment changes at a rapid pace. As a result, the model is less meaningful as they may not accurately address the current investment environment. Finally, data imbalance has been repeatedly reported as an issue by much of the academic literature across different types of studies related to PE, as the number of no-deal companies significantly exceeds the number of deal companies. [7]

Despite these obstacles, there is growing interest in leveraging machine learning and recommendation systems to support investment decisions in PE. Among the various approaches, similarity-based recommendation systems stand out for their potential to iden-

tify comparable companies based on company and sector characteristics or performance metrics. However, their application in the private equity domain presents technical and domain-specific challenges. First, defining similarity itself is non-trivial. Companies may appear similar based on industry classification or product offerings, yet differ drastically in financial stability, growth potential, or strategic fit. Furthermore, sparse and heterogeneous data complicate the construction of reliable feature representations; textual data from company descriptions may be verbose or ambiguous, while structured financial indicators may be incomplete or outdated. Additionally, there is a risk of the model overemphasizing superficial similarities and failing to recommend innovative or contrarian opportunities. Ensuring the interpretability of similarity scores is another key concern, as PE professionals require transparent justifications to validate recommendations against due diligence standards. Therefore, deploying similarity-based systems in this context requires careful feature engineering, domain-aware embedding strategies, and robust evaluation to ensure meaningful and actionable outputs.

To address these challenges, this thesis proposes a content-based recommendation framework for automated, data-driven screening of investment opportunities tailored to private equity contexts. The system integrates structured and unstructured data from multiple sources to assess the financial health and management quality of companies and uses LLMs to generate transparent investment recommendations and applying similarity-based ranking methods. Feature engineering, model evaluation strategies, and class imbalance handling are explored to ensure the robustness and transparency of the recommendations. The core contributions of this work are as follows.

- We define a systematic, data-driven problem formulation for deal screening in PE under real-world constraints.
- We introduce a hybrid recommendation architecture combining content embeddings (via Sentence-BERT), financial and managerial features, and explainability components powered by large language models (LLMs).
- We evaluate techniques to manage data sparsity, class imbalance, and domain-specific requirements, including compliance and transparency.

The remainder of this paper is structured as follows: Section 2 covers the context of the project. Section 3 reviews related literature on recommendation systems in finance, private equity sectors, ranking models. Section 4 describes the dataset construction process, including the integration of textual and financial data for healthcare companies. Section 5 outlines our methodology, including the model architecture, feature engineering strategies, and evaluation metrics. Section 6 presents experimental results, model comparisons, and ablation studies. Section 7 concludes the paper with a discussion of limitations and directions for future research.

2 Context of the Project

This project is conducted within the framework of ARCHIMED private equity. ARCHIMED is a global private equity firm specializing exclusively in the healthcare sector. Founded with a mission to support and scale innovative healthcare businesses, the firm operates across Europe and North America, managing several billion euros in assets under management. ARCHIMED's investment strategy focuses on small to mid-sized companies with

high growth potential, spanning sectors such as medical devices, consumer products, diagnostics, healthcare IT, life sciences, etc. [1] The motivation for this project arises from the objective of streamlining and partially automating the investment evaluation process to enhance efficiency and support decision-making within the investment team.

Currently, identifying promising investment opportunities requires significant manual effort, including browsing trade fairs, conducting extensive web searches, and consulting AI tools to discover companies similar to those already of interest. Interviews conducted with members of the investment team have confirmed that this process is time-consuming and inconsistent, often lacking a systematic approach to identifying relevant companies. The core objective of this project is to develop a system that automatically recommends and ranks potentially relevant healthcare companies for investment by analyzing unstructured textual data, primarily company descriptions using similarity-based methods. The system aims to reduce manual search efforts and enhance the efficiency, scalability, and consistency of the early-stage deal sourcing process.

A fundamental difficulty in automating deal sourcing lies in the concept of similarity. Defining and measuring similarity between companies is inherently complex. This concept is inherently complex, as companies that appear similar based on basic attributes, such as industry, country of operation, employee count, or company age, may still differ significantly in terms of their strategic focus, innovation, or market positioning. While this project does not incorporate deep financial or managerial data, it leverages unstructured textual descriptions alongside structured metadata to approximate strategic and operational similarity. The goal is to move beyond simplistic matching and instead capture meaningful patterns that align with the investment team's interests. We believe this model will support investment professionals, thereby freeing up their time for higher-order analytical tasks and ultimately improving the quality of investment opportunities.

We define the automated deal sourcing task in private equity as a **constrained similarity-based ranking problem** over a candidate set of companies $\mathcal{C} = \{c_1, c_2, ..., c_n\}$, where each company c_i is represented by a high-dimensional feature vector $\mathbf{x}_i \in R^d$, combining structured financial indicators, textual embeddings, and categorical metadata.

3 Survey of the State of the Art

This section reviews recent advancements in recommender systems and company similarity modeling, particularly as applied in private equity and venture capital contexts. It highlights systems that integrate NLP, structured data, and ranking algorithms to support investment decision-making.

3.1 Recommender Systems in Private Equity Firms

In recent years, PE firms are increasingly adopting data-driven tools to streamline their sourcing and investment processes. One emerging class of tools is recommender systems, which assist deal teams in identifying attractive targets by surfacing companies that align with their investment thesis. These systems aim to replicate or enhance the traditional process of identifying lookalike companies based on textual information, proprietary knowledge and network relationships. A prominent example is EQT's platform Motherbrain (launched 2016), which ingests vast multi-source data on companies and uses machine learning to flag potential targets. According to public reports, this AI

system directly identified 7 of the 50 new investments made by EQT Ventures to date, effectively guiding analysts to "hidden gems" early.[6] Complementing proprietary systems like Motherbrain, academic research has proposed scalable frameworks using explainable AI. Petersone et al. (2022) propose a data-driven, explainable AI framework for screening PE opportunities, validating across multiple algorithms and handling class imbalance to mimic human screening efficiency.[7] Luef J. et al. from TU Wien have worked on early-stage enterprise recommenders, that define investor requirements via interviews and formalized five system requirements—leading to a multi-algorithm hybrid solution.[5]

Extending beyond traditional recommender systems, graph-based representations have emerged to model inter-company relationships at scale. Similarly, academic and industry efforts like CompanyKG illustrate this approach: it is a large-scale knowledge graph of companies (built from EQT's platform data) containing 1.17 million firms and 15 types of inter-company relations.[2] CompanyKG was explicitly created to facilitate recommending companies similar to a given query, essentially forming the backbone of a recommender for investment professionals.

Complementing graph approaches, recent advances in language models like BERT and OpenAI have enabled deeper semantic comparisons of textual entities about companies. In the context of private equity, semantic similarity can be used to compare company descriptions, identify comparable firms, and detect strategic alignment. Building on this approach, other researchers and firms have developed semantic similarity tools. For example, Axel Springer's tech team built a "scouting" tool that uses a pre-trained BERT model to find startups similar to a few example companies. Research confirms that LLM-derived embeddings capture meaningful company relationships.[8] Vamvourellis et al. showed that embeddings learned from SEC 10-K descriptions could reproduce industry classifications and cluster firms with similar financial profiles. [9]

3.2 Ranking

Once similarity scores are computed, candidates must be ordered for recommendation. Cosine scores themselves yield a ranking, but systems often apply additional filtering or learning-to-rank. CompanyKG defines a "Similarity Ranking" task (given a query company, decide which of two candidates is more similar) to mimic M & A target prioritization. [2] Excelling at such ranking is critical for good recommendations. In industry, top lists returned by similarity search may be re-ranked using domain rules or a trained model (e.g. weighting by investment criteria or domain relevance).

Text similarity can be combined with structured features (financials, sector codes, etc.). A practical design is to embed free-text separately and concatenate it with numeric/categorical vectors. For example, Sonja Horn in a venture-capital study encoded each company description via a 50-dim pre-trained text model and then concatenated those embeddings with scaled numeric features in a neural predictor. [4] This contrasts with flattening all fields into one text blob.

In summary, leading PE firms increasingly use recommender systems, blending NLP, graph analysis, and business heuristics to suggest similar or interesting companies. Current content-based recommendation methods use advanced contextual embeddings (including OpenAI's text-embedding models and BERT's sentence transformers) as the foundation for semantic similarity search. While existing systems such as CompanyKG and Motherbrain have demonstrated success in general-purpose or technology-focused company recommendation, their methods are not optimized for the unique characteristics of

the healthcare sector. This thesis builds on these foundations to develop a domain-specific recommendation system, where the domain refers to healthcare companies—including those in medical devices, diagnostics, healthcare IT, life sciences, and related subfields. Healthcare companies often involve highly specialized terminology, regulatory considerations, and distinct business models, which differ significantly from other industries. As a result, traditional similarity metrics or keyword-based approaches may fail to capture relevant nuances. This project addresses these domain-specific challenges by incorporating healthcare-relevant features, language models, and data representations tailored to the sector.

Table 1: Summary of Related Recommender Systems

Work/System	**Data Types**	**Embedding Model**	**Domain**
Motherbrain	Multi-source	Unknown	General PE
CompanyKG	Graph + Text	KG embeddings	Tech/General
Axel Springer Tool	Text	BERT	Media/Startups
ARCHIMED	Text + features	BERT/OpenAI	Healthcare PE

4 Datasets

One of the main challenges of this project was acquiring an accurate and comprehensive dataset suitable for building a recommendation system. The data collection process was time-consuming and consisted of many steps, due to the absence of a centralized one source. The final dataset used in this project is a combination of multiple sources: Publicly available CSV files on GitHub, primarily derived from Crunchbase, Tracxn, and PitchBook datasets, and additional data retrieved from LinkedIn profiles of the companies.

We began with a dataset consisting of 127,409 rows. We implemented cleaning tasks, removed duplicated rows, handled missing information. We then applied several data processing steps. We filtered out entries with missing company names or descriptions, discarded companies without LinkedIn information.

After thorough analysis, we decided for better representation of the company, we need to combine website description with LinkedIn description. After combining, to ensure data quality, we retained only companies with descriptions of at least five words to eliminate low-quality or incomplete entries, and removed entries with invalid founding years and unrealistic employee counts. The final dataset contained around 103,429 entries.

As we were working on with a limited GPU using the whole dataset of 103,429 entires was resulting in crashing of the notebook, hence we took a sample of 30,000 companies and worked with this dataset for our projects.

4.1 Exploratory Data Analysis

In this section, we present some interesting insights derived from our dataset using exploratory data analysis (EDA). The following plots provide visualizations that offer a deeper understanding of the data characteristics and distributions.

Table 2, gives us an idea about descriptions. After applying a filtering process, we ensure that the descriptions do indeed contain more than five words. This threshold was

	Number
Minimum words	5
Maximum words	520
Average words	32

Table 2: Statistics on Descriptions

chosen to promote coherence and readability, as shorter entries (typically 2–4 words) tend to be phrases and often lack meaningful context.

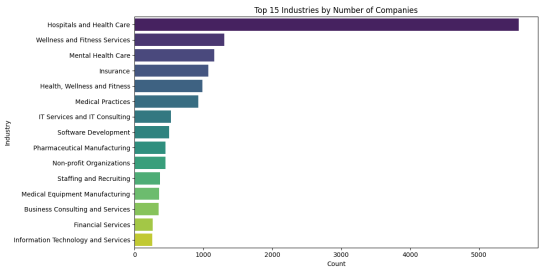

Figure 1: Companies by Industry

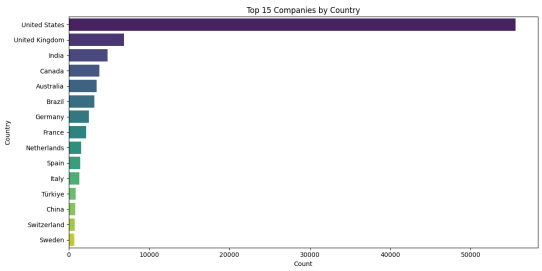

Figure 2: Companies by Country

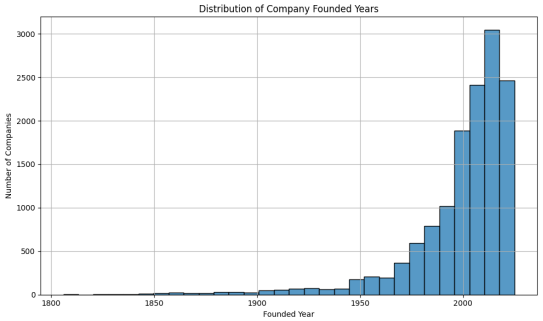

Figure 3: Companies by Founded Year

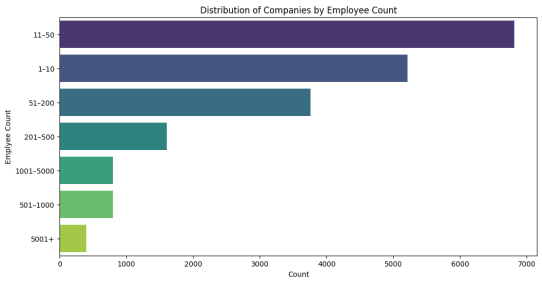

Figure 4: Companies by Employee Count

Figure 1-4 provide insights into the characteristics of our dataset. From Figure 1, we can see that the top 3 industries we have are Hospitals and Health Care, Wellness and Fitness Services and Mental Health Care, which align perfectly with ARCHIMED's domain focus. Next, Figures 2 illustrate the distribution of companies categorized by different countries, and you can see we have visible class imbalance. The class imbalance won't significantly impact our similarity model because the model focuses on measuring textual similarity between company descriptions rather than learning to classify labels. As a result, it is less sensitive to unequal representation across some features in the dataset.

Figure 3, presents the distribution of companies categorized by founded year. We used this information to get the company age and use it as a feature for our model. Figure 4 displays the distribution of companies categorized by employee count. We can see that the majority are mid-sized companies, ranging from 1-200 employees.

To gain a deeper understanding of the textual data, we analyzed the company descriptions using TF-IDF (Term Frequency–Inverse Document Frequency). Specifically,

we extracted and ranked the most common trigrams. Table 3 presents the top trigrams identified, offering insight into the prevalent themes and terminology within the descriptions.

Trigram	TF-IDF
Health care services	2459
Healthcare center offers	1536
Health care company	1425
Health care center	1326
Mental health services	1195
Home health care	1115
Mental health care	1041
Healthcare center provides	1006
Hospital health care	978
Provides mental health	877

Table 3: Statistics on Descriptions

Based on our exploratory analysis, we identified and selected the following categories of features for use in our model:

- **Textual Features**
  - Company descriptions (used for semantic similarity modeling)
- **Categorical Features**
  - Industry sector
  - Specialties
  - Country
- **Numerical Features**
  - Number of employees
  - Company age (calculated from founding year)

Despite extensive data extraction and cleaning, the dataset still contains some limitations, such as missing financial data: revenue, funding rounds, or valuations, which limits financial-based comparisons. Additionally, lack of labeled training data or historical interaction data, restricts the use of supervised learning.

4.2 Data Preprocessing

We explored two main preprocessing approaches:

- **Approach 1: Unified Text Representation**
  In this approach, all relevant company attributes (e.g., description, industry, country, etc.) are concatenated into a single textual string and encoded using NLP-based embeddings such as Sentence-BERT or OpenAI's embedding models. Missing values are replaced with `unknown` to preserve input consistency. An example format is:

```
‘‘Description: ..., Country: ..., Industry: ..., Specialties: ...,
Company Age: ..., Number of Employees: ...’’
```

- **Approach 2: Feature-wise Encoding**
  In this more structured approach, each feature type is processed independently using the most suitable encoding technique:
  - *Company descriptions* are embedded using language models.
  - *Specialties* are treated as a multi-label field and encoded accordingly.
  - *Industry* and *Country* are encoded using OneHotEncoder, with mechanisms in place to handle previously unseen categories.
  - *Number of employees* is log-transformed to reduce skewness and then scaled using `StandardScaler`.
  - *Company age* is also standardized using `StandardScaler` to ensure comparability across features.

5 Methodology

In this section, we outline the model architecture and methodology used for the comparison of our two approaches for suggesting similar companies. To represent companies based on their textual descriptions and business attributes, we needed models capable of generating high-quality semantic embeddings that reflect nuanced business and sector-specific content.

Given the nature of our dataset and workflow, we adopted a content-based similarity approach rather than supervised classification or collaborative filtering. In private equity pipelines, companies typically move through loosely defined stages (e.g., Radar, HOT, WIP, Closed). However, these transitions are nonlinear, sparsely documented, and often lack user interaction logs or decision labels. Out of 30,000 entries only around 3,000 companies had user interaction and final outcome. Out of which only 80 reached mid-stage consideration and just 11 advanced to the final stage. Consequently, the absence of high-quality ground truth data rendered traditional supervised or collaborative filtering methods infeasible. Instead, we opted to use external open-source datasets with accurate metadata and focused on a content-based similarity approach. Furthermore, this allows future collection of user preferences or stage-specific labels for later fine-tuning.

To encode company profiles into dense vector representations, we experimented with several pre-trained models, guided by benchmarks such as the Massive Text Embedding Benchmark (MTEB) leaderboard. We initially tested `multilingual-e5-large-instruct`, a strong multilingual general-purpose embedding model. However, its performance on our healthcare-focused corpus was suboptimal, likely due to domain mismatch. We then transitioned to `Lajavaness/bilingual-embedding-large`, a model fine-tuned for semantic similarity in English and French, which demonstrated improved performance in capturing relevant business relationships. Finally, we evaluated OpenAI's `text-embedding-3-small`.

BERT-based models, such as Sentence-transformers, use a bidirectional encoder trained via masked language modeling and next-sentence prediction. This enables them to understand local syntactic structures effectively. For example, in the sentence *"The patient was discharged from the general ward and referred to general medicine"*, BERT

distinguishes between "general" as a department and "general" as an adjective, depending on its context. **OpenAI embeddings**, such as `text-embedding-ada-002` or `text-embedding-3-small`, are based on autoregressive GPT-style transformers. They excel at capturing global semantic themes across longer documents. For example, given the input *"A biotech company developing gene-editing therapies for rare diseases"*, OpenAI embeddings highlight broader concepts like "biotech," "innovation," and "rare diseases," making them particularly effective for identifying high-level topical similarity across company profiles.

To enable scalable and efficient retrieval of similar companies, we implemented Facebook AI Similarity Search (FAISS). FAISS was used to build an approximate nearest neighbor (ANN) index from the generated embeddings, enabling efficient retrieval of the most similar companies for a given query vector. All embedding vectors were L2-normalized to approximate cosine similarity during search, allowing FAISS to return top-k nearest companies based on semantic proximity. Embeddings were indexed using an optimized vector store, allowing rapid similarity computations even with tens of thousands of entries. After constructing the FAISS index from either text-only or combined features, we queried for the top-N similar vectors per company, while excluding self-matches. The resulting nearest neighbors were directly associated with company names and similarity scores, providing a compact and interpretable output.

To improve ranking precision, we introduced a lightweight re-ranking stage. After initial retrieval using FAISS, we computed exact cosine similarity scores between the query and top companies. Additionally, we incorporated feature-level weighting into the re-ranking process—assigning different importance to components (such as textual embeddings, financial metrics, or categorical tags) to better reflect investment relevance. This two-stage retrieval pipeline (approximate search followed by weighted re-ranking) achieves a balance between speed and accuracy, making the system scalable and robust under current data limitations. It also lays the foundation for future enhancements, such as incorporating user feedback loops, supervised fine-tuning, or personalized recommendation strategies.

5.1 Approach 1: Pure Text Similarity

In this baseline approach, we treated each company as a single textual entity by aggregating all available metadata—including the cleaned business description, sector, and country into a unified, synthetic document. This textual representation was then encoded into dense semantic embeddings using two types of sentence-level language models: Sentence-BERT and OpenAI's `text-embedding-3-small`.

To identify similar companies, we computed similarity scores between the query company and all other companies in the dataset using cosine similarity. For scalability and efficiency, especially with larger datasets, we implemented Facebook AI Similarity Search. This approach is entirely unsupervised and model-driven. It does not require any manual feature engineering or weighting, relying solely on the quality of the embedding model to capture nuanced similarities. The resulting top-N recommendations reflect global semantic similarity as interpreted by the language model, offering a clean and scalable baseline for further evaluation.

5.2 Approach 2: Multi-Feature Similarity Ranking

To address the limitations of a text-only similarity model, this second approach incorporates structured company-level data alongside unstructured textual information. Each company is represented as a concatenation of three distinct feature groups: textual, categorical, and numerical attributes. This multi-feature representation is designed to provide a more holistic view of company similarity, better aligned with private equity (PE) investment criteria.

Textual features were encoded using the same sentence embedding model employed in Approach 1 (e.g., Sentence-BERT and OpenAI), capturing semantic information from company descriptions and mission statements.

Categorical features included variables such as industry sector, company specialities, and country of operation. Given their discrete nature:

- *Specialities* were encoded using multi-label binarization, allowing each company to be associated with multiple relevant categories simultaneously (e.g., diagnostics, imaging, digital health).
- *Industry sectors* were one-hot encoded, capturing mutually exclusive classifications (e.g., pharmaceuticals, medical devices).
- *Country* was also one-hot encoded despite the high cardinality (over 80 distinct values), in order to retain geographic diversity—a factor often deemed strategically significant in healthcare investment decisions.

Numerical features included continuous variables such as company age (in years) and employee count. The employee count variable exhibited a right-skewed distribution, with a small number of companies having disproportionately large workforce sizes. To mitigate this skewness and reduce the impact of outliers, a logarithmic transformation was applied prior to standardization. Both numerical features were then standardized using the `StandardScaler` from the `scikit-learn` library, ensuring zero mean and unit variance. This preprocessing step was critical to ensure that no single feature dominated the similarity computation due to scale differences.

Similarity was computed in a unified, high-dimensional feature space that integrates multiple sources of information. Each company was represented as a concatenated feature vector, combining textual embeddings, categorical encodings, and scaled numerical attributes. Specifically, sentence embeddings derived from company descriptions were horizontally stacked with binarized specialities, one-hot encoded industry sectors, and standardized numerical features (log-transformed employee count and company age).

These vectors were L2-normalized to ensure that inner product similarity corresponded to cosine similarity. Using this normalized matrix, we constructed a FAISS index with `IndexFlatIP`, allowing efficient retrieval of approximate top-k nearest neighbors for each company in the dataset. However, this initial FAISS retrieval was performed over unweighted features, each feature contributed equally to the similarity calculation. To incorporate domain-specific priorities into the ranking, we introduced a post-retrieval reranking step. In this step, we decomposed each candidate pair's feature vector into its original components (text, specialities, industry, country, employee count, company age), computed individual cosine similarities for each group, and then aggregated these using a weighted sum:

$$S(i,j) = \sum_{f \in F} w_f \cdot \cos(\vec{x}_f^i, \vec{x}_f^j)$$

where F is the set of feature groups and w_f is the manually assigned weight for group f. The weights reflected the perceived importance of each feature in healthcare investment screening, with the textual description receiving the highest emphasis. The final top-k most similar companies were thus re-ranked based on this aggregated, weighted similarity score. The chosen weights were:

```
weights = {
    "text": 0.5,
    "specialities": 0.1,
    "industry": 0.1,
    "country": 0.1,
    "employee_count": 0.1,
    "company_age": 0.1
}
```

This two-stage retrieval pipeline, using FAISS for efficient candidate generation and weighted re-ranking for nuanced prioritization allowed us to balance scalability with investment-aligned interpretability.

5.3 Limitations

Despite the overall effectiveness and flexibility of our content-based similarity framework, there are several limitations that may affect the performance and generalizability of the current system. First, although we leveraged powerful pre-trained embedding models such as intfloat/multilingual-e5-large-instruct, Lajavaness/bilingual-embedding-large, and OpenAI's text-embedding-3-small, these models were not fine-tuned on domain-specific healthcare or private equity data. This introduces the risk of domain mismatch, where key investment-relevant semantics (such as clinical trial stages, regulatory context, or investment signals) are either underrepresented or misunderstood by the model. Consequently, companies with highly technical or sector-specific language may be misrepresented in the embedding space.

Second, due to the lack of ground truth labels or explicit user interaction data, our evaluation framework is limited to proxy metrics such as cosine and FAISS. While these provide a useful signal for semantic proximity, they may not accurately reflect real-world investment decision-making. For instance, two companies with similar description may still differ substantially in strategic relevance, business model, or risk exposure. Without curated feedback or historical outcomes, we cannot reliably assess the precision or recall of our recommendations in a business context.

Overall, while the current system offers a scalable and interpretable foundation for content-based company recommendation, these limitations highlight areas for future refinement, including domain-specific model tuning, user feedback integration, having historical interactions and temporal context modeling.

6 Evaluation Metrics and Results

Evaluating the performance of an unsupervised recommendation system is inherently challenging, especially in domains like healthcare private equity where no ground-truth labels or standard benchmarks exist. In our case, the dataset consisted primarily of company information and unstructured metadata. As a result, we adopted a mixed evaluation strategy combining internal quantitative measures with qualitative domain expert assessment. The overall goal is to assess how well each approach returns semantically and strategically relevant company recommendations aligned with the investment sourcing objectives of the firm.

6.1 Quantitative Evaluation

The initial baseline relied on pairwise cosine similarity computed solely from company descriptions using the `intfloat/multilingual-e5-large-instruct` embedding model. While this model performed reasonably well in grouping companies from similar sectors, the results lacked granularity. The descriptions were all quite similar to each other. The simialrity scores ranged from 0.84 to 1.0, with a standard deviation of 0.015 reflecting insufficient discrimination between companies. Motivated by this, we explored alternative embedding models.

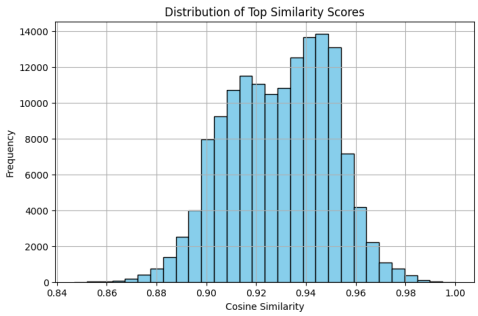

Figure 5: Approach 1: BERT E5 model

Switching to the `Lajavaness/bilingual-embedding-large` model provided more semantically coherent recommendations, especially for non-English entries, likely due to better multilingual handling. Here we have a wider range of simialrity scores, meaning it identifies textual information that are different, with standard deviation of 0.057. Lajavaness is purpose-built for English and French semantic similarity. It uses a higher temperature contrastive loss that spreads cosine scores over a wider range, and does not rely on instruction prefixes. By contrast, E5 is an all-purpose retrieval model for 100 languages, it is great for ranking but not so well for human-readable distance. The Lajavaness model is fine-tuned on tasks and datasets related to semantic similarity, so it learns to emphasize key domain-relevant signals while ignoring superficial similarities. Next, we implemented OpenAI's `text embedding-3-small`, managing to have both semantic precision and diversity in the top-5 lists, generating more relevant and strategically aligned company suggestions. Now, we can dive deeper into the comparison between BERT and OpenAI.

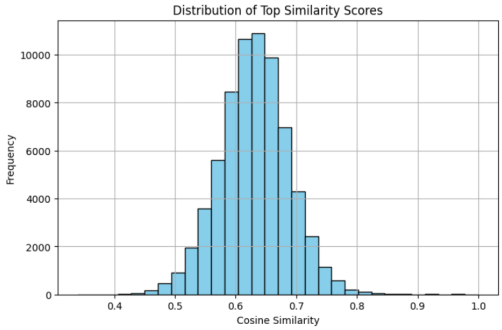

Figure 6: Approach 1: BERT second model

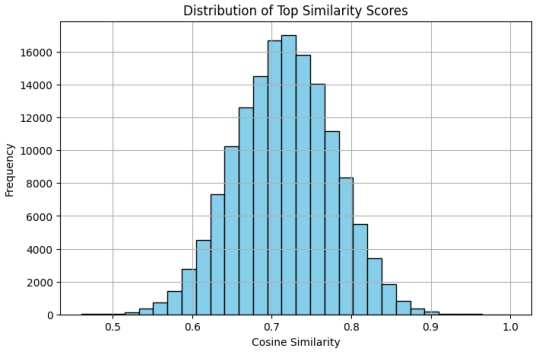

Figure 7: Approach 1: OpenAI

Moving to Approach 2, which integrates multi-feature similarity ranking, the two embedding models exhibited distinct behaviors. The OpenAI model produced a bimodal distribution, with peaks around 0.4 and 0.8, indicating a clear separation between dissimilar and highly similar companies. This suggests that the model is effective at distinguishing strong matches from weaker ones, offering better discriminative power. The wider spread of cosine similarity values (ranging from approximately 0.1 to 1.0) further supports its ability to capture nuanced differences across company features. In contrast, the BERT-based model demonstrated a narrow, unimodal distribution, centered around a cosine similarity of 0.63. The majority of scores clustered around the mean, indicating a tendency to assign moderate similarity to most comparisons. While this behavior results in more stable scoring, it lacks the clear differentiation needed for high-precision matching.

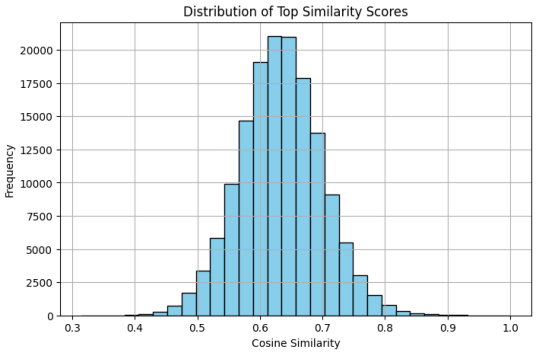

Figure 8: Approach 2: BERT model

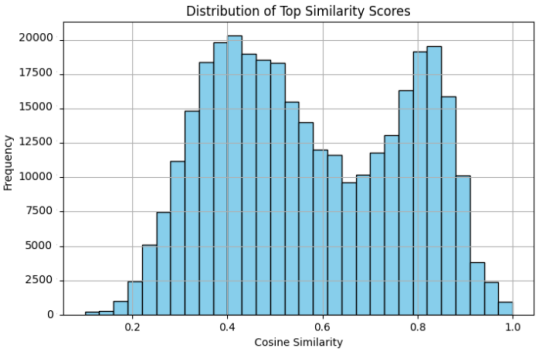

Figure 9: Approach 2: OpenAI

Figures 10 and 11 present t-SNE visualizations of the resulting company embeddings from Approach 2. The Lajavaness embeddings (Figure 10) display substantial cluster overlap, indicating less specialization and more intermixed groups. In contrast, OpenAI embeddings (Figure 11) form more compact, well-separated clusters with minimal overlap, suggesting stronger discriminative power in defining company similarity. These visualizations imply that OpenAI embeddings are more effective for grouping similar companies, potentially enhancing recommendation relevance.

Additionally, to compare results between OpenAI and BERT models we computed the Jaccard similarity of their top-5 most similar companies per query. The average

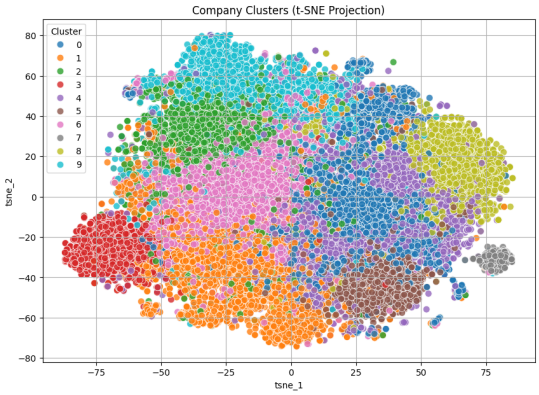

Figure 10: Approach 2: t-SNE for BERT

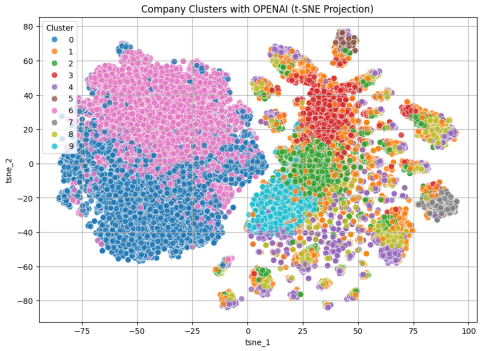

Figure 11: Approach 2: t-SNE for OpenAI

Jaccard overlap was 0.2135, indicating that roughly 21.35% of recommended companies are common between OpenAI and BERT embeddings. The two models often suggest different sets of similar companies. While there is some agreement, roughly 3 to 4 out of the 5 suggestions differ. This relatively low overlap suggests that the models capture distinct notions of company similarity, offering complementary perspectives.

The weighted multi-feature ranking approach provided solid, interpretable results by integrating the relative importance of each feature group into the similarity calculation. In this setting, the OpenAI embedding model outperformed the Lajavaness (BERT-based) model, particularly when combined with FAISS indexing. The synergy of weighted multi-feature similarity and OpenAI embeddings constituted the most effective configuration based on both domain expert assessment and internal quantitative metrics.

In summary, while BERT is powerful for understanding fine-grained sentence context, OpenAI's models offer superior performance when holistic semantic similarity across business descriptions is the priority. This distinction influenced our decision to incorporate both types of embeddings in our experiments and ultimately guided our model selection process.

6.2 Qualitative Review by Investment Experts

The most critical component of our evaluation came from structured feedback sessions with ARCHIMED's investment team. To complement our quantitative evaluation, we conducted a structured qualitative assessment in collaboration with ARCHIMED's investment professionals. This involved a direct feedback exercise designed to simulate real-world use of the recommendation system.

We provided the team with Excel files for each recommendation strategy—corresponding to both Approach 1 (text-only) and Approach 2 (multi-feature)—and for each embedding model: BERT and OpenAI. Each file contained the top 5 suggested similar companies for a curated set of query companies.

Investment professionals were asked to assess each recommended company along the following dimensions:

- **Relevance:** Is the suggested company similar to the query target in a meaningful way, regardless of its attractiveness for investment?

- **Agreement:** A binary Yes/No flag to indicate if the recommendation made sense as a comparable entity.
- **Similarity Score:** An optional numeric score (1–10) reflecting the perceived closeness between the recommended company and the target.

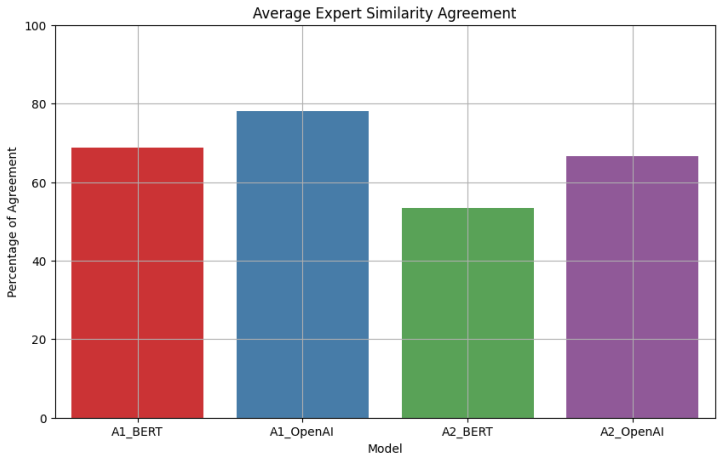

Figure 12: Expert Agreement Percentage

We distributed the evaluation to 23 domain experts, of whom 6 responded. Although the response rate was limited, the feedback provided valuable qualitative insights, articularly around interpretability and alignment with domain expectations—elements that may not be fully captured by automated metrics.

From Figure 12, we can see that OpenAI models outperform sentence-BERT models. In particular, recommendations from Approach 1, using the OpenAI `text-embedding-3-small` model combined with FAISS, consistently received the highest subjective ratings for both accuracy and insightfulness. Overall, expert scores placed the OpenAI + all textual information configuration at the top, with several reviewers noting that the recommendations surfaced both expected and surprising (yet plausible) leads. This blend of accuracy and novelty was cited as critical to the value of the system.

7 Conclusion and Future Work

This work presents a practical and scalable approach to company similarity modeling in the context of private equity deal sourcing, where clean interaction data and outcome labels are often missing. Faced with these constraints, we opted for a content-based similarity system using semantic embeddings to represent companies and retrieve similar entities based on both textual descriptions and structured features comparing BERT models with OpenAI embeddings.

Through comparative evaluation, OpenAI's text-embedding models consistently outperformed BERT-based models, demonstrating superior performance in aligning with expert feedback. The ability of OpenAI embeddings to capture global semantics across longer textual inputs made them particularly effective in understanding nuanced descriptions within healthcare and financial domains.

Our experimental evaluation indicated that the second approach (feature-wise handling) using OpenAI embeddings demonstrated superior quantitative performance. However, expert feedback favored the first approach (whole-document input), citing its alignment with human intuition and better interpretability in some cases. This divergence between empirical metrics and expert judgment underscores a fundamental trade-off between model performance and human-aligned reasoning. Notably, the qualitative assessment was based on a small-scale feedback exercise involving only six experts and a limited sample of 10 target companies out of a dataset of over 30,000. To more reliably determine which approach better serves real-world needs, whether for automated processing or decision support, a more systematic and large-scale user evaluation is necessary. This would allow us to validate whether the preferences observed hold across diverse use cases and user profiles.

Looking ahead, we envision several promising directions for extending this work:

- Collect user feedback in time, by giving them an option to rate the simialr companies.
- Incorporating interaction data (e.g., user validation logs) to build weakly supervised training sets. Integrating collaborative filtering to complement content-based methods once sufficient interaction data is available.
- Retrieve more information about the companies by incorporating richer features, including financial indicators, investment stage data, and macroeconomic trends.

Future work could build on this foundation by developing hybrid models that integrate both empirical similarity metrics and human-in-the-loop signals. For example, user feedback could be used to fine-tune embedding spaces, retrain ranking algorithms, or adapt similarity weighting dynamically based on user profiles or domain-specific relevance. Moreover, semi-supervised or weakly supervised approaches could be introduced to bridge the gap between automated similarity scores and expert judgment, potentially blending the strengths of both global document-level embeddings and feature-wise comparisons. These enhancements would not only improve the precision of recommendations but also promote transparency and user alignment, which is critical for trust and adoption in investment workflows.

In conclusion, this study lays a strong foundation for intelligent company recommendation systems in investment contexts. By combining high-quality embeddings, structured feature modeling, and interactive feedback loops, we move toward more adaptive, explainable, and user-aligned similarity systems that evolve over time.

8 Acknowledgments

I would like to express my deepest gratitude to my supervisor, Ankit, for his unwavering support, guidance, and encouragement throughout this project. I am also sincerely thankful to the investment team members for their invaluable feedback and insights across all versions of the model. Special thanks go to my professors, particularly my academic advisors, for their continuous mentorship and thoughtful advice. Finally, I would like to thank everyone who contributed to my learning and growth during my master's journey—your support has been instrumental in completing this thesis.

References

[1] archimed.

[2] Lele Cao, Vilhelm von Ehrenheim, Mark Granroth-Wilding, Richard Anselmo Stahl, Andrew McCornack, Armin Catovic, and Dhiana Deva Cavalcanti Rocha. Companykg: A large-scale heterogeneous graph for company similarity quantification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4816–4827, 2024.

[3] George W Fenn, Nellie Liang, and Stephen Prowse. The private equity market: An overveiw. *Financial Markets, Institutions & Instruments*, 6(4):1–106, 1997.

[4] Sonja Horn. Deep learning models as decision support in venture capital investments: Temporal representations in employee growth forecasting of startup companies, 2021.

[5] Johannes Luef, Christian Ohrfandl, Dimitris Sacharidis, and Hannes Werthner. A recommender system for investing in early-stage enterprises. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1453–1460, 2020.

[6] Chris O'Brien. How eqt ventures' motherbrain uses ai to find promising startups. *venturebeat*, 2020.

[7] Samantha Petersone, Alwin Tan, Richard Allmendinger, Sujit Roy, and James Hales. A data-driven framework for identifying investment opportunities in private equity. *arXiv preprint arXiv:2204.01852*, 2022.

[8] Axel Springer Tech. Finding startups with bert. *medium.com*, 2020.

[9] Dimitrios Vamvourellis, Máté Tóth, Snigdha Bhagat, Dhruv Desai, Dhagash Mehta, and Stefano Pasquali. Company similarity using large language models. In *2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pages 1–9. IEEE, 2024.