CHALMERS
UNIVERSITY OF TECHNOLOGY

AI Assisted matching in Mergers And Acquisitions

A Data-Driven Approach to Identifying Potential Acquirers

Master's thesis in Data Science and AI

WILHELM JOHNSON SWEGMARK
DIDRIK TVEDT

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2026

www.chalmers.se

Master's thesis 2026

AI Assisted Matching in Mergers and Acquisitions

A Data-Driven Approach to Identifying Potential Acquirers

WILHELM JOHNSON SWEGMARK

DIDRIK TVEDT

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences

Chalmers University of Technology

Gothenburg, Sweden 2026

AI Assisted Matching in Mergers and Acquisitions

A Data-Driven Approach to Identifying Potential Acquirers

WILHELM JOHNSON SWEGMARK

DIDRIK TVEDT

Supervisor: Johan Jonasson, Department of Mathematical Sciences

Examiner: Johan Jonasson, Department of Mathematical Sciences

Master's Thesis 2026

Mathematical Sciences

Chalmers University of Technology

SE-412 96 Gothenburg

Telephone +46 31 772 1000

Cover:

Typeset in LaTeX

Printed by Chalmers Reproservice

Gothenburg, Sweden 2026

AI Assisted Matching in Mergers and Acquisitions

A Data-Driven Approach to Identifying Potential Acquirers

WILHELM JOHNSON SWEGMARK

DIDRIK TVEDT

Department of Mathematical Sciences

Chalmers University of Technology

Abstract

Traditional buyer identification in M&A relies on manual screening and professional networks, making it resource-intensive and naturally limiting the buyer pool. This thesis investigates whether textual embedding models can support the identification of relevant potential buyers in mergers and acquisitions. The study examines how different representation methods, including TF-IDF, Doc2Vec with smooth inverse frequency weighting, and Transformer based models, capture similarity between companies when applied to standardized summaries of portfolio company descriptions. The summaries are created using a large language model with information provided on the portfolio companies websites. The performance of the embedding models is evaluated through visualization of the embedding spaces, cosine similarity search experiments, and an expert review of buyer recommendations. The results indicate that TF-IDF and the Transformer model produced relevant recommendations, with the Transformer model demonstrating the best performance in embedding space separation and alignment with expert judgment, while Doc2Vec models showed weaker differentiation between company types. Overall, the study shows that embedding based similarity search can serve as a useful first step in buyer discovery by expanding the range of potential buyers considered and improving efficiency. The work also highlights that further validation across a larger set of targets and with a more complete dataset would strengthen confidence in these results.

Keywords: M&A, NLP, LLM, Embeddings, Semantic Similarity

Acknowledgements

We would like to express our sincere gratitude to our Chalmers supervisor, Johan Jonasson, whose expertise, guidance, and feedback have been invaluable throughout the course of this thesis. Your support has helped us navigate challenges, refine our ideas, and ultimately shape the direction of our work.

We would also like to extend our appreciation to Merge, with which this thesis was carried out. The opportunity to work closely with the company, along with the access to industry knowledge, data and practical perspectives, has contributed greatly to the development and relevance of this project.

Wilhelm Johnson Swegmark and Didrik Tvedt, Gothenburg, December 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
LLM	Large Language Model
M&A	Mergers and Acquisitions
NLP	Natural Language Processing
PCA	Principal Component Analysis
SIF	Smooth Inverse Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction

Contents

List of Figures

List of Tables

1

Introduction

Identifying potential buyers is a critical first step in the mergers and acquisitions (M&A) process, directly influencing deal success and transaction outcomes (Merge, 2025). Brokers, investment banks, and advisory firms traditionally rely on their networks, relationships, curated databases of past transactions, and manual screening to build buyer lists (Merge, 2025). Although this human expertise is invaluable, it is also resource intensive and prone to biases that limit both scalability and diversity in the buyer lists produced. Consequently, buyer lists may become concentrated around well-known or easily identifiable investors, such as major private equity funds. This can inadvertently limit the consideration of smaller or less conventional buyers, reducing the breadth of potential opportunities. This creates a clear business need for systems that can complement human expertise by discovering non-trivial, high-potential buyers, reducing manual effort, and broadening the coverage of the buyer universe.

A further challenge lies in the fact that potential buyers differ fundamentally in their motivations and acquisition strategies. An important distinction is between financial and strategic buyers (Merge, 2025). Financial buyers, such as private equity firms, venture capital funds, and investment funds, are typically motivated by return on investment, exit strategies, and growth potential. Their acquisition considerations often emphasise cash flows, profitability, and other financial metrics, and they usually hold assets for a limited period before seeking a profitable exit. Strategic buyers, on the contrary, are operating companies that acquire for reasons tied to long-term positioning, such as vertical or horizontal integration, geographic

expansion, or access to new technology and capabilities. These buyers may be willing to pay a premium because synergies raise the potential value of the acquisition, and they often place greater emphasis on non-financial considerations like cultural fit, strategic alignment, and competitive positioning. They also tend to integrate acquisitions more directly into their operations, whereas financial buyers may leave management largely in place. Understanding these distinctions is crucial for designing a recommendation system, since the relevance of a buyer depends on its incentives, constraints, and historical behaviour.

Artificial intelligence and machine learning open up opportunities to address these challenges by enabling more data-driven and scalable approaches to buyer identification. By capturing richer representations of companies, modelling patterns of similarity, and integrating diverse types of information, such methods can reduce the dependence on manual processes. Importantly, these technologies can be designed to support, rather than replace, human expertise and thus help brokers generate more diverse and high-quality recommendations while maintaining interpretability.

This motivates the present study, which explores how AI-driven methods can contribute to a more efficient, data-driven, and insightful process for buyer–seller matching in M&A.

1.1 Aim

The general aim of this thesis is to investigate how AI can be applied to improve the process of buyer identification in mergers and acquisitions. Specifically, the goal is to design a method that generates prioritised and interpretable buyer recommendations for a given target company, thus reducing reliance on manual processes and increasing efficiency in generating leads.

1.2 Research Questions

The research questions for the thesis are stated below:

1. How can we embed companies to best represent the business in operative terms given publicly available data?
   (a) How can dimensionality reduction techniques be used to visualize these embeddings and provide insight into the structure of embedding spaces?
   (b) Which dimensions of a firm's business model (e.g. product offerings and operational domain) are encoded in these embeddings?
2. Can we use these embeddings to find companies with a similar business model?
   (a) How can we use these embeddings to generate a buyer recommendation of a given target company?
   (b) How does performance differ between different embedding models?

1.3 Limitations

Several limitations influence both the scope and the technical approach taken. First, the study is based on a list of financial buyers provided by Merge with a strong concentration of Nordic buyers but also including several major global buyers. This means that the portfolio companies of these buyers may extend internationally. Strategic buyers are not included in this analysis. The primary reason is that no comprehensive or standardized list of potential strategic acquirers exists. Identifying them would require considering virtually every corporation that might acquire a company for strategic reasons. Because of the lack of definable boundaries and the impractical scale of data collection, we chose not to pursue the inclusion of strategic buyers in this study.

On the technical side, the study emphasizes AI methods that can be executed on CPU compute, ensuring that any resulting application can run efficiently in Merge's environment without the need for GPU resources.

Another limitation concerns the lack of labelled data that could be used to directly supervise and benchmark machine learning models. Instead, evaluation must rely on indirect methods, such as conducting manual assessments together with domain experts. While this allows for valuable qualitative insights, it also introduces subjectivity and reduces the extent to which model performance can be measured using standard quantitative metrics.

1.4 Ethical aspects

Training state of the art Large Language Models (LLMs) can carry a large environmental cost. Strubell et al. (2019) estimate that the total carbon emissions associated with training and tuning a single LLM including hyperparameter tuning, architecture exploration, and repeated runs, can exceed the equivalent of 284 tons of CO_2. This amount is comparable to or greater than the lifetime emissions of multiple passenger vehicles and has likely exploded further in recent years. These emissions reflect the electricity consumed by GPUs and data centre infrastructure during extensive experimentation, rather than a single training run alone. Even though using readily available models doesn't infer a full training regime model selection should still consider not only accuracy but also computational and energy demands. Even though using readily available pretrained models does not require repeating the full training regime, model selection should still consider not only KPIs such as accuracy but also computational and energy demands. While this does not alter how foundation models are trained broadly, but a default preference for efficient models helps ensure that environmental cost remains an explicit consideration rather than an afterthought.

A further ethical consideration concerns how differences in data availability across companies may influence the results. Companies with extensive publicly available information, and well-structured websites are more likely to be represented by informative summaries. In contrast, smaller firms or companies operating in less digitally

mature contexts may have limited online information. As the work is based on public information, the models may systematically favour companies with more data, not necessarily because they are more relevant, but because their representations are more complete. This introduces a form of representation bias, where visibility and data richness influence buyer recommendations. Consequently, some potentially relevant buyers or portfolio companies may be overlooked. While this limitation is largely driven by data availability rather than model design, it shows the importance of interpreting model outputs with caution and complementing automated recommendations with human judgement, particularly in high-stakes decision-making situations such as mergers and acquisitions.

1.5 AI Declaration

AI tools (e.g., ChatGPT, Grammarly) were used for language support such as grammar correction, and rephrasing. All content, analysis, and conclusions are the authors own.

2

Theory

This chapter establishes the theoretical foundation for the embedding and similarity methods employed in this study. It covers semantic textual similarity and the algorithms used to measure it, ranging from classical approaches like TF-IDF to modern Transformer-based models, concluding with cosine similarity as the metric for comparing vectorized representations.

2.1 Term Frequency-Inverse Document Frequency

A challenge in representing textual data for machine learning tasks is how to Transform the unstructured text into features that capture the importance of the words. One of the most widely used methods in information retrieval is TF-IDF, originally formalized Salton and Buckley (1988). The method addresses the challenge of quantifying the importance of words in a document relative to a larger collection of documents (a corpus). By combining local and global weighting, TF-IDF captures not only how often a term appears within a document, but also how distinctive it is across the corpus.

The term frequency component measures how often a term t occurs in a given document as shown in equation 2.1. To avoid bias toward longer documents, the frequency is normalized by the total number of terms in that document.

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t,d}} \tag{2.1}$$

Here $f_{t,d}$ is the count of the term t in document d. This ensures that TF represents the relative importance of a word within the document itself. The inverse document frequency component adjusts for the fact that certain words are common across the entire corpus, and therefore carry limited discriminative power. This is described in equation 2.2.

$$IDF(t, D) = log\left(\frac{N}{1 + |\{d \in D : t \in d\}|}\right) \tag{2.2}$$

Here N is the total number of documents in the corpus D, and the denominator counts how many documents contain the term t. The logarithm serves to dampen the effect of very frequent words, while the addition of 1 prevents division by zero. The TD-IDF score is obtained as the product of these two components TF and IDF. This weighting scheme highlights terms that are frequent within a document but rare across the corpus, making them useful for distinguishing that document from others. Common words such as “the”, “or”, and “will” therefore receive low weights, while domain-specific or distinctive terms receive higher values.

Although TF-IDF is defined at the term level the implementation of Salton and Buckley (1988) is primarily used to construct vector representations of entire documents. After computing the TF-IDF weight for every term t in the vocabulary, each document d is represented as a vector as shown in Equation 2.3.

$$\mathbf{v}_d = (w_{1,d}, w_{2,d}, \ldots, w_{V,d}), \tag{2.3}$$

In Equation 2.3 where V is the size of the vocabulary and $w_{i,d}$ is the TF-IDF value of term i in document d. Terms that do not appear in a document receive a weight of zero, resulting in a high-dimensional but sparse vector. These vectors provide a simple yet effective representation of documents and form the basis for tasks such as similarity measurement, clustering, and document ranking.

2.2 Word2Vec

An important contribution to natural language processing (NLP) was made by Mikolov et al. (2013) in their work ´´Efficient Estimation of Word Representations in Vector Space", which introduced the Word2Vec framework. It enables words to be encoded as dense, continuous-valued vectors in a way that they capture semantic meaning. The core idea is to learn these vector representations through a simple neural network trained on a large corpus of text. Instead of representing words as discrete symbols, Word2Vec encodes them in a continuous vector space such that words occurring in similar contexts have similar embeddings (Mikolov et al., 2013). To train the embeddings, the network predicts either a target word given its surrounding context (the Continuous Bag-of-Words, or CBOW, model) or the surrounding context words given a target word (the Skip-gram model) (Mikolov et al., 2013). Both variants share the same underlying architecture of a single hidden layer neural network with a linear transformation from the one-hot encoded input to a dense embedding space. Finally a softmax output layer produces a probability distribution over the vocabulary.

For a skip-gram model, we predict context words N steps away from a given target word in the sequence. To achieve robust training, context words are sampled with probability correlated to their distance from the target word Mikolov et al. (2013). In the standard Word2Vec implementation by Mikolov et al. (2013), each word is represented by two vectors: a target vector v_w and a context vector v_c. The likelihood of observing actual context words given a target word and model parameters θ across the training corpus is modelled using the dot product $v_c \cdot v_w$ as shown in Equation 2.4 where C denotes the set of all context words. During training, the model alternates between fixing one set of vectors and optimizing the other, iterating until convergence (Mikolov et al., 2013).

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}} \tag{2.4}$$

By maximizing the corpus probability of the context c given the target word w and by taking the log of that expression the sum as shown 2.5 is attained. Let T denote the set of all (word, context) pairs extracted from the corpus.

$$\arg\max_{\theta} \sum_{(w,c)\in T} \log p(c|w;\theta) = \sum_{(w,c)\in T} \left(v_c \cdot v_w \; - \; \log \sum_{c'\in C} e^{v_{c'} \cdot v_w} \right) \tag{2.5}$$

As shown in 2.5 this optimization requires summarizing over all context words c' requiring lots of compute for training especially for large context windows. By introducing negative samples denoted T' the task can be modelled as a binary classification task with a likelihood shown in 2.6.

$$p(D = 1|w, c; \theta) = \frac{1}{1 + e^{-v_c \cdot v_w}} \tag{2.6}$$

This removes the worst summation step resulting in a simpler training routine. The objective function to optimize now becomes 2.7

$$\arg\max_{\theta} \sum_{(w,c)\in T} \log \frac{1}{1 + e^{-v_c \cdot v_w}} + \sum_{(w,c)\in T'} \log \frac{1}{1 + e^{v_c \cdot v_w}} \tag{2.7}$$

For the Bag of Words model the context is made up of all terms contained in a symmetric window around the target word where each context word is encoded as a bag of words vector and the neural net outputs a probability vector via softmax activation. The model is trained in the same manner as for skip grams using a cross entropy loss except the context is now the sum of word vectors.

Word2Vec models such as CBOW and Skip-gram provide effective representations for individual words, whereas plenty of applications require vector representations of longer textual sequences, such as sentences or documents. An intuitive approach is to achieve this by averaging the word vectors contained within a text possibly using a weighted average. Although this method produces a fixed-length representation, it misses information about word order and context.

2.3 Doc2Vec

To address limitations of Word2Vec for sequence embeddings, Doc2Vec was introduced by Le and Mikolov (2014) as an extension of Word2Vec. Each paragraph is associated with a unique vector p_j that is shared across all contexts and sampled from the same paragraph (Le and Mikolov, 2014). The word vector matrix on the other hand is shared globally making words retain meaning across different paragraphs (Le and Mikolov, 2014). There are two common implementations Doc2Vec corresponding conceptually to Continous Bag Of Words (CBOW) and Skip-gram.

The first implementation called Distributed Memory Model of Paragraph Vectors (PV-DM) starts by mapping every paragraph to a unique vector, corresponding to a column in matrix P and each word is represented as a vector, corresponding to a column in matrix W (Le and Mikolov, 2014). The combination of the word vectors v_W with the paragraph vector p_j is achieved through concatenation or averaging and allows the model to capture semantic and contextual dependencies (Le and Mikolov, 2014). After training, the learned paragraph vectors can be directly used as features for downstream machine learning tasks. The PV-DM model offers several advantages over bag of words representation as it captures semantic relationships between words and accounts for local word order, similar to high-order n-gram models but without resulting in high-dimensional, sparse representations (Le and Mikolov, 2014).

As opposed to the PV-DM model, that combines the paragraph vector with word vectors as input to predict the next word, the Distributed Bag of Words model (DBOW) simplifies the problem by using only the paragraph vector as input to predict randomly sampled context words (Le and Mikolov, 2014). This approach is conceptually similar to the skip-gram model, where the task is to predict words appearing within a given context window. During training, a text window is sampled from the paragraph, and a random word within that window is used as the target word. The model then performs a classification task to predict this word based only

on the paragraph vector (Le and Mikolov, 2014). This design not only simplifies training but also reduces storage requirements as only the softmax weights has to be stored rather than both the softmax weights and word vectors as in PV-DM (Le and Mikolov, 2014).

2.4 SIF embeddings

To address the limitations inherent in standard averaging of word vectors, Arora et al. (2017) proposed the Smooth Inverse Frequency (SIF) method. The method first uses a Word2Vec model to build embeddings v_w for each word w. The word level embeddings are combined for each sentence $s \in S$ using a weighted average as described in Equation 2.8. sentence is defined as any sequence of tokens and can also represent longer documents. Note that $p(w)$ is the empirical probability of observing word w in our corpus and a is a hyper-parameter to be set.

$$v_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w \tag{2.8}$$

Recall that in the training objective of Word2Vec frequent words are down-weighted because they appear in many contexts and are assumed to represent limited information. The SIF method achieves a similar effect through its weighting function. When $p(w)$ is large, corresponding to highly frequent words, the weigh as described in Equation 2.8 becomes smaller (Arora et al., 2017). Conversely, when $p(w)$ is small, representing rare words, the weight approaches 1 (Arora et al., 2017). The name "smooth inverse frequency" describes this functional behaviour as the weighting is approximately proportional to $\frac{1}{p(w)}$ for frequent words but converges smoothly to 1 for rare words instead of exploding.

The final stage of the SIF embedding method is a normalization that involves the removal of the dominant component shared across sentence embeddings (Arora et al., 2017). After constructing the sentence embedding matrix X from the set of sentence embeddings $v_s : s \in S$, the first singular vector u is computed via singular value

decomposition and normalized to unit length. Each normalized sentence vector $\tilde{v}_s$ is then obtained by subtracting from v_s its projection onto the unit-normalized u, as shown in Equation 2.9.

$$\tilde{v}_s = v_s - (u \cdot v_s)\, u \tag{2.9}$$

This final operation serves as a normalization by filtering out the most common variance amongst the sentence embeddings. Arora et al. (2017) demonstrate through empirical analysis that the leading singular vector u mainly captures common function words, rather than meaningful semantic information. Words exhibiting the highest cosine similarity to u in their study include typical stop words such as "but", "when", and "even". This adjustment ensures that the final representations better reflect the actual semantic content of sentences rather than generic grammatical structures.

2.5 Transformers

The Transformer architecture, introduced by Vaswani et al. (2017) in the paper "Attention Is All You Need", marked a shift in natural language processing (NLP) by replacing recurrent and convolutional structures with a fully attention-based mechanism. The key innovation of the Transformer is the self-attention mechanism, which computes contextual relationships between all tokens in parallel. It allows the model to weigh each token's importance relative to others, capturing long-range dependencies regardless of their position in the sequence. As shown in the multi-head attention block of Figure 2.1, this enables the Transformer to attend to the most relevant parts of the input and build richer contextual representations. Self-attention is computed according to Equation 2.10 (Vaswani et al., 2017).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2.10}$$

Within each head of the multi-head attention block of Figure 2.1, the embedding

of each input token including its positional encoding, is projected into three vector spaces using the weight matrices W_Q, W_K and W_V. These weight matrices are parameters learned during training, specific to each attention head. These projected vectors will be called Query (Q), Value (V) and Key (K). The Q and K vectors are used to calculate how strongly each token should attend to every other token in the sequence. This is done by taking the dot product QK^T, which produces a matrix of attention scores and essentially measures the contextual relationship between all pairs of tokens. To prevent these scores from becoming too large when the dimensionality of the K vectors d_k is high, the result is scaled by $\sqrt{(}d_k)$ (Vaswani et al., 2017). The softmax function is then applied to each row of this matrix to convert the raw scores into normalized attention weights that sum to one. Finally, these weights are used to compute a weighted sum over the corresponding V vector, giving a new representation that integrates information from the entire sequence. By combining multiple attention mechanisms in parallel, the Transformer can learn to focus on different aspects of the sentence structure simultaneously (Vaswani et al., 2017). Additional residual connections and layer normalization are applied around each attention feed-forward block to stabilize optimization and maintain gradient flow during training.

In the full Transformer architecture, the decoder (shown on the right-hand side of Figure 2.1) mirrors the encoder's structure but introduces two key modifications. First, it includes a masked multi-head self-attention mechanism that ensures the model can only attend to previous positions in the output sequence (Vaswani et al., 2017), preserving the autoregressive nature of generation. Second, a cross-attention layer is inserted between the self-attention and feed-forward sublayers. This layer allows the decoder to attend to the encoder's outputs, effectively connecting the encoded source representations with the tokens being generated. The combination of masked self-attention and cross-attention enables the decoder to generate output sequences while conditioning on the full encoded representation of the input.

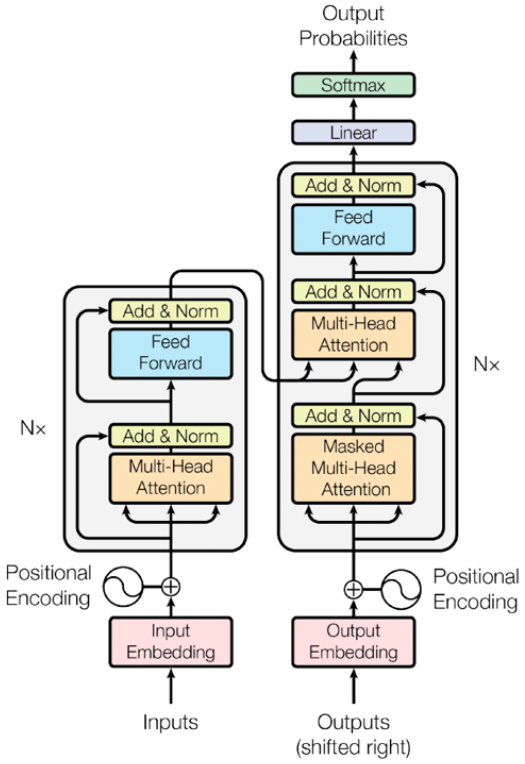

Figure 2.1: Overview of the Transformer architecture (Vaswani et al., 2017), with the encoder part on the left and the decoder part on the right

In the standard training setup, Transformer models are typically first pretrained using self-supervised learning on large unlabeled text corpora, where the objective is autoregressive next-token prediction (Kalyan et al., 2021) . During this phase, the decoder learns to predict the next token given all previously observed tokens, enforced through masked self-attention. The model is then commonly fine-tuned in a supervised manner on smaller, task-specific datasets. In the original Transformer architecture for machine translation, next-token prediction in the decoder is additionally conditioned on the encoder's representation of the source sequence (Vaswani et al., 2017).

2.5.1 Sentence Embeddings

The contextual nature of Transformer representations provides the foundation for obtaining meaningful sentence-level embeddings. Note that a "sentence" in this context can be an arbitrarily long sequence of text, rather than an actual grammatical

sentence. Devlin et al. (2018) proposed Bidirectional Encoder Representations from Transformers (BERT), a model that is pre-trained by masking tokens from an unlabelled input and finally fine-tuned on labelled data for downstream tasks. The Bi-directionality comes from the fact that unlike the sequential pre-training of GPT and RNNs, BERT may use both left and right context of the masked tokens for prediction (Devlin et al., 2018). The input sequence expects a [CLS] token, that represents the entire sequence at the start and a [SEP] token at the end of each sequence. The last hidden state for this token can be used as a sequence representation for classification tasks. Further BERT expects 512 tokens as input so if the sequence exceeds this it must be truncated and conversely in the case it falls of that the sequence is filled with [SEP] tokens to ensure a constant sequence length (Devlin et al., 2018).

While BERT was a big improvement in NLP, it was designed and trained for classification tasks rather than semantic similarity. To address this gap, Reimers and Gurevych (2019) introduced Sentence-BERT (SBERT), modifying the BERT architecture using siamese and triplet network structures to generate fixed-length sentence embeddings suited for similarity comparisons. SBERT can be used in three ways to get sentence embeddings from the Transformer output; using the CLS token representation, computing the mean across all output vectors, or applying max-pooling across the output vectors (Reimers and Gurevych, 2019). Among these approaches, mean pooling has emerged as the most widely adopted strategy in practice (Reimers and Gurevych, 2019). The improvement of SBERT is not only in its architecture and pooling, but more so in its fine-tuning methodology. The SBERT fine-tuning trains the model on sentence pairs with known similarity relationships, enabling it to learn representations where semantically similar sentences are positioned close together in the embedding space while dissimilar sentences remain distant (Reimers and Gurevych, 2019).

2.6 Dimensionality Reduction Methods

High-dimensional data often contain redundant or noisy information that can obscure underlying patterns and relationships. Dimensionality reduction techniques aim to project data from a high-dimensional space into a lower-dimensional representation while preserving as much of the original structure as possible.

Dimensionality reduction approaches can be categorized into *linear* and *non-linear* methods. Linear techniques, such as Principal Component Analysis (PCA), assume that the data lie approximately on a linear subspace of the original feature space and identify directions that capture the most variance. Non-linear methods, including Uniform Manifold Approximation and Projection (UMAP), relax this assumption and instead focus on preserving local and global relationships between data points on a curved manifold. In the following subsections, both PCA and UMAP are described in greater detail, including their mathematical formulations and key intuitions.

2.6.1 PCA

Principal Component Analysis (PCA) is a foundational method for dimensionality reduction that seeks to represent high dimensional data in a more compact form while preserving as much of its original structure as possible. It works by identifying orthogonal directions in the data, known as principal components, that successively capture the greatest possible variance revealing the directions of maximal information content in the dataset (Shlens, 2014). This reduction is achieved by computing the covariance matrix of the data and extracting its eigenvectors and eigenvalues, where the eigenvectors form the orthogonal axes of the new feature space and the eigenvalues quantify the amount of variance each axis accounts for in the dataset (Shlens, 2014).

When applied to vector embeddings, PCA acts as a projection method that compresses high dimensional representations by removing less informative components. Each embedding vector x_i is projected onto the subspace spanned by the top k

principal components, producing a reduced representation $z_i = W_k^T x_i$, where the columns of W_k consists of the eigenvectors associated with the k largest eigenvalues of the covariance matrix (Ringnér, 2008). The fraction of the dataset's total variance preserved through this transformation, the explained variance ratio, is calculated as the sum of these top k eigenvalues divided by the total sum of all eigenvalues. This measure guides the choice of dimensionality k, providing a balance between maintaining the essential informational structure of the embeddings and improving computational efficiency (Ringnér, 2008).

2.6.2 UMAP

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction method introduced by McInnes et al. (2020). The algorithm aims to produce a low-dimensional representation that preserves both local and broader structural relationships present in the original high-dimensional space.

UMAP begins by constructing a weighted k-nearest neighbour graph. For each data point x_i, it defines a local connectivity radius ρ_i as the smallest non-zero distance to any of its neighbours, see Equation 2.11.

$$\rho_i = \min\{d(x_i, x_j) \mid 1 \leq j \leq k,\, d(x_i, x_j) > 0\}. \tag{2.11}$$

Distances are then converted into membership strengths that quantify how strongly two points are connected as shown in Equation 2.12.

$$w((x_i, x_j)) = \exp\left(\frac{-\max(0,\, d(x_i, x_j) - \rho_i)}{\sigma_i}\right), \tag{2.12}$$

σ_i normalizes the local neighbourhood. These weights form a fuzzy simplicial set representing the connectivity structure of the data. The intuition is that the closer two points are in the original space, the higher the probability that they belong to the same local neighbourhood.

UMAP then optimizes a low-dimensional embedding $Y = (y_1, \ldots, y_n)$ that preserves these relationships. In the embedding space, connectivity is modelled using a smooth kernel function. Let $d_Y(y_i, y_j)$ denote the Euclidean distance between two embedded points. The low-dimensional relationship is modelled as Equation 2.13.

$$\tilde{w}((y_i, y_j)) = \frac{1}{1 + a\, d_Y(y_i, y_j)^{2b}} \tag{2.13}$$

with parameters a and b chosen to match the decay of connectivity observed in the high-dimensional graph. This ensures that nearby embedded points receive high membership weights, while distant points contribute minimally. The final embedding is obtained by minimizing the cross-entropy between high- and low-dimensional membership strengths, as seen in Equation 2.14

$$\mathcal{L} = \sum_{i \neq j} \Big[w((x_i, x_j)) \log \tilde{w}((y_i, y_j)) + (1 - w((x_i, x_j))) \log(1 - \tilde{w}((y_i, y_j))) \Big]. \tag{2.14}$$

Through this optimization, strongly connected points remain close in the embedding, while weakly connected points are pushed apart. This makes UMAP particularly suitable for visualizing complex datasets such as text embeddings, where both global clustering patterns and fine-grained neighbourhood relationships provide insights.

2.7 Similarity search

Similarity search refers to the process of identifying objects that are the most alike according to a defined measure of proximity in a vector space. In the context of textual data embeddings are usually expressed as vectors in a high-dimensional space. The task of similarity search in this domain is thus to determine which vectors are closest to each other, which in turn indicates semantic or contextual resemblance.

An important part of similarity search is choosing how to measure the closeness

between two vectors. One of the most common measures used for comparing text embeddings is cosine similarity. It measures the cosine of the angle between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ as shown in 2.15, showing how similar their directions are in the vector space. In Equation 2.15, $\mathbf{x}_i \cdot \mathbf{x}_j$ is the dot product of the two vectors, and $\|\mathbf{x}_i\|$ and $\|\mathbf{x}_j\|$ are their Euclidean norms.

$$\text{sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}, \tag{2.15}$$

The similarity value ranges from -1 to 1, where 1 indicates perfect alignment (identical direction in the embedding space), 0 orthogonality (no similarity), and -1 opposite orientation.

Cosine similarity is particularly well suited for textual representations because it focuses on the orientation of the vectors rather than their magnitude. This property ensures that two documents or descriptions with similar patterns of term importance or semantic meaning are considered close, even if they differ in length or scale.

2.8 Large Language Models for Summarization

Large Language Models (LLMs) are a class of large scale models that builds on the Transformer architecture (Vaswani et al., 2017). While models such as BERT (Devlin et al., 2018) use an encoder-only architecture for deep contextual understanding, decoder-only models are optimized for next-token prediction in an autoregressive manner. This design, introduced in models like GPT (Radford et al., 2018), generates text sequentially by conditioning each token on all preceding ones (Radford et al., 2018). Through extensive pre-training on large unlabelled text corpora where the model learns to predict the next token, these models gain deep linguistic knowledge. Decoder based LLMs thus form the foundation of chatbots, capable of maintaining context and producing fluent responses to prompts.

When generating text, a decoder only LLM computes a probability distribution over

its vocabulary for each token position. The next output token is selected using a sampling strategy, such as greedy sampling (choosing the most probable token) or stochastic sampling, which introduces controlled randomness. The randomness of sampling is typically regulated using a temperature parameter before applying the softmax function. Formally, for a token i with output logit z_i , the probability P_i under temperature $T \in [0, 1]$ is given by the softmax Equation 2.16.

$$P_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{2.16}$$

Lower temperatures (approaching zero) produce more deterministic and focused outputs, preferring tokens with higher predicted probabilities, whereas higher temperatures increase the likelihood of sampling less probable tokens, increasing diversity, and creative text generation (Holtzman et al., 2019).

Because LLM text generation involves stochastic sampling, even identical prompts can produce slightly different outputs. This variability can affect reproducibility, which is important to consider in the embedding analysis. Controlling decoding parameters and standardizing prompts helps reduce this effect. The temperature parameter represents a trade-off where lower values improve reproducibility by favoring high probability tokens, while higher values allow more diverse outputs. Choosing an appropriate temperature balances reliable results with maintaining the model's generative performance.

3

Method

This chapter describes the methodology for developing and evaluating the buyer target recommendation system. It covers the data collection and preprocessing steps, followed by the implementation and analysis of embedding models. Building upon these embeddings, the similarity based retrieval system was then constructed and evaluated through sampling of target companies.

3.1 Data

The primary data used in this study consisted of textual descriptions of companies rather than numerical or financial data. All data were collected from company websites to ensure high relevance and consistency. Websites generally contained comprehensive and current descriptions of a company's activities giving clues about the business model. The tricky part was to find the useful information in the webpage without a too complex logic. Using only textual data was motivated by the study's focus on identifying semantic similarities between businesses, where the textual representation of their activities provides richer descriptive information than financial indicators. Further, the financial data available through public datasets did not cover all markets where portfolio companies were present.

The data collection process began with a curated list of financial buyers, from which all portfolio companies were identified. Because raw website text often was inconsistently formatted with redundant information, a summarization step was introduced before embedding. An intermediate large language model (LLM) was used to trans-

form the scraped content into concise and standardized company descriptions. The standardized summaries form the textual representations used in later stages of the pipeline, including embedding generation and similarity search.

3.1.1 Scraping

Since no centralized or publicly available dataset of portfolio companies existed, a web scraping protocol was developed in collaboration with Merge to collect the necessary data. The process began with a curated list of financial buyers provided by Merge, each associated with a verified company website URL. These websites typically contain sections that describe the buyer's portfolio or list of holdings, which served as the starting point for the data retrieval routine.

The first step of the scraping involved identifying the specific webpage that listed the buyer's portfolio companies. This was achieved by searching the homepage of each buyer for links containing relevant keywords such as "Portfolio or "Holdings." When such a link was found, the crawler followed it to access the portfolio section of the website.

Then, a crawler systematically traversed these portfolio pages to locate subpages that contained information about individual portfolio companies. From these subpages, all external hyperlinks (href attributes) were extracted. To ensure relevance, only links pointing to external company domains were retained, while links associated with social media platforms (e.g., LinkedIn, Twitter, Facebook) or general navigation elements were filtered out. The output of this stage was a structured mapping between each financial buyer and the corresponding list of portfolio company URLs. When analysing how often each portfolio company link appeared across portfolios, certain domains reoccurred at unusually high frequencies. These were not actual portfolio companies but external sources such as news sites or financial data pages. Because these invalid records usually were among the most frequently occurring domains, the dataset could be improved by systematically inspecting and removing entries with the highest counts.

Once the list of portfolio company websites had been established, the next step was to extract descriptive text for each company. For every company, the raw HTML content was retrieved from text-bearing elements which contain the main written material on a webpage. The extraction focused on two key sources of information: the company's landing page and its "About us" page, when available. The landing page was easily identified as the root domain and provided a concise overview of the company's offering and positioning. Locating the "About us" page required an additional search step, as its structure and URL varied between companies. To identify it, the crawler searched for links within the site containing indicative keywords such as "about", "who we are", or "what we do". When such a link was found, the corresponding page was scraped and its textual content extracted.

This approach ensured that the collected text captured both the general presentation of the company and its self-described purpose and activities. The resulting HTML texts from the landing page and the identified "About us" page were later parsed and processed into clean textual representations for the summarization and embedding steps described in subsequent sections.

3.1.2 Summarization via LLM

After having retrieved the full HTML content from the selected webpages using `BeautifulSoup`, the textual material was extracted through the `text` attribute of the parsed HTML object. This unstructured text served as input to a large language model (LLM), which was used to generate standardized and coherent summaries suitable for embeddings. The purpose of this step was to convert noisy website text into concise, comparable descriptions that consistently capture the key characteristics of each company.

The summarization was performed using the GPT-4o-mini model, which provided a 128,000-token context window (OpenAI, 2025). This capacity ensured that both the landing-page text and, when available, the "About us" section could be included in their entirety. The API was called using the OpenAI library for python with the

provided API key to first create a client. Then a request could be sent by providing the model, prompt, and desired temperature as arguments. Several temperature values were tested during experimentation, and the temperature of 0.5 was ultimately selected over the default value of 0.7. After testing a range of temperatures from 0-1 the selected temperature built coherent sentences and kept more of the provided information, reducing the risk of hallucinations.

The summarization prompt used in this study consisted of two components: a structured instruction block outlining the required tone, content, and formatting, followed by the raw text extracted from each company's landing page and, when available, its "About us" section. The instruction block ensured that all summaries adhered to a consistent structure and level of detail, while the inclusion of the full extracted text allowed the model to base its output solely on information explicitly provided in the source material. The complete prompt is shown below.

The Summarization Prompt

You are an M&A analyst. Your task is to create a company description from the information given into a concise, neutral and standardized summary. The style should be factual and objective, write it in free text and not as a list.

Instructions:

- Length: 300 words.
- Tone: neutral, objective and professional.
- Content focus: Industry and business model, Core products or services, Geographic focus and main markets, Customer segments or end-users.
- Avoid marketing laguage, exaggerations or subjective adjectives.

Do not infer or invent information not explicitly mentioned in the text.
If information is missing, omit it rather then guessing.

Information below:

{WEB PAGE CONTENT}

3.1.3 Preprocessing and cleaning

As the LLM generated summaries form the primary input to all embedding models, it was important to ensure that they were both accurate and consistent. Although the web scraping pipeline retrieved text from the landing pages and the "About us" sections for most companies, several issues appeared in the raw data that required additional cleaning.

Some webpages contained faulty or redirected links, which led to empty or unusable HTML content. In these cases the LLM could not produce a meaningful summary and instead returned placeholder phrases such as "information not found" or "not specified." To identify such cases in a systematic way, the summaries were examined using a combination of keyword searches and length based filters. An inspection of the distribution of summary lengths showed that nearly all summaries shorter than about 86 words were failed generations. These entries typically corresponded to missing webpage content, non-English source text, or very limited material that the LLM could not expand into a proper summary. Based on this observation, all summaries below the threshold of 86 were removed. A keyword filter was also applied to detect low quality summaries even if they were slightly longer. These keywords or phrases included phrasing like "Not found", "Not specified", and words in foreign languages.

Language inconsistencies created another source of noise. Although the summarization prompt was written in English, pages written entirely in other languages sometimes resulted in short or partially untranslated summaries. These were identified through manual language detection checks and removed in order to maintain a coherent dataset.

After having applied these cleaning steps, including the removal of invalid outputs,

filtering by length, keyword detection, and exclusion of non English summaries, the resulting dataset consisted of high quality and comparable company descriptions. This cleaned corpus serves as the basis for all embeddings. Further preprocessing of the summaries was done individually for the models. This included tokenization, lemmatization, and punctuation removal. As the summaries were generated using a LLM the textual quality was generally very high with few misspellings and special characters.

3.2 Models

After obtaining the textual descriptions of each portfolio company, the next step was to generate numerical vector representations that could be used for similarity analysis. This was done by applying several embedding models through a Python-based pipeline. A dedicated script was developed to load the summarized company texts and organize them into a pandas DataFrame. From there, each model was applied to the corpus to produce a corresponding set of embedding vectors.

Three main types of models were applied: TF-IDF, Doc2Vec models, and several Transformer-based models. All models followed the same data pipeline and storage setup to ensure comparability.

3.2.1 TF-IDF

TF-IDF was used to generate vector representations of the standardized company summaries. The summarized texts were read into Python and processed using the `TfidfVectorizer` from the Scikit-learn library. For the TF-IDF representation, preprocessing was handled directly within the `TfidfVectorizer`. Instead of using an external tokenizer, a custom regular-expression-based token pattern was applied. This pattern ensures that only alphabetic tokens with a minimum length of two characters was included, effectively filtering out numbers, isolated letters, and other non-informative fragments.

The vectorizer was fitted on the full corpus of documents, with the minimum document-frequency parameter min_df tuned prior to finalizing the configuration. We evaluated values in the range $4 \leq min\_df \leq 7$, and selected 5 as it provided an effective balance between vocabulary coverage and noise reduction. Terms appearing in fewer than five documents were therefore excluded from the vocabulary, mitigating the impact of extremely rare words that offer limited discriminative value. The resulting transformation produced a sparse TF–IDF matrix in which each document was represented as a weighted vector reflecting both term frequency and inverse document frequency.

3.2.2 Doc2Vec

Document embeddings were generated using the Doc2Vec implementation from the Gensim library. Each business summary was first preprocessed through lemmatization, where punctuation were removed using spaCy, after which the cleaned tokens were wrapped into Gensim TaggedDocument objects with unique integer identifiers. The Doc2Vec model implemented a Distributed Bag of Words (dm=0), as described in 2.3. The model was set to build 100 dimension vectors, a context window size of 4, a minimum token frequency threshold of 5, and a negative sampling rate of 10. As no numerical target such as an accuracy could be used to set optimal hyper-parameters we started with recommended values and tweaked them based on the resulting distribution. The model object was trained for 5 epochs using `train()`, a reasonable number for a smaller corpus. After vocabulary construction with `build_vocab()`, the model was trained using stochastic gradient descent with Gensim's optimized routines. Final document representations were produced using the model's `infer_vector` method, which applies several gradient descent steps to derive stable embeddings that align with the learned semantic space.

Building on the trained Doc2Vec embeddings, two additional procedures were implemented to enhance and further analyze the resulting document representations. First, Smooth Inverse Frequency (SIF) embeddings were computed to reduce the influence of high-frequency, non-informative terms. This was achieved by re-weighting

token contributions based on inverse frequency and then applying NumPy's SVD function to remove the first principal component, which captures the most most generic information across documents. Second, to investigate the effects of dimensionality reduction on representation quality, Scikit-learn's PCA function was applied to the embedding matrix. The number of components retained was set to preserve 90% of the explained variance, enabling the construction of more compact vectors while maintaining the core informational content of the original Doc2Vec representations.

3.2.3 Transformer Models

Transformer-based embeddings were generated using pretrained models from the SentenceTransformer framework, which includes architectures derived from SBERT as well as variants influenced by GPT-style embedding designs and more. Three prospective models were considered to illustrate the range of Transformer-based embedding approaches: the lightweight "all-MiniLM-L6-v2", the intermediate "all-mpnet-base-v2", and the larger "Qwen/Qwen3-Embedding-0.6B". All models op-

Model Name	**Parameters**	**Output Dim**
MiniLM-L6-v2	22M	384
all-mpnet-base-v2	110M	768
Qwen3-Embedding-0.6B	600M	1024

Table 3.1: Model specifications for Transformer embedding models

erated directly on raw text inputs and incorporated their own tokenization and normalization procedures and therefore no additional preprocessing was applied. Embeddings were obtained using the framework's encode() method. Among these considered models, all-mpnet-base-v2 was selected for the primary analysis as it provided a balance between a good representation and computational cost, and was suitable for environments without access to GPU resources. This model took about 15 minutes to embed all 10 000 samples.

3.3 Implementation

The aim of this study was to design a system that generates suggested buyers for a given target company that was to be sold. When evaluating this system it seemed reasonable to sample target companies from the dataset of portfolio companies and then do a similarity search disregarding the target company from the set of portfolio companies. Given a target company the embeddings could be analysed both in terms of direct similarity search and through buyer suggestion. To perform similarity search the embedding vectors were queried using cosine similarity against the target embedding to find top K most similar companies. The system identified companies to add to the portfolio of the buyer so that they shared similar business models with companies already in their existing portfolio. This assumes that buyers are interested in what they have experience with and know works. To implement this idea for a given target company the system began by computing the pairwise cosine similarity for the target to every other portfolio company. Then the portfolios were grouped by their financial buyer (owner) and only the top three portfolio companies were kept for each buyer. The score for each buyer were then equal to the geometric average of the cosine similarities over these top three holdings. Using these buyer scores the system could then suggest buyers for the given target company based relevant parts of the buyers portfolios.

For a practical implementation the target company was not sampled randomly but provided outside of the dataset. This meant that the target company needed to be scraped, summarized, and embedded separately from the previously embedded portfolio companies. Further, the vector embeddings could be stored in a vector database so that each run only needed to embed a single summary. This put a requirement on the models used as the new target summary needed to be embedded on the same terms as the other companies. For TF-IDF that used corpus word counts in the IDF term this posed a problem. As this was a very fast model it was feasible to simply re-embed all summaries including the new one given that all the summaries can be stored instead of the vectors. Doc2Vec used a neural network to

build the embeddings and was pre-trained for our entire corpus of summaries so it was necessary to store the neural network in a pickle file for example to be used for new summaries. This did however assume that the vocabulary in the summaries were consistent with new ones as the training did not consider these newly added summaries. The Transformer model was not fine-tuned and the weights were pre-trained and imported so this model could be used to embed new target summaries on the same terms as the old ones.

3.4 Evaluation

The models were evaluated both visually and through similarity-based retrieval. Visual inspection allowed for studying how companies were positioned relative to one another in the embedding space, while retrieval tests measured how well each model surfaced similar companies based on cosine similarity. Together, these methods give an indication of whether the embeddings capture business-level similarity.

3.4.1 Visualizations

For evaluation by visualization, dimensionality reduction techniques were used to plot the vectors in 2D. Using UMAP we were able to represent the embedding vectors in 2D to enable scatter plotting. As UMAP works with stochastic initialization it was given a seed for comparability. The UMAP was used via the UMAP-learn library taking parameters of the target dimensionality (=2), the random seed, and some parameters to tweak the resulting distribution.

Taking the full set of companies and plotting them resulted in a very large number of samples giving an overview of clusters. There was no industry codes associated with the data so all that could be done at this level was to analyse the distribution by hovering over samples to check if the sample has reasonable neighbours.

The final visualization examined subsets of companies manually labelled by industry based on their summaries. Five industries were selected to test both similar-

ity and dissimilarity: 'Insurance', 'Asset Management', 'Industrial', 'Realtors', and 'Healthcare'. Insurance and Asset Management were chosen for their financial sector similarities, while Healthcare, Realtors, and Industrial represent materially different operational domains. This selection allowed investigation of whether embeddings capture both clear distinctions between dissimilar sectors and the relationships between related ones.

3.4.2 Similarity Search

To compare companies, similarity scores were computed directly on the embedding vectors generated by the models. For each company that was evaluated, its embedding was first retrieved and then compared with all embeddings in the buyer dataset. The comparison was carried out using cosine similarity, which offers a normalized measure of closeness between vectors and allows for a consistent interpretation of similarity across different embedding types.

The implementation followed a straightforward procedure. Once the target company had been embedded using the chosen model such as TF-IDF or a Transformer-based encoder, its vector representation was compared against every portfolio-company embedding using cosine similarity. Cosine similarity was calculated pairwise, and the resulting scores form the basis for ranking the potential portfolio companies. The companies in the buyer set was sorted according to their similarity scores in descending order, after which the top-k most similar candidates were returned as the system's recommendation.

3.4.3 Expert Evaluation

To evaluate the practical relevance of the buyer recommendations produced by the models, a structured expert review was conducted. The goal of this step was to assess whether the buyer recommendations generated by the system align with the expectations of experienced analysts and to identify failures not captured by numerical metrics comparing the different models presented in the thesis.

First a sample of portfolio companies were selected to cover a mix of industries and business models. For each of these companies, the three different models generated two suggested buyers each. These buyers were identified by first locating the three most similar portfolio companies within each buyer's portfolio and then averaging their cosine similarity scores to form a buyer-level relevance measure. In some cases, the three underlying portfolio companies contributed evenly to the similarity score, while in others a single highly similar portfolio company had a disproportionate influence on the buyer's ranking.

The motivation behind using the three most similar portfolio companies, instead of relying only on the single closest match, was to obtain a more stable and representative measure of buyer relevance. By considering the three closest portfolio companies and taking the average similarity score, the measure better reflects the overall investment profile of the buyer. This method gives a more reliable picture of what the buyer normally invests in, reduces the effect of outliers, and avoids placing too much weight on one unusually strong match. There are several possible approaches to constructing a buyer relevance score, such as using larger sample groups or applying weighted similarity. In dialogue with Merge this setup was chosen because it offered a clear and balanced way of evaluating buyer interest while still being practical to work with.

A sample of 71 target companies was included in the evaluation. To ensure a manageable and consistent review process, only the first paragraph of each target company's summary was provided, as this was deemed sufficient for forming a clear understanding of the business. It is important to note that reviewers were aware that the system gives suggestions purely on textual similarity.

Experts assigned a relevance score to each suggested buyer using a three-point scale:

- **3** – Highly relevant
- **2** – Relevant
- **1** – Not relevant

The expert rankings offer a qualitative assessment of how well each model identifies strategically meaningful buyers. By focusing directly on buyer-level suggestions, rather than individual portfolio companies, the evaluation captures experts' judgement of sector alignment and business model fit. These results serve as the primary qualitative benchmark for comparing the models' performance.

4

Results

This chapter presents the empirical results of the implemented buyer-matching pipeline. We report findings from the summarization stage, embedding performance, dimensionality-reduction visualizations, similarity-search experiments, and the expert evaluation. The results follow the methodological structure outlined in the Method chapter.

4.1 Dataset statistics and quality

The final dataset contains a total of 9,492 company summaries after all preprocessing steps. Each observation corresponds to a cleaned textual summary. Basic descriptive statistics of token lengths for these summaries are presented in Table 4.1.

Statistic	**Value**
Number of documents	9,492
Average number of tokens	199.36
Standard deviation	22.98
Minimum number of tokens	86
Median number of tokens	202
Maximum number of tokens	287

Table 4.1: Statistics for the company summaries

The summaries vary between 86 and 287 tokens, with most values concentrated around 200 tokens. These statistics describe the textual characteristics of the dataset used in the embedding and similarity-search stages.

4.1.1 Quality of LLM Summaries

The summarization step produced one structured summary per company, resulting in 9,492 text outputs. The outputs follow a consistent format extracted from the scraped website content and the examples in Tables 4.2 and 4.3 illustrates the typical structure of the generated summaries in the dataset.

Example Summary - BSI Software
BSI Software is a European company specializing in customer relationship management (CRM) and customer experience (CX) solutions. Founded in 1996 in Switzerland, BSI Software offers the BSI Customer Suite, a modular platform that integrates artificial intelligence to enhance customer engagement, data insights, and relationship management. The company focuses on industries such as banking, insurance, retail, and energy, providing tailored solutions that comply with regulatory standards and meet specific market needs. The BSI Customer Suite is designed to facilitate seamless customer interactions across sales, marketing, and service channels, ensuring data protection and digital sovereignty for its users. The platform includes features such as customer data management, enterprise integration, and AI-driven analytics, allowing businesses to derive actionable insights from their customer data. BSI Software emphasizes flexibility and scalability, enabling clients to adapt the platform to their evolving requirements. Geographically, BSI Software operates primarily in the DACH region (Germany, Austria, and Switzerland) and Italy, serving a diverse range of customer segments from large corporations to smaller enterprises. The company prioritizes customer-centricity and collaboration, fostering a networked approach to project management without traditional hierarchies. BSI Software's commitment to quality and precision is reflected in its extensive industry expertise and its focus on long-term partnerships with clients.

Table 4.2: Example of a generated company summary for BSI Software

Example Summary - Apotea
Apotea is an online pharmacy operating in Sweden, specializing in the sale of pharmaceutical products, health and beauty items, and various wellness solutions. The company offers a wide range of products, including prescription medications, over-the-counter drugs, dietary supplements, and personal care items. Apotea's business model is centered around e-commerce, providing customers with the convenience of shopping for health-related products from home, with options for fast delivery and free shipping. The geographic focus of Apotea is primarily within Sweden, serving customers nationwide. The company caters to diverse customer segments, including individuals seeking health and wellness products for themselves and their families, as well as pet owners looking for veterinary medications and supplies. Apotea also provides professional advice through its licensed pharmacists, ensuring customers receive guidance on their purchases and health inquiries. With an extensive inventory that includes over 50,000 quality-checked products, Apotea positions itself as one of the largest online pharmacies in Sweden. The product categories range from allergy relief and skincare to nutritional supplements and household items. The company emphasizes customer service, offering support via email, phone, and chat, and aims to meet the needs of various consumer demographics, including those with chronic health conditions and specific wellness requirements.

Table 4.3: Example of a generated company summary for Apotea.

The summary in Table 4.2 provides a clear overview of BSI Software's core business areas by identifying CRM and CX solutions as its primary focus with the main offering being the BSI Customer Suite. It also states the company's focus toward regulated service industries which helps situate its target markets. Geographically, the summary specifies that the firm was founded in Switzerland and operates in Europe. In terms of business model, the description implies a modular, AI-enhanced software platform that allows businesses to derive actionable insights from their

customer data. As a result everything desired seems to be mentioned but not at a detailed level.

The summary in Table 4.3 provides a clear overview of Apotea's core operations by identifying its role as a Swedish online pharmacy with an extensive assortment of pharmaceutical, health, and wellness products. It highlights e-commerce as the central business model, emphasizing convenience, fast delivery options, and broad product availability as key value propositions. The summary also situates Apotea geographically by noting its exclusive focus on the Swedish market and its nationwide customer base. In terms of customer segments, the description covers both general consumers seeking health and personal care items as well as pet owners requiring veterinary products. The mention of licensed pharmacists adds context to the company's service offering, suggesting a model that combines digital retail with professional guidance. Overall, the summary captures the main business areas, customer focus, and operational model, though it remains high-level rather than detailing specific logistics or competitive differentiators.

4.2 Models

Table 4.4 reports the approximate runtime for each embedding model when generating representations for the dataset and half of the dataset to compare scaling. TF-IDF runs in a under a second in both cases, while the Doc2Vec variants complete in around two minutes for half the dataset and three minutes for the whole. The MPNet Transformer however requires substantially more time due to its higher model complexity taking ca 16 minutes to train over the full dataset. The runtime for TF-IDF and MPNet Transformrer more than doubles for the full dataset, while for doc2vec the runtime is slightly less than twice the amount. Overall, the results confirm the expected trade-off between model expressiveness and runtime.

	Runtime (seconds)	
Model	**5,000 documents**	**10,000 documents**
TF-IDF	0.3 s	0.8 s
Doc2Vec	102.7 s	192.3 s
Doc2Vec PCA	111.4	199.4 s
Doc2Vec SIF	110.3	198.3 s
MPNet Transformer	410.6	990.2 s

Table 4.4: Runtime statistics

4.3 Embedding Visualization

Using UMAP dimensionality reduction the distribution of companies in the embedding space can be plotted in two dimensions.

4.3.1 All Portfolio Companies

The UMAP plots of all the portfolio companies are shown in Figure 4.1. As the dataset includes no industry codes it is difficult to draw any real conclusions from this data other than some partial cluster formation. It appears that Transformer MPNet and TF-IDF embeds more dense clusters whilst the Doc2Vec models appear more dispersed. TF-IDF seems to have the most outliers and MPNet have some outliers also.

4.3.2 Subset visualization

In order to show how the models embed information about the business and its industry a subset of companies within different sectors where selected. The selection was done manually to find subsets of companies with different and similar business models. The selection was made as 'Insurance', 'Asset Management', 'Industrial', 'Realtors', 'Healthcare'. Some of the companies where consciously selected with vague industries such as Healthcare Insurance, Insurance Brokers, and Asset Man-

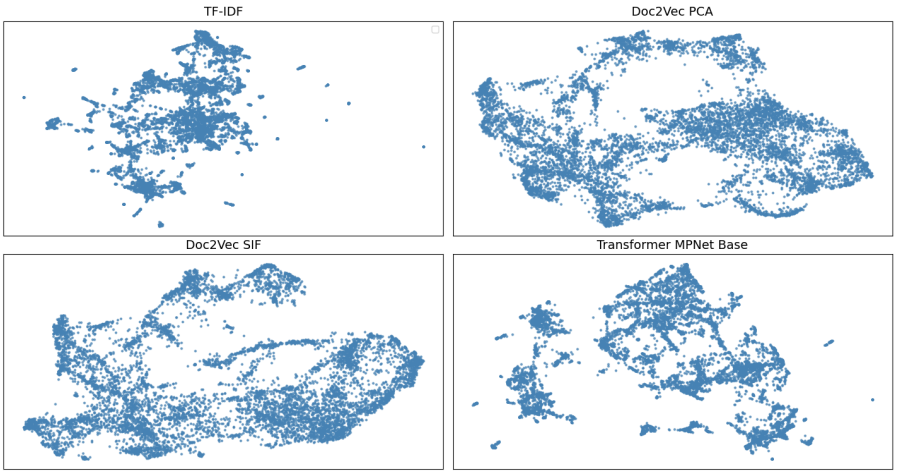

Figure 4.1: UMAP visualization of all companies for the different models

agers with pension insurance operation. The full list of these companies are attached in Appendix A.

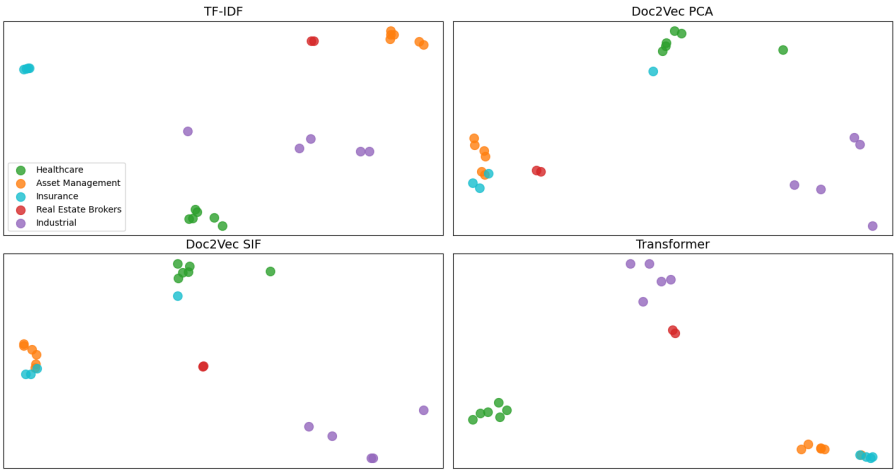

Figure 4.2: UMAP visualization of a subset of companies within 5 industries for the different models

Figure 4.2 shows the UMAP plots for each embedding method, and a few patterns stand out. The industrial companies are somewhat clustered in all models except TF-IDF, where the groups blend more horizontally. The industry names were densely clustered only for Transformer model while in the other appeared more dispersed. Healthcare names cluster tightly across all embeddings, but for Doc2Vec one name stands out as it is more focused on products (Spinal Discs) rather than providing healthcare services or pharmaceuticals. Also both Doc2Vec models embed

a consumer insurance company close to the healthcare names. The asset management cluster is also interesting because most of these companies also do pension insurance or insurance brokerage, and both the Doc2Vec and Transformer models pick up on that by placing them near the insurance cluster, whereas TF-IDF does not capture this relationship as well, placing them closer to the real estate brokers. Overall the Transformer model gives the most separated clusters in terms of the given subset. In Figure 4.3 the subset overlays the plot of all companies to show

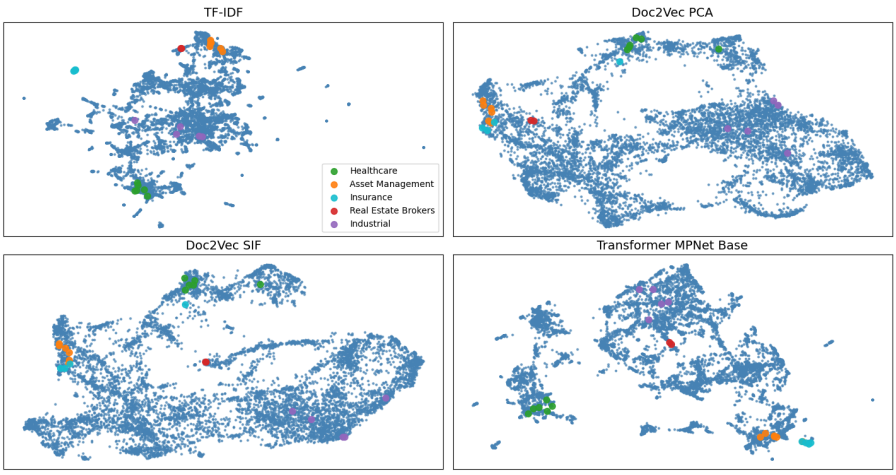

Figure 4.3: UMAP visualization of all companies in the embedding space, with selected companies from five industries highlighted

how the labelled data points conforms to the total structure of the dataset. This gives some context about the cluster formation. The TF-IDF appear to have some clustering but also plenty of outliers where the insurance names appear together in one of these outlier clusters. Both Doc2Vec models seems to place the industrial names in a large dispersed cluster whilst the healthcare names appear in two more dense clusters. The two are very similar where the Real Estate brokers are slightly more separated for SIF and also there seems to be slightly more cluster formation.

4.4 Similarity Search Experiments

This section presents the results of the similarity search experiments, evaluating how effectively each embedding model retrieves relevant companies for a given set of test queries.

4.4.1 Similarity Score Distributions

To better understand how each embedding model represents companies in the vector space, the distribution of cosine similarity scores between all pairs of companies are shown in Table 4.5. These distributions provide insights into how densely or sparsely the models cluster the representations, which in turn influences how sensitive each model is when identifying relevant buyers.

Model	**Min**	**Max**	**Mean**	**Median**
TF–IDF	0.000	0.940	0.127	0.120
Doc2Vec PCA	-0.76685	0.94113	-0.00016	-0.00926
Doc2Vec SIF	-0.652	0.911	-0.00012	-0.00766
MPNet	-0.139	0.994	0.251	0.239

Table 4.5: Statistics of cosine similarity distributions for each embedding model

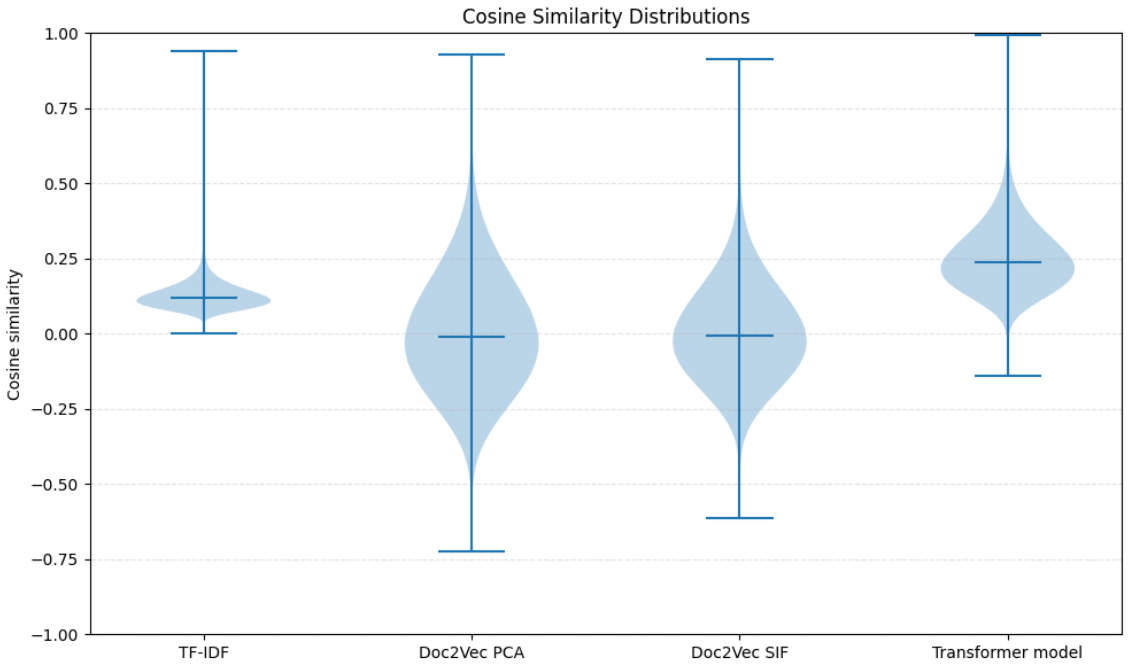

Figure 4.4: Distribution of similarity scores

For the TF-IDF model, the similarity scores are concentrated around low values

with a long tail toward higher similarities. This behaviour is expected, as TF-IDF produces high-dimensional and sparse vectors where most company pairs share few terms. As a result, only companies with strongly overlapping vocabulary achieve high similarity scores, while the majority remain close to zero.

The SIF-Doc2Vec model shows a distribution centered around zero. Unlike TF-IDF, which only contains non-negative values and therefore produces mostly positive similarities, Doc2Vec vectors contain both positive and negative components. After applying SIF (removal of the first singular vector), the embeddings become more isotropic, further pushing cosine similarities toward a normal distribution centered around 0. This results in high contrast between similar and dissimilar companies, but also means that random pairs will have similarity close to zero.

In contrast, the Transformer-based embeddings (SBERT) show a distribution skewed toward higher similarity values compared to TF-IDF. These models place semantically related companies closer in the embedding space, even when textual descriptions do not share explicit vocabulary. The transformer embeddings therefore produce higher baseline similarity scores.

4.4.2 Example of target similarity suggestions

The tables below illustrate an example of the top-k retrieved companies for a selected target, note that the target in this case is a sampled portfolio company. For the chosen target, cosine similarity scores are computed against all portfolio companies in the dataset, and each model returns its highest-ranked matches based on these scores. A larger set of top k results for target companies can be found in Appendix A.

The purpose of the example is to provide a qualitative impression of how the models behave in practice. By examining one representative case, it becomes possible to observe the types of semantic or textual patterns that lead to high similarity scores under TF-IDF, Doc2Vec SIF, and the Transformer-based embeddings. Below is a short description of the sampled target company, Klarna, followed by the top-2 sug-

gestions produced by each model.

Klarna summary

Klarna is a financial technology company specializing in payment solutions for both consumers and businesses. Operating primarily within the e-commerce sector, Klarna provides a variety of payment options designed to enhance the online shopping experience. Its core offerings include immediate payments, deferred payment solutions, and installment plans, enabling consumers to manage purchases according to their financial preferences.

TFIDF Summary	**Similarity**
Paysafe is a global payment solutions provider that offers a range of services designed to facilitate online transactions for businesses and consumers. The company operates within the financial technology industry, focusing on payment processing, digital wallets, and online cash solutions. Its core offerings include card processing, eCommerce solutions, local payment methods, and various digital wallet services such as Skrill, Neteller, and PaysafeCard.	0.518
Qliro is a financial technology company that operates in the payments and savings industry, providing a platform designed to facilitate both online and in-store transactions. The company offers a range of payment solutions, allowing customers to choose their preferred payment methods, including options for immediate payment or flexible payment plans. The Qliro app serves as a comprehensive tool for users to manage their payments and finances, enabling them to track invoices, schedule payments, and communicate with customer support.	0.495

Table 4.6: TFIDF Model Matches for Klarna

The results of TF-IDF top two similarity search is presented in table 4.6. TF-IDF appears to focus on the transaction side of the Klarna business picking two companies providing payment solutions. Both offer a checkout solution for e-commerce which is a big part of the Klarna business. Only Qliro also offers the deferred payment service's making it a more direct competitor to Klarna.

Doc2Vec PCA Summary	**Similarity**
Curve Pay is a digital wallet service that consolidates multiple payment cards into a single, secure platform, enabling users to manage their finances more effectively. The service allows customers to switch between cards even after a purchase, thereby eliminating hidden foreign exchange fees and enhancing the rewards associated with existing bank cards. Curve Pay is designed for both online and in-store transactions, as well as for international spending, offering features such as cashback on purchases and flexible payment options.	0.784
Freecharge is a financial services and payment technology company based in India, operating as a subsidiary of Axis Bank. The company primarily focuses on providing a wide range of payment solutions, including mobile and DTH recharges, utility bill payments, and UPI transactions, catering to over 100 million users across the country. Freecharge's business model integrates various payment methods, allowing users to transact using wallets, UPI, net banking, debit cards, and credit cards, thereby facilitating seamless payments for both online and offline merchants.	0.678

Table 4.7: Doc2Vec PCA Model Matches for Klarna

The PCA reduced Doc2Vec model produces 2 more diverse financial payment companies presented in table 4.7. Curve pay seems in line with Klarna's offering of e-commerce payment solutions but Freeharge appears more focused on general personal finance.

Doc2Vec SIF Summary	**Similarity**
Curve Pay is a digital wallet service that consolidates multiple payment cards into a single, secure platform, enabling users to manage their finances more effectively. The service allows customers to switch between cards even after a purchase, thereby eliminating hidden foreign exchange fees and enhancing the rewards associated with existing bank cards. Curve Pay is designed for both online and in-store transactions, as well as for international spending, offering features such as cashback on purchases and flexible payment options.	0.629
Splitit USA Inc. operates in the financial technology sector, specializing in a buy now, pay later (BNPL) service that allows consumers to split their purchases into smaller monthly payments using their existing credit cards. The company's business model is designed to facilitate flexible payment options without the need for new loans or credit checks, thereby avoiding additional interest or fees.	0.612

Table 4.8: Doc2Vec SIF Model Matches for Klarna

The SIF Doc2Vec model results presented in table 4.8 gain presents Curve Pay but includes a new suggestion. Splitit USA seems to offer buy now pay later services which is a big part of Klarna's offering. The focus on providing consumer credit makes Splitit an interesting suggestion.

Transformer Summary	Similarity
Klar Technologies (Klar) is a regulated financial entity in Mexico, authorized by the Comisión Nacional Bancaria y de Valores (CNBV) and operating under the Law of Savings and Popular Credit. The company primarily focuses on providing financial services, including credit cards, personal loans, and investment accounts. Klar offers a variety of credit card options, such as the Klar Plus and Klar Platino, which feature benefits like no annual fees, cashback rewards, and flexible payment terms.	0.739
Karbon is a fintech company based in Bengaluru, India, specializing in foreign remittance services and corporate expense management solutions. The company primarily focuses on facilitating international payments for businesses, including exporters, importers, freelancers, and direct-to-consumer (D2C) e-commerce enterprises. Karbon's service offerings include a prepaid corporate card designed for expense tracking, an AI Accountant to streamline accounting tasks, and competitive foreign exchange remittance solutions.	0.667

Table 4.9: Transformer Model Matches for Klarna

The MPNet Transformer model results presented in table 4.9 shows two new suggestions. The first suggestion appears to have more of a traditional bank offering of credit cards, loans, and investment accounts. The second option seems to have a broader financial offering more targeted toward B2B.

Although nearly all suggested peers to Klarna operate in the financial sector, their business models show notable variation. Klarna's diversified operations span e-commerce checkout solutions, consumer payment cards, and micro-lending services, creating a complex profile for similarity matching. Notably, multiple models returned pairs of suggestions with comparably high cosine similarity scores, indicating substantial representation of fintech companies within the dataset. However, Klarna's major international competitor Affirm does not appear among the retrieved

peers, as verification confirmed its absence from the portfolio dataset.

4.5 Expert Review Results

In designing the expert evaluation, it was necessary to balance the number of models assessed with the practical constraints of the reviewers' time. As mentioned in the method, each target gave one suggested buyer per model and each of these included three portfolio companies to be considered. Increasing the number of models would therefore have reduced the number of target companies possible to review meaningfully, and a trade-off was required. The SIF Doc2Vec and PCA Doc2Vec showed very similar results in the visualization and cosine similarity search. Given the minor benefits of SIF Doc2Vec model variant this was selected as the third model together with MPNet Transformer and TF-IDF for the expert evaluation.

Table 4.10 shows the results of the expert evaluation with a total of 71 samples evaluated by professionals from Merge. Note that each model outputs two examples and so the total count for each model is therefore 142. The full evaluation set with scores per target can be found in Appendix B.

Model	**Count 1**	**Count 2**	**Count 3**	**Average**	**Median**
Doc2Vec (SIF)	36	52	54	2.13	2
TF–IDF	16	43	83	2.47	3
Transformer (MPNet)	3	43	96	2.65	3

Table 4.10: Expert ratings for each model. Scores range from 1 (not relevant) to 3 (highly relevant)

5

Discussion

The purpose of this thesis was to investigate whether textual embedding models can support the identification of relevant potential buyers in M&A processes. The central question was whether different embedding approaches create vector representations that meaningfully reflect business similarities between companies, and whether those representations can be used to recommend buyers whose portfolio profiles align with a given target. To address this, the study relied on generating a large dataset of scraped and LLM-generated summaries describing the operations, value propositions, and geographic footprints of portfolio companies. Across both quantitative analyses and expert evaluation, the results indicate that the choice of embedding representation significantly influences the quality and relevance of the recommendations. In particular, the Transformer model demonstrated the most coherent clustering and achieved the strongest alignment with expert judgement.

5.1 Discussion of Results

This chapter interprets and reflects on the results presented in the previous section. The aim is to evaluate how well the proposed methodology addresses the problem of buyer-seller matching in M&A, assess the relative performance of the three embedding models, and relate the findings back to the overarching research questions and practical context at Merge.

5.1.1 Embedding space and Visualizations

A central objective of this thesis was to explore how companies can be embedded in a vector space in a way that reflects their underlying business characteristics. Since all embeddings were genereted from LLM-produced summaries of publicly available website content, the quality and content of this data directly shaped the geometry of each embedding space. The UMAP visualisations presented in Figure 4.1-4.3 provide an interpretable view of these high dimensional structures and give an indication of how well the different models captures the semantics and content of the summaries.

Although the models differ in how their embeddings are constructed, they also differ substantially in dimensionality, which influences the structure of the resulting embedding spaces. TF-IDF produces very high-dimensional sparse vectors (in our case 8000) and Doc2Vec creates dense vectors with substantially lower dimensionality with a standard length of 100. The MPNet Transformer model lies between these extremes, generating 768-dimensional dense semantic embeddings. These dimensional differences affect how much variation each model can encode and how tightly companies can be positioned relative to one another. TF-IDF representations tend to create separations based on vocabulary patterns, whereas lower-dimensional dense vectors may compress information and produce smoother boundaries. To visualise these high-dimensional structures, UMAP was applied to project each embedding space into two dimensions. UMAP preserves local neighbourhoods while maintaining aspects of global organisation. Although some information is inevitably lost, meaningful and coherent clusters in the two-dimensional projections indicate that the original embeddings capture relevant business similarities.

Figure 4.1 shows how the different models embed all the data in the embedding space, showing clear differences in the structure between the models. Both the TF-IDF and MPNet transfomer embeddings show more distincitve cluster seperation, suggesting that these models capture more consistent patterns in how companies

relate to one another. In contrast, both Doc2Vec variants generate more evenly spread shapes, indicating that their representations are less sharply defined and may struggle to separate companies clearly. The stronger cluster formation observed for TF-IDF and especially for the Transformer model suggests that these embeddings retain more distinctive information.

The subset visualizations in Figure 4.2 provides an example of how well the models seperate companies within different sectors. In this example, the differences between the models become clearer. TF-IDF manages to separate the selected sectors relatively well, with insurance, healthcare, real estate, and asset management forming distinct clusters, while the industrial companies appear spread out on a line. This pattern can be expected, as industrial firms often span a wide range of activities and therefore have a wide range of textual information. Both Doc2Vec models show a similar overall structure, producing clear clusters for real estate, industrial, and healthcare companies, but asset management overlaps with insurance. Additionally one insurance company is positioned closer to healthcare. In this example the MPNet Transformer model clearly stands out when it comes to performance. Each sector is clearly clustered and additionally the relative positions of the clusters stands out compared to the other models. The insurance and asset management companies appear close to one another. This aligns with their shared financial orientation, and the fact that assets managers often offer pension insurance. Further, real estate ends up between industrial and the finance sectors and healtcare ends up a in a corner of its own.

Taken together, the findings indicate that publicly available text, when processed through LLM-generated summaries and encoded using embeddings techniques, can serve as a meaningful basis for representing companies in operational terms. Dimensionality reduction techniques such as UMAP preserve much of this structure and offer an intuitive way to inspect and validate these embeddings. While the precision of the representations depends heavily on the quality of the underlying textual data, the overall evidence indicates that embedding based representations

can capture relevant business similarities. The relative clarity and separation observed in the Transformer embeddings further support the use of Transformer models for applications such as buyer seller matching, where identifying subtle strategic or operational overlaps is essential.

5.1.2 Similarity Search and Buyer Recommendation

The study also explored using the trained embeddings to find similar companies and also finding suitable buyers for target companies. The similarity search gives insight into how a recommendation system can use the embeddings to make suggestions and what kinds of patterns the embeddings actually encode. In the results Klarna was used as an example where all models found similar financial companies. TF-IDF returned companies with explicit references to payment processing and checkout solutions, reflecting its reliance on textual overlap. The Doc2Vec models, both PCA reduced and SIF, identified firms sharing functional characteristics such as digital wallets or BNPL services, indicating it makes semantic connections between related concepts. The Transformer embeddings highlights companies with a broader financial-service offerings, suggesting a focus on contextual similarities. These distinctions illustrate how each representation emphasizes different nuances of a firm's business profile. This example gives some insight into what aspects of the business model are encoded into the embeddings. However this is only one example and cannot alone be used to determine which models performs best or which dimenions of the company that are encoded for the different models. More examples of similarity search are included in Appendix A.

Several factors create a nuance to the interpretation of these results. Companies with multiple operational segments may present a challenge for similarity search, as their broad scope may not be fully captured through information gathered from their website. This issue that is especially relevant for large firms like Klarna, whose many activities span payments services, consumer credit and technological solutions. In addition, the generated suggestions are inevitably limited by the companies of the portfolio dataset. A corpus dominated by tech investments naturally increases

the likelihood of retrieving similar firms for targets like Klarna, but limits the identification of peers for a target outside this investment profile.

The expert review results presented in Table 4.10 show clear differences in performance across the models. Doc2Vec (SIF) received the lowest average and median scores and the highest number of low-relevance ratings (1s), suggesting that it generally struggled to identify suitable buyers in this setting. TF–IDF performed noticeably better, producing more highly relevant suggestions and fewer clearly unsuitable ones. MPNet achieved the most consistent performance, with the fewest poor ratings and the highest average and median relevance. These patterns suggest that Transformer-based embeddings are better aligned with expert judgments of suitable buyers, whereas simpler models like Doc2Vec may fail to capture the characteristics that experts consider important for buyer matching.

The results highlight the practical trade-offs in using embedding-based recommendations. MPNet's consistent outputs make it more reliable for systematic recommendation workflows, whereas TF–IDF offers a simpler alternative that can still identify relevant peers but with slightly less precision. Further, TF-IDF has a significantly faster runtime than the rest of the models, indicating it is suitable for time constrained applications. The longer runtime of the Transformer suggests that it more useful in applications where time is not an issue and the goals is to the best accuracy possible.

5.2 Limitations

This chapter reflects on the limitations of this study presented in the introduction. The purpose is to highlight design choices made regarding the methods and scope and how this may have affected the results of the study.

5.2.1 Impact of LLM Summarization

A central design choice in the thesis was to base all models on LLM generated summaries rather than raw website text. The motivation was to obtain a more uniform and comparable representation of companies, both in terms of length and structure. By enforcing a common prompt and requiring the model to describe aspects such as sector, business model, and customer segments, the summaries provided a standardized input representation that is better suited for large scale similarity search than heterogeneous website content with varying tone, formatting, and level of detail.

However, the reliance on LLM summarization could introduce sources of uncertainty that may influence the model outputs. The summaries are not direct compressions of the website text, but model based interpretations, which means that certain details may be omitted or receive disproportionate emphasis. In some cases, important characteristics could be underrepresented, while generic vocabulary and broad descriptions might appear more often than in the original content. For example, a clothing brand with extensive sustainability related marketing could receive a summary that places more weight on sustainability than is representative of its core business. This could increase the risk of matching it with buyers focused on sustainability oriented portfolios. Similar effects could occur when a specialized company is summarized in very general terms such as “technology solutions provider” or “consulting services,” potentially reducing the ability of the embedding models to differentiate between companies operating in different niches.

The use of LLM generated summaries nevertheless reduced substantial noise from raw website data and created a more consistent input space. This uniformity allowed the models to focus on conceptual content rather than website layout or stylistic differences. Still, the presence of overly generic or underspecified summaries influenced how clearly the embeddings could separate industries or business models. The Transformer based embeddings appeared more resilient to this issue, showing better cluster separation in the visualizations.

5.2.2 Limitations in Scraped Data

A substantial limitation of the study arises from the incompleteness of the scraped dataset. A large number of portfolio companies could not be scraped at all, often due to non English websites, dynamic content, or complex site structures that prevented reliable extraction of portfolio pages. As a result, the dataset underrepresents the true set of companies owned by several buyers, particularly international buyers. This reduces the diversity and completeness of the buyer portfolios and may affect matching performance for certain buyers.

Beyond missing companies, the quality of information available on scraped pages varied considerably. Some portfolio companies, especially smaller holdings, provided very limited "About us" information or relied heavily on marketing oriented landing pages that contained little structural or operational detail. In such cases, the LLM had limited material from which to generate a meaningful summary, increasing the risk of vague, overly promotional or incomplete descriptions. Systematic cleaning was applied to remove clearly faulty summaries, such as those containing phrases like "information not found" or "not specified," or those below a minimum length threshold. However, with more than ten thousand pages, manual verification was not feasible, and it is possible that some long summaries still contained irrelevant marketing language or partial misinformation.

While such cases introduce noise into the embedding space, their practical impact on matching may be limited. Portfolio companies with low informational quality tend not to become close neighbours to any target and therefore rarely appear among the top ranked buyers. In practice, these weak summaries typically fail to match strongly with any company, reducing but not entirely eliminating the risk of misleading recommendations. Nonetheless, the variability in scraped data quality represents an inherent limitation of using web derived company descriptions at scale.

5.2.3 Limitations in Method

The reliance on unsupervised similarity scoring as a proxy for buyer fit introduces several inherent limitations to this study. While the absence of labelled historical transaction data resulted in this approach, it constrains the model's ability to capture the full complexity of acquisition decision making. The similarity based method operates under the assumption that buyers primarily seek targets resembling their existing portfolio companies, thereby overlooking other acquisition patterns driven by diversification intent, emerging market themes, capability gaps, or exploratory investments in adjacent domains. A supervised learning approach trained on historical transaction data could have identified these nuanced patterns by learning from actual buyer behaviour, potentially uncovering non obvious relationships between buyer characteristics and acquisition targets. Furthermore, the exclusion of financial metrics and deal-specific features limits the interpretability of results, as the model cannot account for valuation considerations, financial synergies, or other quantitative factors that significantly influence real-world transaction decisions.

The evaluation in this study is constrained by resources and time limitations. While the analysis provides insights based on the target companies examined, a more comprehensive evaluation across a larger and more diverse sample of targets would be necessary to establish complete certainty regarding the generalizability of findings. The resource requirements to evaluate all possible model combinations and parameter configurations exceeded available resources, requiring a more limited evaluation than would be ideal for certain conclusions. Having labelled data could also be an approach to allow for large scale evaluation.

5.3 Recommendations for Future Work

This study faced constraints related to both the scope of the company dataset and the absence of labeled data for supervised validation. While expanding the number of companies in the dataset would likely give slight improvements in coverage

and retrieval quality, a more substantial improvement would come from access to a large labelled dataset linking companies to their industries or transaction outcomes. Such labelled data would enable better hyperparameter tuning and provide objective benchmarks for model performance evaluation. Additionally, with sufficient labelled examples, models could be trained in a supervised way to predict company characteristics such as industry classification directly from embeddings, potentially capturing patterns that unsupervised similarity methods cannot detect. Future work might also explore approaches using models trained on historical acquisition data to learn more complex patterns underlying acquisition decisions beyond portfolio similarity.

Further, as the scraped dataset is based on a wide range of buyers, the area of operations of the portfolio companies are spread out all over the world. Often, when receiving a target company it might not be suitable to consider all geographies, therefore an approach could be to incorporate hard filtering. This way the user of the recommendation system could limit the search to portfolio companies based in a specific country or area and thus get matches based on this. There is no trivial way to get a reliable dataset for the geographies of all portfolio companies. Future work could focus on extracting geographic information from websites or using named-entity recognition to identify locations in the text. Even a partially accurate location layer would allow initial geographic filtering and make the recommendations more practical in real M&A workflows.

5.4 Conclusion

This thesis examined whether textual embedding models can help identify relevant buyers in M&A by capturing business similarities between companies. The results show that embeddings built from standardized LLM-generated summaries contain enough structured information to support meaningful similarity search and to form the basis of buyer recommendations. The evaluation revealed that the Transformer model produced the most coherent embedding space, showed the strongest clus-

tering behaviour, and delivered buyer suggestions that experts found most relevant. TF–IDF also produced several good matches but with less consistency. The Doc2Vec models were less able to separate different types of companies and generally performed weakest in the expert review.

These findings suggest that Transformer-based embeddings are well suited for assisting the early stages of buyer discovery. They can efficiently highlight potential buyers whose portfolios resemble the target company, helping reduce manual screening and broaden the initial search. At the same time, the limitations of the approach must be recognised. The models depend on the quality of website data and LLM summaries, and the unsupervised similarity scoring used in this study captures only a small part of real buyer behaviour. Important factors such as financial metrics, strategic motives, and diversification considerations remain outside the scope of the embeddings.

In summary, embedding-based methods, especially Transformer models, can provide useful recommendations in M&A processes. They should not be viewed as a replacement for expert judgement but as a tool that can support and speed up early analysis by offering useful suggestions.

Bibliography

S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SyK00v5xx.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

A. Holtzman, J. Buys, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019. URL http://arxiv.org/abs/1904.09751.

K. S. Kalyan, A. Rajasekharan, and S. Sangeetha. Ammus : A survey of transformer-based pretrained models in natural language processing, 2021. URL https://arxiv.org/abs/2108.05542.

Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014. URL http://arxiv.org/abs/1405.4053.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Merge, 2025. Personal communication, 2 September 2025.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

OpenAI. Api-pricing. https://openai.com/svSE/api/pricing/, 2025. Accessed: 2025-11-12.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL http://arxiv.org/abs/1908.10084.

M. Ringnér. What is principal component analysis? *Nature Biotechnology*, 26 (3):303–304, 2008. doi: 10.1038/nbt0308-303. URL https://doi.org/10.1038/nbt0308-303.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: https://doi.org/10.1016/0306-4573(88)90021-0. URL https://www.sciencedirect.com/science/article/pii/0306457388900210.

J. Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014. URL http://arxiv.org/abs/1404.1100.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355/.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

A

Appendix A: Competitor Similarity Results

This appendix shows top three nearest neighbours in the embedding space for a list of target companies. The list consists of the subset of companies discussed in chapter 4.2.2. This represents the three portfolio companies that have the most similar embeddings.

Table A.1: Top-3 Similar Companies for http://www.partnerre.com (Insurance)

Model	**Similar Company**	**Similarity**
TF-IDF	http://www.archcapgroup.com	0.54
	https://www.bmsgroup.com	0.52
	https://premiaholdings.com	0.45
PCA Doc2Vec	https://www.bmsgroup.com	0.75
	https://www.balanceuw.com	0.70
	https://specialistrisk.com	0.65
SIF Doc2Vec	https://www.bmsgroup.com	0.66
	https://www.balanceuw.com	0.56
	https://www.optiogroup.com	0.56
Transformer	https://www.balanceuw.com	0.74
	https://flemingih.com	0.74
	https://foundationrp.com	0.73

Table A.2: Top-3 Similar Companies for http://www.maxm.se (Insurance)

Model	Similar Company	Similarity
TF-IDF	http://www.sakra.se	0.49
	http://www.avidinsurance.co.uk	0.48
	https://www.alperseguros.com.br	0.46
PCA Doc2Vec	http://www.sakra.se	0.71
	https://www.risk-strategies.com	0.65
	https://www.mrh-trowe.com	0.65
SIF Doc2Vec	https://www.soderbergpartners.se	0.66
	http://www.sakra.se	0.65
	https://www.newportgroup.com	0.56
Transformer	https://www.mrh-trowe.com	0.72
	https://www.capitalgroup.com	0.71
	https://www.mandatum.fi	0.70

Table A.3: Top-3 Similar Companies for https://www.hedvig.com (Insurance)

Model	Similar Company	Similarity
TF-IDF	http://www.hedvig.com	0.80
	https://www.acko.com	0.61
	https://sureapp.com	0.60
PCA Doc2Vec	http://www.hedvig.com	0.84
	https://www.lassie.co	0.75
	https://www.acko.com	0.63
SIF Doc2Vec	http://www.hedvig.com	0.80
	https://www.lassie.co	0.67
	https://www.acko.com	0.56
Transformer	http://www.hedvig.com	0.98

Model	Similar Company	Similarity
	https://ludvig.se	0.61
	http://getsafe.de	0.60

Table A.4: Top-3 Similar Companies for https://www.epicbrokers.com (Insurance)

Model	Similar Company	Similarity
TF-IDF	https://www.risk-strategies.com	0.52
	https://foundationrp.com	0.49
	https://honan.com.au	0.42
PCA Doc2Vec	https://www.bishopfleming.co.uk	0.69
	https://www.bmsgroup.com	0.69
	http://www.sumer.co.uk	0.68
SIF Doc2Vec	https://www.risk-strategies.com	0.61
	http://www.sumer.co.uk	0.61
	https://www.bishopfleming.co.uk	0.58
Transformer	https://www.embarkmga.com	0.69
	https://www.worldinsurance.com	0.68
	https://surer.sg	0.67

Table A.5: Top-3 Similar Companies for https://www.brookfield.com (Asset Management)

Model	Similar Company	Similarity
TF-IDF	https://www.ardian.com	0.51
	https://www.oaktreereif.com	0.50
	https://www.amundi.fr	0.45
PCA Doc2Vec	http://www.gcmlp.com	0.76

Model	Similar Company	Similarity
	https://www.bcapgroup.com	0.71
	https://firstreserve.sharesecurely.com	0.70
	http://www.gcmlp.com	0.60
SIF Doc2Vec	https://firstreserve.sharesecurely.com	0.57
	https://apolloacademy.com	0.55
	https://www.oaktreereif.com	0.75
Transformer	https://wilshire.com	0.71
	https://powersustainable.com	0.70

Table A.6: Top-3 Similar Companies for http://cworldwide.com (Asset Management)

Model	Similar Company	Similarity
	https://www.waystone.com	0.37
TF-IDF	https://www.efront.com	0.37
	https://www.thinkproject.com	0.36
	http://www.artisanfunds.com	0.68
PCA Doc2Vec	https://www.impact-expansion.com	0.66
	https://ambientasgr.com	0.66
	https://www.caledonia.com	0.57
SIF Doc2Vec	https://www.velcan.lu	0.54
	https://www.baincapitalcredit.com	0.53
	https://www.cvc.com	0.70
Transformer	https://www.orlando-management.com	0.67
	https://www.21invest.com	0.63

Table A.7: Top-3 Similar Companies for https://www.oaktreesicav.com (Asset Management)

Model	Similar Company	Similarity
TF-IDF	https://fisherfunds.co.nz	0.55
	http://www.fsfund.com	0.54
	http://www.vcm.com	0.54
PCA Doc2Vec	https://greshamhouse.com	0.80
	https://www.rielloinvestimenti.it	0.77
	https://www.deacapital.com	0.76
SIF Doc2Vec	https://greshamhouse.com	0.72
	https://www.deacapital.com	0.64
	https://deacapitalaf.segnalazioni.net	0.61
Transformer	https://www.oaktreereif.com	0.76
	http://www.lingotto.com	0.75
	https://www.aeo-se.de	0.74

Table A.8: Top-3 Similar Companies for https://www.spiltanfonder.se (Asset Management)

Model	Similar Company	Similarity
TF-IDF	https://fisherfunds.co.nz	0.56
	http://www.ultimusfundsolutions.com	0.56
	http://www.fsfund.com	0.55
PCA Doc2Vec	https://www.holberg.no	0.64
	https://inveready.com	0.64
	http://www.maxm.se	0.62
SIF Doc2Vec	https://www.holberg.no	0.64
	https://www.easyvest.be	0.58

Model	Similar Company	Similarity
	https://formuepleje.dk	0.57
Transformer	https://www.capitalgroup.com	0.68
	https://formuepleje.dk	0.66
	https://atle.se	0.66

Table A.9: Top-3 Similar Companies for https://www.mandatum.fi (Asset Management)

Model	Similar Company	Similarity
TF-IDF	http://www.sakra.se	0.41
	https://www.soderbergpartners.se	0.38
	http://www.edelmanfinancialengines.com	0.38
PCA Doc2Vec	http://www.sakra.se	0.73
	https://www.fairstone.co.uk	0.68
	https://www.wealthenhancement.com	0.68
SIF Doc2Vec	http://www.sakra.se	0.68
	https://www.afhwm.co.uk	0.61
	https://mcfadvisors.com	0.61
Transformer	https://www.mrh-trowe.com	0.71
	http://www.maxm.se	0.70
	https://www.capitalgroup.com	0.68

Table A.10: Top-3 Similar Companies for https://www.soderbergpartners.se (Asset Management)

Model	Similar Company	Similarity
TF-IDF	http://www.sakra.se	0.62

Model	Similar Company	Similarity
	https://savantwealth.com	0.47
	https://www.fairstone.co.uk	0.47
	http://www.sakra.se	0.70
PCA Doc2Vec	https://www.advantawealth.co.uk	0.67
	https://www.clearstead.com	0.66
	http://www.sakra.se	0.67
SIF Doc2Vec	http://www.maxm.se	0.66
	https://www.mandatum.fi	0.60
	https://www.capitalgroup.com	0.70
Transformer	http://www.maxm.se	0.67
	https://formuepleje.dk	0.66

Table A.11: Top-3 Similar Companies for https://sjukhus.sophiahemmet.se (Healthcare)

Model	Similar Company	Similarity
	https://platform24.com	0.43
TF-IDF	https://groupenovamed.com	0.41
	https://www.doctify.co.uk	0.39
	https://www.med-on-mvz.de	0.72
PCA Doc2Vec	https://www.london-gynaecology.com	0.70
	https://avimedical.com	0.70
	https://www.med-on-mvz.de	0.61
SIF Doc2Vec	https://www.london-gynaecology.com	0.59
	https://www.medicum.ee	0.56
	https://www.varden.se	0.60
Transformer	https://www.bergmanclinics.nl	0.58

Model	Similar Company	Similarity
	https://www.mindoktor.se	0.58

Table A.12: Top-3 Similar Companies for https://www.landmarkhealth.org (Healthcare)

Model	Similar Company	Similarity
TF-IDF	https://www.marathon-health.com	0.52
	https://www.mediq.nl	0.49
	https://andhealth.com	0.47
PCA Doc2Vec	http://compassus.com	0.77
	https://www.centerwellprimarycare.com	0.75
	https://www.imamedicalgroup.com	0.73
SIF Doc2Vec	http://compassus.com	0.72
	https://www.centerwellprimarycare.com	0.69
	https://www.advita.de	0.66
Transformer	https://www.centerwellprimarycare.com	0.74
	https://welbehealth.com	0.72
	https://eventuswholehealth.com	0.72

Table A.13: Top-3 Similar Companies for http://www.vamed-care.com (Healthcare)

Model	Similar Company	Similarity
TF-IDF	https://www.medicum.ee	0.44
	https://www.gemma.lt	0.43
	https://oyora.de	0.41
PCA Doc2Vec	https://pentahospitals.com	0.76

Model	Similar Company	Similarity
	https://sanecum.de	0.74
	https://www.gruppoanimalia.it	0.70
	https://www.gruppoanimalia.it	0.64
SIF Doc2Vec	https://pentahospitals.com	0.64
	https://www.corius.de	0.59
	https://www.med-on-mvz.de	0.69
Transformer	https://careflexzorggroep.nl	0.69
	https://pentahospitals.com	0.67

Table A.14: Top-3 Similar Companies for https://www.highridgemedical.com (Healthcare)

Model	Similar Company	Similarity
	http://www.spineart.com	0.53
TF-IDF	https://spine-innovations.com	0.41
	https://www.onwd.com	0.40
	https://www.merillife.com	0.71
PCA Doc2Vec	http://www.ndd.ch	0.69
	https://www.clinicaamo.com.br	0.65
	https://www.clinicaamo.com.br	0.62
SIF Doc2Vec	https://www.merillife.com	0.62
	https://dorcglobal.com	0.57
	https://spine-innovations.com	0.75
Transformer	http://www.spineart.com	0.75
	http://www.mypainsolution.com	0.66

Table A.15: Top-3 Similar Companies for http://www.reliant-rehab.com (Healthcare)

Model	Similar Company	Similarity
TF-IDF	https://www.rehaneo.de	0.40
	https://www.gemma.lt	0.39
	https://quicklypro.it	0.37
PCA Doc2Vec	https://www.usacs.com	0.74
	http://www.interdent.com	0.74
	http://thesteppingstonesgroup.com	0.73
SIF Doc2Vec	http://thesteppingstonesgroup.com	0.63
	http://www.elara.com	0.63
	https://www.aoncology.com	0.60
Transformer	https://www.rehaneo.de	0.80
	https://ivyrehab.com	0.79
	https://www.numotion.com	0.72

Table A.16: Top-3 Similar Companies for https://www.cloverhealth.com (Healthcare)

Model	Similar Company	Similarity
TF-IDF	https://www.getduos.com	0.45
	https://gohealth.com	0.43
	https://www.marathon-health.com	0.41
PCA Doc2Vec	http://www.inmedica.lt	0.66
	https://www.bergmanclinics.nl	0.66
	https://www.libertydentalplan.com	0.66
SIF Doc2Vec	https://www.london-gynaecology.com	0.56
	https://www.bergmanclinics.nl	0.56

Model	Similar Company	Similarity
	http://www.compbenefits.com	0.54
Transformer	https://includedhealth.com	0.60
	http://www.greenwayhealth.com	0.60
	https://quantum-health.com	0.60

Table A.17: Top-3 Similar Companies for https://www.valimmobilier.ch (Real Estate Brokers)

Model	Similar Company	Similarity
TF-IDF	http://www.iadfrance.fr	0.52
	http://www.diakrit.com	0.46
	https://casavo.com	0.46
PCA Doc2Vec	https://bskimmobilier.com	0.72
	http://www.iadfrance.fr	0.69
	https://casavo.com	0.60
SIF Doc2Vec	https://bskimmobilier.com	0.60
	http://www.iadfrance.fr	0.59
	http://www.arthur-loyd.com	0.53
Transformer	http://www.iadfrance.fr	0.66
	https://www.svefa.se	0.63
	https://hosman.co	0.62

Table A.18: Top-3 Similar Companies for https://bskimmobilier.com (Real Estate Brokers)

Model	Similar Company	Similarity
TF-IDF	https://laposteimmobilier.fr	0.55

Model	Similar Company	Similarity
	https://www.bouygues-immobilier-corporate.com	0.54
	http://www.iadfrance.fr	0.53
	http://www.iadfrance.fr	0.79
PCA Doc2Vec	https://www.valimmobilier.ch	0.72
	https://www.laddercapital.com	0.62
	http://www.iadfrance.fr	0.68
SIF Doc2Vec	https://www.valimmobilier.ch	0.60
	http://www.arthur-loyd.com	0.58
	https://www.bouygues-immobilier-corporate.com	0.70
Transformer	http://www.iadfrance.fr	0.66
	https://laposteimmobilier.fr	0.64

Table A.19: Top-3 Similar Companies for https://www.renson.fr (Industrial)

Model	Similar Company	Similarity
	https://www.selectwater.com	0.41
TF-IDF	https://crbwater.com	0.39
	http://norwesco.com	0.38
	https://emmeti.com	0.68
PCA Doc2Vec	https://www.sigmaco.com	0.66
	https://www.ecoclim.net	0.65
	https://emmeti.com	0.63
SIF Doc2Vec	https://cutnordic.com	0.62
	http://norwesco.com	0.59
	https://rensman.se	0.72
Transformer	https://genieflexion.com	0.65
	https://www.dsl.fr	0.64

Table A.20: Top-3 Similar Companies for https://azekco.com (Industrial)

Model	Similar Company	Similarity
TF-IDF	https://barretteoutdoorliving.ca	0.37
	http://buitenhof-tuinmeubelen.nl	0.37
	http://hartman.nl	0.36
PCA Doc2Vec	http://edelcarpets.com	0.70
	https://www.ricchetti-group.com	0.65
	https://www.vaggmaterial.se	0.64
SIF Doc2Vec	https://www.ricchetti-group.com	0.57
	http://edelcarpets.com	0.57
	https://www.royalrobbins.com	0.55
Transformer	http://www.zytechtruss.com	0.65
	https://www.vhzgroep.nl	0.65
	https://www.flexiteek.com	0.60

Table A.21: Top-3 Similar Companies for https://www.globalppi.com (Industrial)

Model	Similar Company	Similarity
TF-IDF	https://precision-werks-group.de	0.43
	http://www.bostoncenterless.com	0.42
	https://www.malouvis.com	0.41
PCA Doc2Vec	http://www.dukeempirical.com	0.69
	https://www.tegeno.pl	0.68
	http://dumurindustries.com	0.68
SIF Doc2Vec	https://www.tegeno.pl	0.63
	http://dumurindustries.com	0.59
	http://www.dukeempirical.com	0.57
Transformer	https://precision-werks-group.de	0.81

Model	Similar Company	Similarity
	https://www.asindus.com	0.79
	https://www.ptf-group.com	0.77

Table A.22: Top-3 Similar Companies for https://www.hubs.com (Industrial)

Model	Similar Company	Similarity
TF-IDF	https://geomiq.com	0.51
	https://www.gtk.co.uk	0.36
	https://www.asindus.com	0.35
PCA Doc2Vec	https://www.tecnotion.com	0.70
	https://geomiq.com	0.67
	https://www.emeraldtechnologies.com	0.66
SIF Doc2Vec	https://www.tecnotion.com	0.64
	https://geomiq.com	0.60
	https://elcee.com	0.55
Transformer	http://www.sybridgetech.com	0.72
	https://geomiq.com	0.72
	https://www.kickmaker.fr	0.70

Table A.23: Top-3 Similar Companies for https://www.rotomon.fi (Industrial)

Model	Similar Company	Similarity
TF-IDF	http://www.armtec.com	0.47
	https://www.usalco.com	0.37
	https://www.evoqua.com	0.37
PCA Doc2Vec	https://www.novatek.no	0.69
	https://tecomatic.com	0.67

Model	**Similar Company**	**Similarity**
	https://ewers.de	0.65
SIF Doc2Vec	https://tecomatic.com	0.61
	https://www.novatek.no	0.60
	https://www.anolytech.com	0.59
Transformer	https://swehydro.se	0.73
	http://www.flowplus.fi	0.70
	https://va-gruppen.se	0.70

B

Appendix B: Expert Evaluation Summary

This Appendix show the expert valuation scores. The numeric values represent the average score across the two suggested buyers per target and embedding model.

Target	TF-IDF	Doc2Vec	Transformer
https://www.zizzi.de	3	2	2
https://www.worldaware.com	3	3	3
https://www.wastequip.com	2.5	2	3
https://www.vatterledenlogistik.se	2	1.5	2
https://www.tops-products.com	1	1	1.5
https://www.tnp.net.uk	2.5	2	3
https://www.tier.app	2	2	2.5
https://www.tidningskungen.se	2	1.5	2
https://www.stegra.com	2.5	2.5	2.5
https://www.shawgibbs.com	3	3	3
https://www.sentech.nl	2	2	3
https://www.sensome.com	2.5	2.5	3
https://www.sdclgroup.com	2.5	2	3
https://www.satelliteparis-boutique.com	2.5	1.5	2.5
https://www.qima.com	2	3	3
https://www.payflows.io	2.5	3	3

Table B.1: Expert evaluation scores (1/3)

Target	TF-IDF	Doc2Vec	Transformer
https://www.pathstone.com	3	3	3
https://www.ohalo.co	3	2.5	3
https://www.merlincycles.com	3	1.5	3
https://www.ledil.com	2	1	2.5
https://www.lafeemaraboutee.fr	2	3	2.5
https://www.jdepeets.com	2.5	1	2.5
https://www.inatech.com	2	2.5	2.5
https://www.impresoftgroup.com	2.5	3	2.5
https://www.imaginelearning.com	3	3	3
https://www.iii.com	2.5	1	2.5
https://www.ictparts.be	2	2	2.5
https://www.fresha.com	3	2	2.5
https://www.foodtel.com	2	1.5	2.5
https://www.farfield.net	2.5	2.5	2
https://www.embarkmga.com	3	2	2.5
https://www.elitees.com	2.5	1.5	2.5
https://www.dietrichs.com	2	1.5	2
https://www.constructions-socopa.fr	2.5	1	3
https://www.circlecvi.com	3	1.5	3
https://www.cedes.com	1.5	2.5	2.5
https://www.bdvsolutions.com	2	2.5	3
https://www.bcn.co.uk	3	2	3
https://www.autodata-group.com	2	2.5	2.5
https://www.antser.com	3	2.5	3
https://www.ameco.com	2.5	1.5	3
https://www.adomex.nl	2	1	2.5
https://reladyne.com	1.5	1.5	2
https://mtderm.de	3	2.5	2.5
https://mellifiq.com	2.5	3	2.5

Table B.2: Expert evaluation scores (2/3)

Target	TF-IDF	Doc2Vec	Transformer
https://mayhealth.com	2.5	2.5	2.5
https://laborfirst.com	3	2	3
https://homerez.com	2.5	2.5	2.5
https://groupe-can.com	1	2.5	3
https://globeteam.com	2.5	2.5	2.5
https://exalta.com	3	2.5	3
https://dpsi.com	3	3	3
https://bombas.com	2.5	1.5	2
https://boards-and-more.com	2	2	2.5
https://bioserenity.com	3	2.5	3
https://axcell.se	2.5	2	2.5
https://astranis.com	3	2.5	3
https://ambinex.nl	3	2.5	2.5
http://www.vegafruits.fr	2	2.5	3
http://www.tectaamerica.com	3	3	2.5
http://www.targit.com	3	1.5	3
http://www.stats.com	3	3	3
http://www.prochilab.com	2.5	1.5	3
http://www.phoenix-spa.com	2	2.5	3
http://www.norskgjenvinning.no	3	2.5	3
http://www.integral-corp.com	3	2.5	2.5
http://www.iberconseil.es	2.5	1.5	2.5
http://www.dimex.fi	1.5	1	1.5
http://www.diabgroup.com	2.5	2	3
http://www.ap-ag.com	3	3	3
http://24pesula.fi	2.5	1	2.5

Table B.3: Expert evaluation scores (3/3)

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden - www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY