AI-Driven Peer Company Identification: A Semantic Text-Similarity Approach Beyond Traditional Industry Classification Systems

Timotej Jagrič
Institute of Finance and Artificial Intelligence, Faculty of Economics and Business, University of Maribor
Aljaž Herman
Institute of Finance and Artificial Intelligence, Faculty of Economics and Business, University of Maribor

Abstract

Background: Traditional classification systems, such as NACE and NAICS, primarily classify businesses by industry, limiting their ability to identify related companies. **Objectives:** This research aims to improve the identification of related companies by analysing their descriptions, utilising a more semantic approach. **Methods/Approach:** A pre-trained BERT model was employed to assess semantic text similarity for suggesting peer companies. The goal was to create a system that assists experts in comparing companies based on their descriptions, rather than developing a perfect classification tool. **Results:** Trained on publicly available data, the model achieved 73.6% accuracy in identifying related companies, with accuracy exceeding 90% for selected industry-pair combinations. **Conclusions:** While the system demonstrates promise, its outputs are intended to guide professionals who must ultimately validate the results. The findings emphasise the strengths and limitations of using AI models for this purpose, providing a foundation for future enhancements and real-world applications. However, the solution remains a conceptual idea, limited to only 13 industry categories, highlighting the need for broader testing and development.

Keywords: industrial classification schemes, peer companies, artificial intelligence, semantic text similarity, BERT

JEL classification: C55, G20, M15, N20
Paper type: Research article

Received: Oct 21, 2024
Accepted: Aug 15, 2025

Citation: Jagrič, T., Herman, A. (2026). AI-Driven Peer Company Identification: A Semantic Text-Similarity Approach Beyond Traditional Industry Classification Systems. *Business Systems Research, 17*(1), 204-222.
DOI: https://doi.org/10.2478/bsrj-2026-0010

Introduction

Recent advancements, particularly in technology, have significantly transformed the way we live and work. Tasks that once took months or even years can now be completed in mere days or hours. However, these rapid developments require individuals to remain adaptable and continuously informed about ongoing changes. Many still struggle to keep pace with innovation or are reluctant to adopt new technologies due to traditional mindsets. One of the most notable areas of progress is artificial intelligence, which, when applied appropriately, can offer substantial benefits (Makridakis, 2017). Related recent research has shown that ChatGPT-based classification can support decision-making in sensitive categorisation tasks, such as distinguishing urgent from non-urgent patient conditions (Fildor & Pejić Bach, 2023).

This article introduces a conceptual idea for identifying peer companies based on textual similarity in company descriptions, rather than relying on traditional industrial classification systems. This shift reflects broader changes in the business landscape, where firms are increasingly diversifying beyond their original industry labels. Driven by technological innovation, evolving market demands, and the integration of artificial intelligence and automation (Wright & Schultz, 2018), businesses are expanding their operational scope. As routine tasks are automated, companies can redirect human resources toward strategic and innovative functions. As a result, many firms now engage in a broader range of activities than when they were founded, blurring the boundaries between traditional industry categories.

The question is which industrial class or sector this company belongs to. This could be adequately determined using existing classification schemes (Phillips & Ormsby, 2016) such as Global Industry Classification Standard (GICS), Statistical classification of economic activities in the European Community (NACE), Standard Industrial Classification (SIC), or North American Industry Classification System (NAICS). The problem is this: companies have grown and diversified, so they may belong to more than one industry class, but classification schemes fail to capture the full scope of a company's activities. As a result, an enterprise falls into a single industrial category under the classification scheme, even though it also conducts activities in other categories. Consequently, the question is whether two companies can be related, even if they are classified in different sectors. The answer, as demonstrated, is affirmative, highlighting a significant challenge in identifying related enterprises for valuation purposes.

The objective of this study is to propose an artificial intelligence-based model. The key would be company profiles, in which companies tell us what they do and which activities they cover. The proposed model, which belongs to the Natural Language Processing group of language models, converts the descriptions into a self-consistent form and, based on this, computes the semantic similarity with other companies and their descriptions. In doing so, it considers all a company's activities, not just its primary activity. Therefore, the final results would be much more accurate, as similar companies would be identified based on all their activities, not just their primary activity. The degree of relatedness for such firms is therefore much higher.

In the following sections of this paper, a literature review is presented to provide context and highlight relevant research in the field. Next, in the materials and methods section, the dataset utilised for analysis is introduced, along with details of the implementation of the BERT transformer model. This is followed by the results and discussion, where the findings and their implications are analysed and compared with existing studies. Finally, a summary of key insights and potential directions for future research is provided.

Literature Review

Identifying peer companies is a critical task in finance, strategy, and market analysis. Traditional approaches typically rely on industry classification systems that group companies by their primary sector of operation. While convenient, this method often fails to capture the multidimensional nature of modern businesses. To address these limitations, recent research has explored more nuanced and data-driven techniques. This review synthesises the existing literature across three thematic categories: limitations of traditional classification; clustering and machine learning-based alternatives; and approaches leveraging textual and semantic analysis. This provides the foundation for our proposed Natural Language Processing (NLP)-based model, which evaluates firm similarity through semantic analysis of company descriptions.

Limitations of Traditional Classification

Standard classification systems, such as the SIC codes, are commonly used to define peer groups based on sectoral affiliation. However, they often overlook important dimensions such as geographic scope, customer base, or business model diversity. For instance, Eaton et al. (2022) criticise the reliance on SIC codes in mergers and acquisitions, finding that advisors instead strategically select high-valuation peers, regardless of industry code, to justify acquisition premiums. Similarly, Ding et al. (2019) argue that industry-based classifications are too rigid for varied research goals and propose that peer selection should be adapted to context-specific needs. These critiques collectively point to the inadequacy of industry labels in reflecting operational or financial similarity.

Data-Driven Clustering and Machine Learning Techniques

In response to these shortcomings, several studies have turned to clustering techniques that group firms based on shared quantitative features. Momeni et al. (2015) applied the K-means clustering algorithm to companies listed on the Tehran Stock Exchange using financial ratios. They supplemented it with the Analytic Hierarchy Process (AHP) to prioritise criteria. Similarly, Ding et al. (2019) employed K-medians clustering to align peer groups with financial deviations relevant to the research question. Recent business-systems research similarly demonstrates the value of deep learning for predictive modelling in decision-support contexts, including the early identification of at-risk students in online education (Shala Riza et al., 2025).

Machine learning techniques have further enhanced clustering capabilities. Husmann et al. (2022) used t-distributed stochastic neighbour embedding (t-SNE) and spectral clustering to visualise and group companies, facilitating more intuitive peer comparison in finance. These unsupervised learning methods reveal patterns in high-dimensional data that traditional classifications miss.

An innovative alternative is proposed by Noels et al. (2023), who developed a graph-based distance metric based on the Earth Mover’s Distance (EMD) to measure similarity between financial statements. This method demonstrates improved accuracy and efficiency in peer identification by leveraging the standardised structure of financial data.

Toward Text-Based and Perceptual Similarity

Beyond numerical data, a few studies have explored how firms may be grouped based on market perception or textual descriptions. Text-mining approaches are particularly relevant in this context, as they enable the extraction of meaningful patterns from large volumes of unstructured textual data and have already shown value in financial-sector applications (Pejić Bach et al., 2020) Bonne et al. (2022)

introduced an AI-driven peer-grouping system that clustered companies based on their market perception. Their model created continuous similarity measures to form hedged portfolios, showing that AI-generated clusters often align better with risk-return profiles than traditional sector-based classifications.

Our approach builds on this movement toward perceptual and semantic similarity by proposing a peer-identification model based on Natural Language Processing (NLP). Unlike prior studies that cluster firms based solely on financial metrics or external perception, our method extracts and analyses the semantic content of company descriptions. By transforming descriptions into a unified vector space using NLP techniques, we can evaluate similarity across all stated activities of a firm, capturing multidimensional roles that traditional classifications ignore.

Synthesis and Contribution

The reviewed literature reveals a growing consensus that traditional classification systems are insufficient for nuanced peer identification. Clustering and machine learning offer improvements, particularly when grounded in financial or performance data. However, many models still overlook what companies say they do – the textual self-representation of their activities, goals, and market positioning.

By introducing a semantic-based, NLP-driven approach, our work contributes to this evolving landscape by offering a scalable and adaptable framework. It integrates linguistic data into the peer-identification process, aligning firms based on how they describe themselves rather than on external labels. This approach better reflects the complexity of modern businesses and allows for more meaningful benchmarking, valuation, and strategic analysis.

Materials and Methods

Dataset

To develop a model to identify related companies to a given company, a freely accessible database available on the Kaggle platform was utilised (Puvvala, 2019). The dataset contains various types of information for nearly 70,000 companies. However, two specific variables were of primary importance for this study: (1) the company description and (2) the industry in which the company is classified. The criterion for firm-relatedness was that firms are related if they belong to the same industry. In this study, industry membership is used as a pragmatic proxy for relatedness during model training, rather than as a definitive measure of business similarity. This is an important limitation, since the purpose of the proposed model is precisely to move beyond rigid industry labels and support expert-based identification of comparable firms. However, firm descriptions were crucial because they serve as input to our model.

Our work was constrained to 13 industry categories due to limitations in publicly available data, whereas traditional classification schemes encompass a far broader range of sectors. However, this limitation is not intended as a critique of traditional industry classification systems, which serve an important role in organising firms by sector. Rather than seeking to replace these established systems, our objective is to introduce a complementary model that enables firm comparison based on textual company descriptions, extending beyond mere industrial affiliation. This approach focuses on the actual similarities between companies, enabling a more precise search for related companies, regardless of traditional classification. Additionally, in our model, similar companies are assigned solely based on their descriptions, eliminating the need for additional information such as country, size, or other attributes. To develop our model, a custom database was created, consisting of three columns: (1) the description of the first firm, (2) the description of the second firm, and (3) the relatedness, which had values of 1 if the two firms belonged to the same industry and 0 if the two firms belonged to different industries. An example of the database is presented in Table 1.

Table 1
Dataset example

Business description 1	Business description 2	Similarity
Heine Brothers' is a 100% fair trade & organic coffee roaster, with 17 shops in the Kentuckiana area. From the beginning, Heine Brothers has been committed to operating responsibly and sustainably – donating to numerous community groups, recycling, being a founding member of the world's 1st fair-trade coffee-buying cooperative and buying locally. With your support, we strive to do even more while offering exceptional drinks, perfectly roasted coffee, and friendly service.	Think it is about coffee? It is not. Really. It is not. We deliver a quality unlike anything in the coffee community. We have a vision to meet global needs and provide AMAZING coffee. When we import our coffee beans, they are not from a company; they are from a farmer with a family. Our partnership gives hope at a reasonable wage and a commitment to eradicate poverty from their lives. We know you could spend a lot less on coffee. You could even drive down to the big green store and purchase a good cup. Our vision is to bring together GREAT farmers to make a GREAT cup for a GREAT cause. It is our calling. It is not about coffee. It is not even about good coffee. It is about amazing people who want to make a huge difference.	1
Ethika Investments is a real estate investment management firm that specialises primarily in full-service hospitality and resorts, as well as office and multifamily/residential. Ethika's business model encompasses activities ranging from the acquisition of existing assets and property renovations to the redevelopment of existing properties, the purchase of raw land for improvements, and the sale of improved property. The firm acts as a principal and seeks to capitalise on its experience by focusing on the selective acquisition of well-located hotels and resorts, as well as institutional-quality assets that can be repositioned and rebranded to increase value. We tailor our acquisitions and strategies to meet the needs of our high-net-worth and family-office clientele.	Hospital Curtain Solutions understands the importance of completing your Hospital Cubicle Curtain and Track project on time and under budget. Many of the cubicle curtain fabrics used in the manufacture of Hospital Curtains Solutions Green hospital privacy curtains are made from recycled content, fire-retardant polyesters, or are manufactured with processes that reduce chemical use and consider methods such as waste management and recyclability. Our anodised aluminium cubicle curtain track is not only of the highest quality but also 100% recyclable. Hospital Curtain Solutions is also proud to be a woman-owned business enterprise and a leader in bringing both the most innovative eco-conscious hospital cubicle curtains and stylish.	0

Note: 1 represents similar business descriptions, 0 represents dissimilar business descriptions
Source: adopted from (Puvvala, 2019)

However, not all possible combinations of the two companies in the original database were included, as this would have resulted in a database of almost 2.5 billion records, which is far too large for developing a successful model and would have required excessive financial and time resources. Another problem that has arisen is that the original database contains descriptions that have not been "cleaned up", i.e., they contain special characters, numbers, repetitions of words, grammatical errors, etc. To address this issue, 30 random company descriptions were selected from each industry (13 industries in total), ensuring that the descriptions were of a reasonable length (averaging 700-1,000 characters) and that only those meeting specific criteria were utilised. These criteria included excluding special characters, duplicates, non-English words, and meaningless content.

This approach was followed to ensure the quality and consistency of our dataset, which is crucial for training the BERT transformer effectively. By randomly selecting 30 company descriptions from each of the 13 industries, a balanced representation across sectors is ensured. Filtering for reasonable length and excluding descriptions with special characters, duplicates, non-English words, or meaningless content helped eliminate noise and maintain linguistic coherence. These steps enhanced the dataset's relevance, enabling the model to identify meaningful similarities in company descriptions better.

At the same time, this ensured that enough examples of companies and their descriptions were included from each industry. In total, 390 descriptions were selected. To build the database used to train the model, all possible pairs of descriptions were generated. This resulted in a database of 75,855 description pairs. Of these, there were 5,655 pairs of descriptions between companies that were similar, so they were marked as 1, and 70,200 pairs of descriptions that were marked as 0, because the companies were not similar.

Although selection criteria were applied to ensure the business descriptions used for model training were appropriate, the final selection remained random, as the pool of suitable descriptions greatly exceeded 30 for each industry.

When examining the similarities between firms, it was evident that the ratio between similar (1) and dissimilar (0) firms strongly favoured dissimilar firms, with more than 90% of dissimilar firms and less than 10% of similar firms. Therefore, extra attention was paid to ensure sufficient descriptions of similar companies were included in the test set, which was then used to assess the model's goodness-of-fit. To address the significant imbalance in the dataset, a pre-existing function for the train-test split was used, maintaining the original proportion of similar and dissimilar firms in both the training and test sets. This approach ensured that the imbalance was consistently reflected throughout the model development process, allowing the model to learn and be evaluated under realistic conditions. By preserving the natural distribution of the dataset, artificial bias was avoided while maintaining the integrity of the original data proportions.

To construct a set for evaluating our model's performance in identifying similar businesses, a thorough approach was employed that ensured both within- and cross-category representation. The dataset was divided into training and testing subsets while preserving the class distribution of similar and dissimilar pairs. In addition, a separate manually structured evaluation set was prepared to assess the model's performance across all 13 industry categories and selected cross-category combinations. This distinction is important because the structured evaluation set was designed to provide balanced insight into both similar and dissimilar cases, rather than to reproduce the original class imbalance.

Within each category, 400 pairs of descriptions were randomly sampled, thereby ensuring enough cases for assessing similarity within each category. Additionally, to evaluate the model's ability to distinguish between descriptions of dissimilar companies across categories, pairs were crafted in which one description belonged to the current category. In contrast, the other belonged to a different one. This cross-category sampling yielded 408 pairs per category combination, for a total of 5,304 across all 13 categories. Combining these with the within-category pairs, the final test set comprised 10,504 description pairs. This comprehensive approach ensured that the test set was balanced and representative, encompassing a diverse range of similarity and dissimilarity scenarios, thereby providing a rigorous evaluation framework for our similarity model.

BERT Transformer

For our purposes, an advanced pre-trained transformer-based AI model, BERT (Bidirectional Encoder Representations from Transformers), is used. The BERT model is pivotal in natural language processing (NLP) for tasks like classification (Jagrič & Herman, 2024), named entity recognition (Chang et al., 2021), and question answering (Qu et al., 2019). Its strength lies in bidirectional transformer-based training, which contrasts with the unidirectional nature of conventional language models. Bidirectional training allows the model to better understand text meaning by considering both left and right contexts simultaneously. This capability enhances the model's understanding of contextual nuances, making it a powerful tool in NLP tasks (Devlin et al., 2019). Recent evidence also indicates that the framing of textual inputs substantially affects the quality of generative AI outputs, which further supports the need for carefully structured company descriptions in semantic-similarity applications (Car et al., 2025).

The Transformer architecture, fundamental to models like BERT, contains two key components: the encoder and the decoder (Vaswani et al., 2023). The encoder processes input sequences, generating contextualised representations of each token by considering the entire input sequence simultaneously. On the other hand, the decoder is primarily used for tasks that require autoregressive generation, predicting subsequent tokens based on previously generated ones. However, in the context of text-similarity tasks such as ours, the decoder component is omitted. This decision is rooted in text-similarity tasks, where the goal is to compare and measure the likeness between pairs of input sequences rather than generating new ones.

Furthermore, authors (Devlin et al., 2019) clarify how BERT is designed to handle various downstream tasks effectively. It features a flexible input representation that can accommodate single sentences or pairs of sentences. BERT utilises Word Piece embeddings (Wu et al., 2016) for tokenisation, with a vocabulary of 30,000 tokens, striking a balance between computational efficiency and effective representation. The Word Piece algorithm breaks words into smaller units called subword tokens. These subword tokens are then used as the basic units for building the model's vocabulary. This allows the model to represent a wider range of words, including those that it has not encountered during training. Each input sequence begins with a distinct classification token ([CLS]), whose final hidden state serves as a comprehensive representation of the entire sequence.

For additional semantic value, after tokenisation, each token is mapped to its corresponding token embedding. These embeddings are learned during BERT's pretraining and capture the semantic meaning of each token within the input sequence. Along with token embeddings, positional embeddings are added to the input representation. Positional embeddings encode the position of each token within the sequence, allowing BERT to understand the order of words in the input sentences. Positional embeddings are crucial for capturing sequential information in the Transformer architecture, where recurrence or convolution is absent. In the end, token embeddings and positional embeddings are typically concatenated to form the final input representation for BERT. This combined representation contains both token-level semantic information and positional information, enabling BERT to understand the meaning of each token within the context of the entire input sequence (Devlin et al., 2019).

After the input data undergoes tokenisation and positional embedding, several additional steps occur within the BERT architecture. These steps include (1) multi-head self-attention (Vaswani et al., 2023) to weigh the importance of each token's contextual information with respect to other tokens in the sequence, (2) residual connections (He et al., 2016), which enable the gradient flow during training, mitigating the vanishing gradient problem and facilitating the training of deep neural networks (Pascanu et al., 2013), (3) layer normalization (Lei Ba et al., 2016) to normalize the activations of the neurons within a layer, ensuring that the distribution of values remains stable during training, and (4) feedforward neural networks (Vaswani et al., 2023) that consist of two linear transformations separated by an activation function (typically ReLU), allowing the model to learn complex nonlinear relationships between tokens.

All of the above-mentioned procedures, tokenisation, positional embedding, multi-head self-attention, residual connections, layer normalisation, and feedforward neural networks, integral to semantic text similarity tasks with BERT, are comprehensively clarified in the paper authored by Jagrič & Herman (2024). In their work, they detail how BERT processes input sequences to generate contextualised representations that capture the complex semantic value of sentences. Although their research primarily focused on BERT for text classification, the same procedures are employed for semantic text similarity tasks, ensuring robust contextual understanding and comparison of input sentences.

After extracting contextual representations, various similarity metrics can be used to measure the semantic similarity between a pair of input sentences. These metrics may include cosine similarity (Rahutomo et al., 2012), Euclidean distance (Krislock & Wolkowicz, 2012), or other specialised similarity measures tailored to the specific task requirements (Santini & Jain, 1999).

In our case, the Binary Cross-Entropy Loss (BCEL) function is used, a special form of the Cross-Entropy Loss (Ruby et al., 2020). Authors (Mao et al., 2023) present Cross-Entropy as a cornerstone across various applications, serving as a prevalent loss function in neural networks. It aligns seamlessly with the logistic loss function, particularly when coupled with the softmax activation (Banerjee et al., 2020) at the neural network's output layer. This symbiotic relationship enables quantifying the discrepancy between predicted probabilities and observed outcomes, making cross-entropy a pivotal tool for tasks such as classification and probability estimation. Ruby et al. (2020) explained that the BCEL function is commonly used in machine learning tasks involving binary classification. It measures the difference between probability distributions, specifically when the labels are binary (0 or 1). In the context of our case for semantic text similarity with BERT, BCEL was used to evaluate the similarity between two pieces of text encoded by BERT. More specifically, between two pieces of encoded business descriptions. Binary cross-entropy measures the difference between two probability distributions over the same underlying set. The formula for BCEL is (Hurtik et al., 2022):

$$C(p,t) = \begin{cases} -\log(p)\,; & if\ t = 1 \\ -\log(1-p)\,; & if\ t = 0 \end{cases}, \tag{1}$$

where $tis\ an\ element\ of\ cap\ T, \in T$ is a label and $T = \{0, 1\}$ (In our case, label 1 is assigned when two business descriptions are similar and 0 if not) and $p\ \in (0,1)$ is the probability determined by our model. We can rewrite the above-mentioned formula and present it as (Hurtik et al., 2022):

$$C(p,t) = -\,(t \cdot \log(p) + (1-t) \cdot \log(1-p)). \tag{2}$$

During training, the goal is to minimise this loss across all sentence pairs in the training dataset. The model parameters are adjusted via backpropagation (Rojas, 1996) and an optimisation algorithm; in our case, we used AdamW (Kingma & Ba, 2017; Zhuang et al., 2022) to improve predictions and thereby reduce the loss. Eventually, the model learns to predict similarity scores closer to the ground-truth labels, thereby improving its performance on the similarity task.

Evaluation metrics

To evaluate the performance of our model, a systematic approach is employed by presenting a confusion matrix alongside several standardised performance metrics. The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering insights into the model's classification capabilities.

In addition to the confusion matrix, key performance metrics are calculated, including precision, recall, and the F1 score (Goutte & Gaussier, 2005). Precision measures the proportion of correctly identified related companies (true positives) out of all companies predicted as related (true positives + false positives). In the context of company similarity, high precision means that when the model predicts two companies are related, the prediction is likely to be accurate, which is crucial in applications where false similarities could mislead decision-making, such as benchmarking or investment comparisons.

$$Precision = \frac{TP}{TP + FP}. \tag{3}$$

Recall, also referred to as sensitivity, measures the ratio of true positive predictions to the total number of actual related companies (true positives + false negatives). In the context of company similarity, a high recall indicates that the model identifies most of the truly related companies. This is particularly important in scenarios where missing a relevant company (a false negative) could lead to missed opportunities.

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

F1 score balances precision and recall, providing a single metric for assessing the classifier's overall performance. A high F1 score indicates that the model is both accurate and comprehensive in its identification. This means it correctly identifies companies that truly belong to the same industry (high precision), while also capturing most of the relevant companies that should be considered similar (high recall). Such a result suggests that the model is effective not only at minimising false matches but also at avoiding the exclusion of relevant peers.

$$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{5}$$

Given the complexity of identifying truly comparable companies, all the mentioned evaluation metrics – precision, recall, and the F1 score – play a vital role in assessing the model's effectiveness. Each captures a different aspect of performance, from avoiding false matches to ensuring relevant peers are not overlooked. For this reason, these metrics were systematically applied to evaluate and refine our conceptual model, ensuring it provides both accurate and comprehensive identification of related companies based on textual descriptions.

Results and Discussion

After successfully training the model, an evaluation is conducted using our manually built test dataset. This evaluation aimed to assess the efficiency and effectiveness of the trained model in accurately predicting the similarity between pairs of descriptions. As previously outlined, the test dataset comprised 10,504 pairs of descriptions. Within this dataset, a balanced representation of similarity cases is ensured, with 5,200 pairs featuring descriptions deemed similar based on their category. Additionally, the dataset included 5,304 pairs of descriptions, intentionally selected to represent dissimilarity, with one description from each pair originating from a different category. By incorporating both similar and dissimilar cases in the test dataset, we were able to evaluate the model's performance thoroughly. The model results are presented in Table 2.

Out of the total 10,504 cases, the model made 7,736 correct predictions. These correct predictions encompassed scenarios in which the model accurately identified pairs of descriptions as either similar (indicating businesses from the same industry) or dissimilar (suggesting businesses from different industries). From the relative perspective, this indicates that the model demonstrated an overall accuracy of approximately 73.6% in its predictions.

Furthermore, when the model encountered pairs of business descriptions that were indeed similar, it correctly predicted 3,798 cases. However, it also misclassified 1,402 cases, erroneously identifying them as dissimilar, resulting in an accuracy of 73% for similar business descriptions. Conversely, when presented with dissimilar pairs of business descriptions, the model performed admirably. It accurately identified 3,938 cases as dissimilar, effectively identifying instances in which the corresponding businesses operated in different industries. However, there were 1,366 cases where the model incorrectly classified dissimilar descriptions as similar. From the perspective of non-similar business descriptions, the model accurately predicted in 74.2% cases.

This level of accuracy demonstrates the potential of this proposed method to significantly enhance the identification of both similar and dissimilar companies, offering a more nuanced, data-driven approach than traditional classification systems. While an overall accuracy of 73.6% and strong performance across both similar (73%) and dissimilar (74.2%) classifications indicate promising initial results, it is important to contextualise these findings within the scope and already mentioned limitations of the study.

Table 2
Model Output

Description 1	Description 2	Category 1	Category 2	Similarity	Predicted Similarity
Innovation. Techno-logy. Innovation. Industrial Electric Mfg™ (IEM) is the nation's largest independent full-line manufacturer of electrical distribution equipment and fully integrated systems. We offer quality…	BW Electrical are the leading domestic and commercial electrical contractor, serving the heart of…	Commercial Services & Supplies	Commercial Services & Supplies	1	1
Home | Visions has been committed to providing its customers with the highest-quality products through continuous innovation and product development…	Our team of Denver business brokers will help you buy, sell, and value small Colorado-owned businesses. Let us be your trusted broker during the process…	Materials	Corporate Services	0	1
A warm welcome to Abbey Veterinary Group – Your friendly independent two-centre veterinary practice based in Chaddesden and Chellaston, Derby…	Cheesecake Recruitment are a South Wales-based specialist recruitment agency providing jobs in Wales and beyond…	Healthcare	Corporate Services	0	0
World Cancer Research Fund UK is a UK cancer charity which funds and supports scientific research on how to prevent cancer with exercise, alcohol and diet…	Clear Coaching provides bespoke executive and business coaching for leaders in the corporate, public and third sectors. Mentoring | Supervision | Team Coaching…	Professional Services	Professional Services	1	0

Note: Category 1 = industry category of first business description, Category 2 = industry category of second business description. Both business descriptions are only partially presented.
Source: adopted from (Puvvala, 2019)

Despite those limitations, the results highlight the conceptual design's viability, demonstrating its ability to accurately process and compare company profiles even at this early stage. For potential users, this accuracy suggests a robust starting point for practical applications, with the understanding that future iterations incorporating a wider range of industries and larger datasets could further enhance its precision and generalizability, making it a more versatile tool for real-world decision-making.

To provide a more nuanced understanding of our model's performance, accuracy metrics are presented by industry class. In Table 3, the rows represent the first business description, while the columns represent the second business description. The entries in the table indicate the accuracy of the model's predictions for pairs of business descriptions from different industry classes.

Table 3
Model accuracy by industry-pair combination (%)

	CSS	H	M	F	EU	PS	CS	MMS	IT	CD	I	TL	C
CSS	71.5	91.2	76.5	70.6	82.4	82.4	70.6	85.3	88.2	79.4	76.5	94.1	88.2
H	85.3	83.0	76.5	85.3	88.2	91.2	76.5	94.1	82.4	85.3	88.2	88.2	85.3
M	76.5	94.1	69.0	85.3	88.2	97.1	94.1	97.1	76.5	91.2	91.2	91.2	94.1
F	82.4	73.5	76.5	79.8	70.6	85.3	88.2	91.2	82.4	76.5	82.4	88.2	76.5
EU	50.0	85.3	73.5	76.5	82.0	88.2	97.1	76.5	73.5	88.2	79.4	94.1	73.5
PS	67.6	76.5	61.8	52.9	70.6	73.3	67.6	70.6	67.6	82.4	82.4	79.4	94.1
CS	55.9	55.9	82.4	73.5	76.5	52.9	63.3	76.5	79.4	85.3	91.2	88.2	76.5
MMS	73.5	85.3	58.8	79.4	55.9	76.5	61.8	81.0	73.5	82.4	85.3	85.3	88.2
IT	67.6	58.8	67.6	67.6	64.7	64.7	67.6	52.9	65.8	76.5	73.5	79.4	67.6
CD	70.6	64.7	70.6	79.4	82.4	52.9	70.6	67.6	88.2	64.3	61.8	76.5	70.6
I	67.6	91.2	41.2	82.4	50.0	79.4	67.6	82.4	58.8	88.2	58.3	64.7	85.3
TL	20.6	52.9	52.9	35.3	47.1	20.6	29.4	50.0	41.2	50.0	47.1	75.3	70.6
C	76.5	76.5	52.9	73.5	79.4	73.5	55.9	70.6	67.6	73.5	73.5	76.5	83.3

Note: CSS = Commercial Services & Supplies; H = Healthcare; M = Materials; F = Financials; EU = Energy & Utilities; PS = Professional Services; CS = Corporate Services; MMS = Media, Marketing & Sales; IT = Information Technology; CD = Consumer Discretionary; I = Industrials; TL = Transportation & Logistics; C = Consumer Staples.
Source: Authors' work

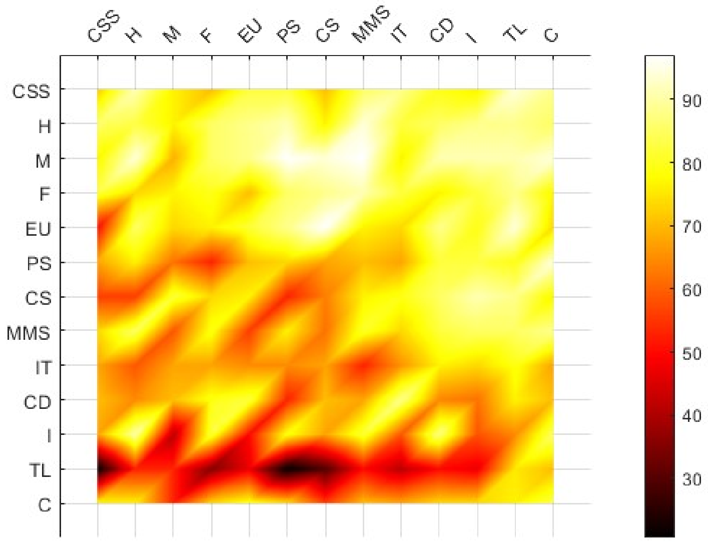
From Table 3, it is visible that the model performed exceptionally well in 18 cases, achieving accuracy rates of over 90% when predicting the similarity or dissimilarity between two business descriptions. More specifically, the highest accuracy of 97.1% was achieved when the model had to identify dissimilarity between pairs of Materials and Professional Services, Energy & Utilities and Corporate Services, and Materials and Media, Marketing & Sales.

However, there were some challenges, particularly in 3 cases where the model's accuracy dropped below 30%. This happened in cases where the model should recognise dissimilarity between business descriptions from Transportation & Logistics and Commercial Services & Supplies, Transportation & Logistics and Professional Services, and Transportation & Logistics and Corporate Services. To provide a clearer, more effective representation of the findings, a surface plot is presented in Figure 1.

Notably, the plot reveals an interesting pattern: higher accuracy is concentrated along the upper diagonal. This indicates that the model performs better when evaluating similarities between companies within specific clusters of two descriptions. However, it achieves inferior results when applied to the opposite combinations of these descriptions.

This issue could be mitigated by expanding the dataset's industry diversity and ensuring more accurate, professional business descriptions. Such enhancements would likely improve the model's ability to generalise and evaluate company similarities more effectively across a broader spectrum.

Figure 1
Surface Plot of Model Accuracy by Industry-Pair Combination



Source: Authors' work

Along with assessing the overall accuracy of this semantic text similarity model, a comprehensive analysis is conducted, including the preparation of confusion matrices and the calculation of various performance metrics to provide deeper insights into its effectiveness. The confusion matrices allowed us to visualise the distribution of correct and incorrect predictions made by the model across 13 industry classes (similar and dissimilar pairs of business descriptions). In Table 4, the components of the confusion matrix are shown in the first four columns, and the three performance metrics are shown in the last three columns.

Table 4
Confusion Matrix Components and Performance Metrics by Industry Category (%)

	TP	TN	FP	FN	Precision	Recall	F1
CSS	35.4	41.5	9.0	14.1	79.7	71.5	75.4
H	41.1	43.2	7.3	8.4	84.9	83.0	83.9
M	34.2	45.3	5.2	15.3	86.8	69.0	76.9
F	39.5	41.0	9.5	10.0	80.6	79.8	80.2
EU	40.6	40.2	10.3	8.9	79.8	82.0	80.9
PS	36.3	36.8	13.7	13.2	72.5	73.3	72.9
CS	31.3	37.6	12.9	18.2	70.9	63.3	66.9
MMS	40.1	38.1	12.4	9.4	76.4	81.0	78.6
IT	32.5	34.0	16.5	17.0	66.4	65.8	66.1
CD	31.8	36.0	14.5	17.7	68.7	64.3	66.4
I	28.8	36.1	14.4	20.7	66.8	58.3	62.3
TL	37.3	21.8	28.7	12.3	56.5	75.3	64.6
C	41.2	35.8	14.7	8.3	73.7	83.3	78.2

Note: The acronyms in the rows represent the industries as before. TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. Values for TP, TN, FP, and FN are reported as percentages.
Source: Authors' work

Caelen (2017) states that confusion matrix is a table that provides a summary of the performance of a classification model by presenting the counts of true positive (cases where the model predicted the positive class correctly), true negative (cases where the model predicted the negative class correctly), false positive (cases where the model predicted the positive class incorrectly), and false negative (cases where the model predicted the negative class incorrectly) predictions. Those values are presented in columns 2-5.

From Table 4, it can be seen that 3 typical performance metrics are calculated to validate our model and its performance further: (1) precision, (2) recall, and (3) F1 score. These matrices are used as additional performance indicators.

The calculated metrics from Table 4indicate that (1) the overall performance seems reasonable, with F1 scores ranging from around 62% to 84% and (2) a good balance of our model. The latter follows from the trade-off between precision and recall (Gordon & Kochen, 1989). When precision and recall values are similar, it suggests that the model is achieving a balance between correctly identifying relevant instances (precision) and capturing all relevant instances (recall). In other words, the model is making fewer mistakes in terms of both false positives (items incorrectly identified as belonging to a class) and false negatives (items incorrectly not identified as belonging to a class). In this case, the difference between precision and recall was acceptable in 8 out of 13 industry categories.

However, there were cases where the trade-off favoured either precision or recall. Authors explain that, when precision is much higher than recall, the model is conservative and cautious in its predictions (Gordon & Kochen, 1989). On the contrary, when recall is much higher than precision, the model is more liberal in its predictions, aiming to capture as many relevant instances as possible, even at the cost of higher false positives. In this case, in 5 out of 13 industry classes, the trade-off between precision and recall was slightly off. In Commercial Services & Supplies, Materials, and Industrials, precision was higher compared to recall. In industry classes of Transportation & Logistics and Consumer Staples, the recall significantly exceeded precision.

A closer examination of precision, recall, and F1 scores across industries highlights specific strengths and areas for improvement in the model's performance. For example, in industries like Materials and Commercial Services & Supplies, where precision is higher than recall, the model excels at avoiding false positives, suggesting it is conservative in classifying these industries as similar. This approach is beneficial in applications where misclassifying dissimilar companies is costly. On the other hand, in Transportation & Logistics and Consumer Staples, where recall significantly exceeds precision, the model prioritises capturing similar pairs over fewer false positives. This trade-off may be suitable for applications where identifying all potential peers is more critical than avoiding overgeneralization. These variations suggest that the model's performance is context-sensitive, with its behaviour influenced by the underlying characteristics of industry-specific data, and that it could be further optimised by fine-tuning its parameters or introducing industry-specific adjustments.

In this context, one limitation of our model is that it was trained on data from only 13 industries, which is insufficient for widespread practical use. This limitation limits its generality, as it does not cover all industrial sectors to which companies could be compared. As a result, relying solely on this AI-driven approach for peer company identification in uncovered sectors may lead to inaccuracies or biased results. Nevertheless, the main objective was to present the conceptual design of the model, which is already performing well at this early stage. Further extension to a larger set of industries and data could further improve its usefulness and accuracy in a real-world setting.

Additionally, this approach does not focus on where a company would be classified in the traditional industrial classification, but rather on finding related companies based on descriptions. A potential source for later development could be the company's annual reports, which provide comprehensive and reliable information. The model is therefore not designed as a classification system, but as a tool for suggesting similar companies. The model's precision and recall results reflect its ability to identify companies with descriptive similarity, but they should not be taken as a final judgment on the relevance of the suggested companies.

The final evaluation should be carried out by an expert who, based on the model's suggestions, assesses whether the proposed companies are sufficiently comparable for a particular analysis or need. The model serves as a support in the search for similar companies, but human expert review is necessary to ensure relevance and accuracy.

Conclusions

This research addressed the challenge of identifying peer companies beyond traditional industry classification systems. While conventional schemes such as NACE, GICS, SIC, or NAICS define relatedness solely on the basis of shared industry codes, this approach often overlooks meaningful similarities between companies that operate across sectors or have diversified activities. To overcome these limitations, our study proposed a novel, text-based method that leverages Natural Language Processing (NLP) to analyse company descriptions and identify peers based on actual business activities rather than predefined industrial categories. This approach offers a more nuanced and flexible way to capture company-relatedness, with potential applications in finance, market analysis, and competitive benchmarking.

The accelerating pace of technological and economic change has highlighted limitations in traditional classification systems. This is particularly evident in the case of large companies that continually diversify their operations, making it increasingly difficult to assign them to a single, well-defined industry category. This prevents the proper identification of similar enterprises, because two enterprises may be similar but are labelled as dissimilar due to overly narrow categorisation. However, the rapid development of the world does not necessarily pose problems only in terms of finding similar enterprises as they expand their activities. With rapid development, new technologies such as machine learning and artificial intelligence methods have emerged, helping us move beyond classical classification schemes and providing a more optimal tool for discussing the search for related companies. To this end, this study presents the conceptual design of the model used to identify related companies.

Therefore, in this study, a pre-defined BERT model based on the transformer architecture is used. To train the model, a database containing company descriptions and the corresponding industry category is included. From this database, we generated a new database consisting of three columns: the first two columns contained the descriptions of the first and second companies, and the third column was assigned the value 1 if the two companies were from the same industry and 0 otherwise.

To test the accuracy of our model, we methodically crafted a test dataset comprising enough of both similar and dissimilar pairs of descriptions, enabling a comprehensive assessment of the model's performance. Overall, the model achieved approximately 73.6% accuracy, with nuanced performance across industry classes. Additionally, our analysis revealed instances of exceptional accuracy, with rates exceeding 90%. However, challenges were also observed when the model's accuracy dropped below 30%.

Further validation using confusion matrices enabled a detailed examination of the model's performance, including the calculation of precision, recall, and F1 scores. These metrics highlighted a reasonable overall performance range of F1 scores between 62% and 84%. Additionally, a notable observation was the trade-off between precision and recall, with acceptable balances in 8 out of 13 industry categories. However, in 5 out of 13 industry classes, the trade-off between precision and recall deviated, with instances of either higher precision or higher recall. This discrepancy indicates areas where the model's cautiousness or liberality in predictions may need refinement.

Additionally, the proposed model is not intended to classify enterprises into traditional industry categories, so it does not serve as a critique of existing classification schemes like NACE or NAICS, which still play an important role in organising companies by industry. Instead, the focus of our model is entirely different; it aims to analyse a company's description and identify peer companies based on descriptive similarity rather than industry affiliation. This approach allows for more flexible and nuanced comparisons between companies, especially in dynamic or emerging sectors where traditional classifications may not fully capture evolving business activities.

Finally, it is important to emphasise that our model's output is intended to be used in conjunction with professional judgment. The model serves as a tool to help evaluators find peer companies based on descriptive similarities, but it is not a substitute for expert evaluation. The final decision on whether the suggested companies are truly comparable should be made by the valuator, who can apply their knowledge and insights to assess the relevance and suitability of the model's suggestions in a specific context.

Overall, the proposed model should be interpreted as a proof-of-concept rather than a fully deployable classification system. Its main value lies in demonstrating how semantic text similarity can support peer-company identification beyond rigid industry labels. Future research should extend the dataset to more industries, use more reliable company descriptions such as annual reports, and include expert validation of the suggested peer companies.

References

1. Banerjee, K., Prasad C, V., Gupta, R. R., Vyas, K., Anushree, H., & Mishra, B. (2020). *Exploring Alternatives to Softmax Function*. https://doi.org/10.48550/arXiv.2011.11538
2. Bonne, G., Lo, A. W., Prabhakaran, A., Siah, K. W., Singh, M., Wang, X., Zangari, P., & Zhang, H. (2022). An Artificial Intelligence-Based Industry Peer Grouping System. *Journal of Financial Data Science*, *4*(2), 9–36. https://doi.org/10.3905/jfds.2022.1.090
3. Car, T., Šimac, I., & Šuman, S. (2025). Framing prompts as user stories: Effects on the output quality of generative AI. ENTRENOVA - ENTerprise REsearch InNOVAtion, 11(1). https://doi.org/10.54820/entrenova-2025-0057
4. Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3–4), 429–450. https://doi.org/10.1007/s10472-017-9564-8
5. Chang, Y., Kong, L., Jia, K., & Meng, Q. (2021). Chinese named entity recognition method based on BERT. *Proceedings of 2021 IEEE International Conference on Data Science and Computer Application, ICDSCA 2021*, 294–299. https://doi.org/10.1109/ICDSCA53499.2021.9650256
6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 16. https://doi.org/10.48550/arXiv.1810.04805
7. Ding, K., Peng, X., & Wang, Y. (2019). A machine learning-based peer selection method with financial ratios. *Accounting Horizons*, *33*(3), 75–87. https://doi.org/10.2308/acch-52454

8. Eaton, G. W., Guo, F., Liu, T., & Officer, M. S. (2022). Peer selection and valuation in mergers and acquisitions. *Journal of Financial Economics*, *146*(1), 230–255. https://doi.org/10.1016/j.jfineco.2021.09.006
9. Fildor, D., & Pejić Bach, M. (2023). Testing the ability of ChatGPT to categorise urgent and non-urgent patient conditions: Who ya gonna call? ENTRENOVA - ENTerprise REsearch InNOVAtion Journal, 9(1), 101–112. https://doi.org/10.54820/entrenova-2023-0010
10. Gordon, M., & Kochen, M. (1989). Recall-precision trade-off: A derivation. *Journal of the American Society for Information Science*, *40*(3), 145–151. https://doi.org/10.1002/(SICI)1097-4571(198905)40:3<145::AID-ASI1>3.0.CO;2-I
11. Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Lecture Notes in Computer Science*, *3408*, 345–359. https://doi.org/10.1007/978-3-540-31865-1_25
12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
13. Hurtik, P., Tomasiello, S., Hula, J., & Hynar, D. (2022). Binary cross-entropy with dynamical clipping. *Neural Computing and Applications*, *34*(14), 12029–12041. https://doi.org/10.1007/s00521-022-07091-x
14. Husmann, S., Shivarova, A., & Steinert, R. (2022). Company classification using machine learning. *Expert Systems with Applications*, *195*. https://doi.org/10.1016/j.eswa.2022.116598
15. Jagrič, T., & Herman, A. (2024). AI Model for Industry Classification Based on Website Data. *Information*, 19. https://doi.org/10.3390/info15020089
16. Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization*. 15. https://doi.org/10.48550/arXiv.1412.6980
17. Krislock, N., & Wolkowicz, H. (2012). Euclidean distance matrices and applications. *International Series in Operations Research and Management Science*, *166*, 879–914. https://doi.org/10.1007/978-1-4614-0769-0_30
18. Lei Ba, J., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization*. 14. https://doi.org/https://doi.org/10.48550/arXiv.1607.06450
19. Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006
20. Mao, A., Mohri, M., & Zhong, Y. (2023). Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *40th International Conference on Machine Learning*, 23803–23828.
21. Momeni, M., Mohseni, M., & Soofi, M. (2015). Clustering Stock Market Companies via K-Means Algorithm. *Kuwait Chapter of Arabian Journal of Business and Management Review*, *4*(5), 1–10. https://doi.org/10.12816/0018959
22. Noels, S., De Ridder, S., Viaene, S., & De Bie, T. (2023). An efficient graph-based peer selection method for financial statements. *Intelligent Systems in Accounting, Finance and Management*, *30*(3), 120–136. https://doi.org/10.1002/isaf.1539
23. Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *30th International Conference on Machine Learning, ICML 2013*, *PART 3*, 2347–2355.
24. Phillips, R. L., & Ormsby, R. (2016). Industry classification schemes: An analysis and review. *Journal of Business and Finance Librarianship*, *21*(1), 1–25. https://doi.org/10.1080/08963568.2015.1110229
25. Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2020). Text mining for big data analysis in financial sector: A literature review. Sustainability, 12(3), 1155. https://doi.org/10.3390/su12031155
26. Puvvala, C. (2019). *Company classification*. Retrieved from: https://www.kaggle.com/datasets/charanpuvvala/company-classification/data
27. Qu, C., Yang, L., Qiu, M., Bruce Croft, W., Zhang, Y., & Iyyer, M. (2019). BERT with history answer embedding for conversational question answering. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1133–1136. https://doi.org/10.1145/3331184.3331341

28. Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic Cosine Similarity. *The 7th International Student Conference on Advanced Science and Technology*. Retrieved from https://www.researchgate.net/profile/Faisal-Rahutomo/publication/262525676_Semantic_Cosine_Similarity/links/0a85e537ee3b675c1e000000/Semantic-Cosine-Similarity.pdf
29. Rojas, R. (1996). The Backpropagation Algorithm. *Neural Networks*, 149–182. https://doi.org/10.1007/978-3-642-61068-4_7
30. Ruby, U. A., Theerthagiri, P., Jacob, J. I., Vamsidhar, Y. (2020). Binary cross entropy with deep learning technique for Image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 5393–5397. https://doi.org/10.30534/ijatcse/2020/175942020
31. Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(9), 871–883. https://doi.org/10.1109/34.790428
32. Shala Riza, L., Abazi Bexheti, L., & Zoroja, J. (2025). Early identification of at-risk students in online education: A deep learning approach to predictive modelling. Business Systems Research, 16(2), 69–91. https://doi.org/10.2478/bsrj-2025-001
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N., G., Kaiser, Ł., & Polosukhin, I. (2023). *Attention Is All You Need*. 15. https://doi.org/10.48550/arXiv.1706.03762
34. Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons*, *61*(6), 823–832. https://doi.org/10.1016/j.bushor.2018.07.001
35. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 23. https://doi.org/10.48550/arXiv.1609.08144
36. Zhuang, Z., Liu, M., Cutkosky, A., & Orabona, F. (2022). *Understanding AdamW through Proximal Methods and Scale-Freeness*. 24. https://doi.org/10.48550/arXiv.2202.00089

About the authors

Timotej Jagrič received his first PhD from the University of Maribor and his second PhD from the University of Primorska. He is certified in quantitative risk management (CQRM), a full professor of applied economics and econometrics, and a full professor of finance. He works at the University of Maribor, in the Department for Finance and is Head of the Institute for Finance and Artificial Intelligence. He teaches at the undergraduate, graduate, and PhD levels at the Faculty of Economics and Business, the Faculty of Natural Sciences and Mathematics, and at other Universities in and outside Slovenia. The author can be contacted at **timotej.jagric@um.si**.

Aljaž Herman is currently pursuing a PhD in Economics. He has a master's degree in mathematics and is proficient in statistics, Python, MATLAB, and SQL. He is a researcher at the Institute of Finance and Artificial Intelligence and an assistant at the Faculty of Economics and Business, University of Maribor. His passion for data science, analytics, and risk management drives his work, blending academic rigour with real-world applications. The author can be contacted at **aljaz.herman@um.si**.